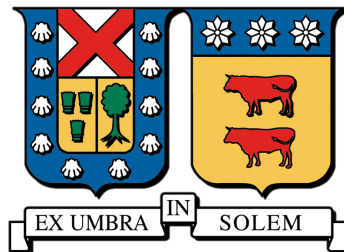


Módulo Web para la Automatización de Inventario entre Proveedores y Pymes (B2B) con Soporte para Toma de Decisiones



Juan Pablo Caniguante Cordero

Departamento de Informática
Universidad Técnica Federico Santa María

Memoria para la obtención del grado de
Ingeniería Civil Informática

2024

Índice general

1. Introducción	1
1.1. Definición del problema	2
1.1.1. Caracterización del problema	2
1.2. Objetivos	3
1.2.1. Objetivo general	3
1.2.2. Objetivos específicos	3
2. Estado del arte	5
2.1. Modelos de negocio	5
2.1.1. Comercio electrónico (e-commerce)	7
2.1.2. Business to business (B2B)	7
2.2. Sistemas de información	9
2.3. Aplicaciones y módulos Web	10
2.4. Interfaz de programación de aplicaciones (API)	12
2.4.1. API Rest	13
2.5. Sistemas de recomendación	14
2.5.1. Desafíos en sistemas de recomendación	15
2.5.2. Tipos de datos	17
2.5.3. Modelos en sistemas de recomendación	18
2.5.4. Evaluación del sistema de recomendación	24
2.5.5. Toma de decisiones en un e-commerce	25
3. Desarrollo del módulo	27
3.1. Consideraciones previas	27
3.1.1. Modelo de negocio de CGTEC	28
3.1.2. Servicios de Mercado Libre	28
3.2. Sistema de Recomendación	29
3.2.1. Preprocesamiento de datos	30
3.2.2. Exploración de datos	32
3.2.3. Construcción del Modelo de recomendación	42
3.2.4. Plataforma de Mercado Libre	47
3.2.5. Interpretación y recomendaciones	48
3.2.6. Iteración y mejora	48
3.3. Desarrollo de aplicación web	49

3.3.1. Modelo de Dominio	50
3.3.2. Tecnologías	51
3.3.3. Interfaz Visual	53
4. Resultados y Análisis	61
4.1. Gráficos de Varianza y Tiempo	62
4.2. Tabla de vector TF-IDF	62
4.3. Tabla de Resultados	64
4.3.1. Cargadores	64
4.3.2. Baterías	65
4.3.3. Pantallas	66
4.3.4. Discos Duros y SSD	68
4.4. Test del Sistema de Recomendación	69
4.4.1. Por categoría	69
4.4.2. Por técnica	70
4.5. Análisis y Discusión	72
4.6. Limitaciones	75
5. Conclusiones y Trabajos Futuros	77
5.1. Conclusiones	77
5.2. Trabajo Futuro	80
Bibliografía	81
Apéndice A. Gráficos de varianza de SVD y similitudes	83
Apéndice B. Tabla de resultados de vector TF-IDF	89
B.1. Cargadores y Fuentes	89
B.2. Baterías	91
B.3. Pantallas LCD	93
B.4. Discos duros y sólidos	95
Apéndice C. Tabla de resultados ordenados por Categoría	97
C.1. Cargadores y Fuentes	98
C.2. Baterías	103
C.3. Pantallas LCD	108
C.4. Discos Duros y SSD	114
Apéndice D. Gráficos Box Plot de similitud ordenados por categoría	119
D.1. Cargadores y Fuentes	120
D.2. Baterías	121
D.3. Pantallas	122
D.4. Discos Duros y SSD	123

Apéndice E. Gráfico de barras similitud v/s producto	125
E.1. Cargadores y Fuentes	126
E.2. Baterías	127
E.3. Pantallas	128
E.4. Discos Duros y SSD	129
Apéndice F. Resultados de evaluación de Sistema de recomendación	131
F.1. Cargadores y Fuentes	132
F.2. Baterías	133
F.3. Pantallas LCD	134
F.4. Discos Duros y SSD	135

Le dedico este trabajo a mis padres, que tanto se han esforzado por darme educación y que no me faltara nada, los quiero un montón.

Agradecimientos

Me gustaría agradecer a todos quien me han dado apoyo durante estos años de carrera de universidad, a mi mamá por estar en mis peores días, por mi papá por motivarme a siempre seguir adelante y a mi familia por no dejar que me hunda en mi mismo.

Resumen

Los modelos de negocio *e-commerce* consisten en la venta, compra, distribución y suministro de información relacionada a productos y servicios mediante la utilización de medios electrónicos. En este contexto, el uso y la necesidad de sistemas que permitan la compra de productos, gestión de inventario, actualización de catálogo, visualización de proveedores y monitoreo de compras y ventas, son procesos claves en tiendas con almacenamiento físico, por ello han emergido varios sistemas Business to Business (B2B) que buscan satisfacer estas necesidades. Sin embargo, estos no realizan otros procesos de vital importancia, como: extracción de información relevante y específica de fuentes estimadas (catálogos de proveedores), clasificación de productos e información personalizada a cada empresa/cliente, estimación de productos que entreguen una mayor ganancia a la empresa, priorización de aspectos del producto al momento de la compra. Es por esto, que en esta memoria proponemos una aplicación WEB, que permita obtener los catálogos de los proveedores de forma eficiente, gestione el inventario en tiempo real de la empresa, permita calificar cada proveedor y producto, recomiende órdenes de compra (en base al contexto de cada empresa) y sea lo suficientemente escalable para agregar otras plataformas y proveedores. Adicionalmente, este sistema contribuirá a la toma de decisiones de los clientes, apoyándose en el uso del vector TF-IDF, cálculo de similitud entre productos y la exploración de diferentes técnicas, con la finalidad de determinar cuál de estas entrega una mejor recomendación para la empresa.

Capítulo 1

Introducción

En los tiempos antiguos, el comercio se inicia como un sistema de trueque, en donde se intercambiaban productos por otros, siendo principalmente objetos de esas épocas (pieles de animales o alimentos) el cual gradualmente se fue estableciendo con el uso de monedas y con la globalización y la extensión de los medios de comunicación (principalmente internet) el comercio adaptándose a las nuevas tecnologías digitales.

La evolución del comercio electrónico desde sus inicios en Estados Unidos en los años 1920 mediante la venta por catálogo ha experimentado grandes cambios y avances, incluyendo tecnologías de información y comunicación para el apoyo de procesos, siendo en el año 1960, el nacimiento de una herramienta que permitiría la transacción de datos financieros de manera electrónica, el Electronic Data Interchange (EDI) [Cárdenas, 2020]. Entre sus numerosas ventajas se encuentra la capacidad de realizar compras y ventas en medios no físicos, mostrar catálogos a una mayor cantidad de personas y generar transacciones a nivel mundial y generando mejoras a los modelos de negocio convencionales, definiéndose múltiples modelos de comercio electrónico, siendo los principales: **Business To Client (B2C) y Business to Business (B2B)**, siendo este último el que desarrolla este trabajo.

Dentro del campo del comercio electrónico existen varios softwares y sistemas que son parte de este mismo modelo, entregando diferentes soluciones con otras ventajas y desventajas, siendo Mercado Libre el principal representante en el comercio electrónico latinoamericano, siendo un servicio entregado netamente vía internet y su página web, siendo principalmente enfocadas al modelo B2C; además, con el surgir de este tipo de comercio, empresas como ebay, aliexpress, wish (entre otras) entrega servicios de compra/venta con alcance internacional. Siendo las principales plataformas que tienen una gran importancia para las pequeñas y medianas empresas, dada la ventaja de poder seleccionar y visualizar los productos y repuestos entre diferentes proveedores, brindando una flexibilidad en la compra ya sea de productos de mercado o tecnológicos, permitiendo obtener productos a menor precios de otros países sin necesidad de tener convenios especiales con la empresa particular.

Lo destacable de estas mismas plataformas es que ocupan la información del usuario (como las compras, productos que muestren interés, gustos o productos más visitados, etc) para predecir o calcular productos/proveedores que podrían entregarle un valor a este usuario, mostrándolos en la página inicial recomendando productos relacionados a los ya comprados. Por otro lado, existen otros mecanismos de recomendación basado en la información del producto

en vez del usuario, siendo el caso que algún item tenga muchas ventas u otra información que denote un mayor valor para los usuarios, por lo que Mercado Libre los marca como “productos recomendados” en su lista de catálogo, entregando un descuento de precio al sobrepasar cierto umbral de compras del item; siendo este último modelo un sistema de recomendación basado en items o filtrado colaborativo, y el primero basado en usuarios, siendo los dos principales de distintos modelos de recomendación, entre ellos el basado en contenido, el cual aplica para casos con información restringida, siendo el tipo de modelo que se ocupa en este trabajo.

1.1. Definición del problema

Dentro de este modelo de negocio, existen varias clasificaciones según los actores en la comunicación, los más reconocidos son: Business To Business (B2B), Business to Client (B2C) y Business to Employee (B2E), entre otros [L., 2021]. Estos modelos tienen en cuenta el contexto, características y participantes en una relación comercial, como la valorización del comprador y el vendedor, afectando en como se interaccionan y que productos son comprados. Dado esto, los principales beneficiados son aquellas empresas que son parte del sistema, ya sea comprando o vendiendo sus productos o servicios, aumentando considerablemente su distribución, producción y venta de artículos a otras entidades.

1.1.1. Caracterización del problema

Muchas de estas empresas no poseen algún sistema que les de apoyo en estos procesos ni tampoco para gestionar su inventario, dada su dificultad de emplear las nuevas tecnologías, estiman como conveniente seguir utilizando métodos antiguos, tal sería el uso de registros físicos en la tienda. Medio que acarrea problemas tales como, información de los productos en la sucursal con difícil acceso, dificultad para realizar análisis numérico de inventarios, nula integración de datos al realizar una emisión de compra al proveedor y métodos no fiables para determinar fechas con mayores ofertas. Con esto, negocios de gran magnitud se posicionan en una clara ventaja frente a otros que sí posean los recursos para obtener un sistema que proporcione apoyo y soporte en el problema.

El proceso de abastecimiento de productos es fundamental para una empresa que vende artículos regularmente, por lo que, tener una eficiente comunicación con proveedores le entrega mayores beneficios. Sin embargo, las fuentes de información (catálogos de productos) están desplegadas y no centralizadas, resultando incómodo e ineficaz encontrar otro proveedor el cual iguale o mejore la calidad del producto, del servicio de entrega, que proporcione seguridad de distribución y mayores garantías en caso de fallos, dado esto, el criterio de las Pymes para seleccionar al proveedor se basa en su propia experiencia y la opinión que obtienen de empresas análogas. Hecho que genera un sesgo en las opciones de compra disponibles, evitando el obtener los productos con mayor ganancia, que concluye en una desventaja competitiva frente a sus rivales de mercado. Esto se puede apreciar en la Figura 1 1.1 que representa la situación.

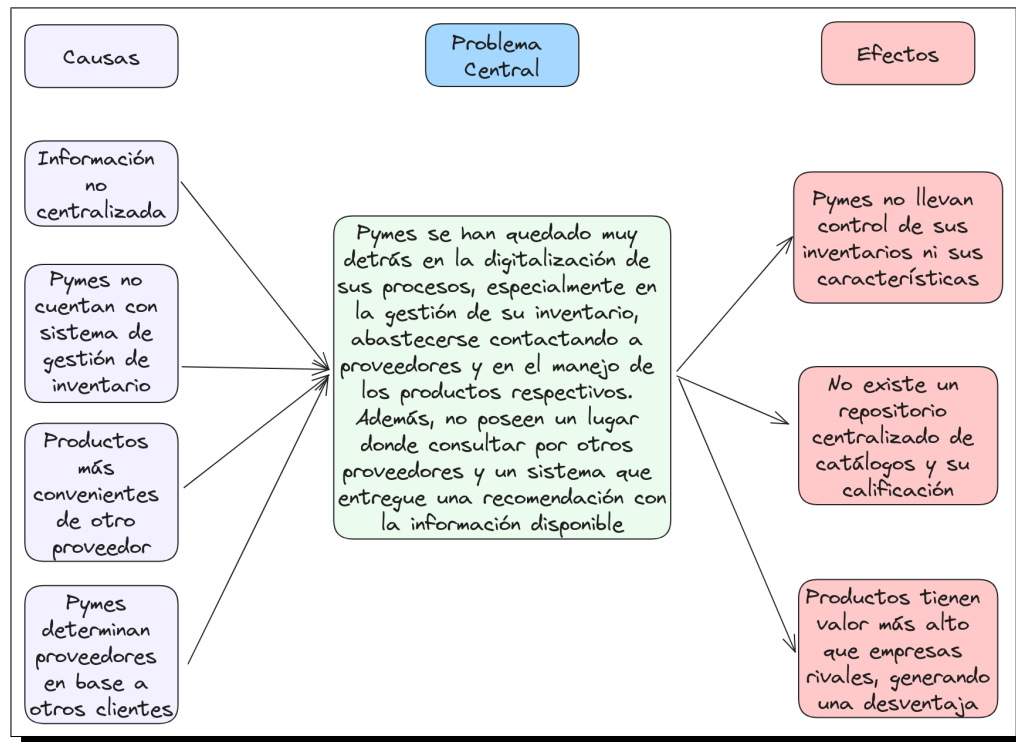


Figura 1.1 Árbol del problema que representa causa, problema y consecuencias. Fuente: creación propia

1.2. Objetivos

En esta sección se describirá los objetivos a cumplir en este trabajo, primero de forma general para luego definir punto a punto lo que se realizará.

1.2.1. Objetivo general

Diseñar y desarrollar un módulo informático basado en una aplicación web que permita a los proveedores distribuir su catálogo de productos y a los clientes emitir órdenes de compra seleccionando a qué proveedor comprar ciertos productos. Estas decisiones serán tomadas en base a recomendaciones propuestas por el sistema usando la información disponible de los productos y de los proveedores.

1.2.2. Objetivos específicos

- Desarrollar una aplicación web que permita gestionar a los proveedores la distribución de su catálogo de productos desde fuentes similares.
- Modelar y crear una base de datos la cual permita almacenar los productos, proveedores, usuarios e inventario de las empresas registradas en el sistema, la cual tendrá un servidor que permita gestionar esta información.
- Diseñar y desarrollar un esquema de recomendación basado en contenido que permita determinar los beneficios monetarios de comprar cierto producto entre varios provee-

dores, en base a sus atributos (e.j., calidad del producto, precio, eficiencia de entrega), prioridad que defina el cliente y calificación entregada por los usuarios.

- Diseñar y constituir una Application Programming Interface (API) que entregue el listado de productos obtenidos en el esquema de recomendación, con el objetivo de que la aplicación web pueda consumirla.
- Analizar la información y prioridades que tenga la empresa usuaria para entregar una recomendación de productos, que apoye a la toma de decisiones de compra.

Capítulo 2

Estado del arte

En este capítulo se exponen aquellos conceptos vitales para el entendimiento del módulo propuesto, dando su inferencia en los sistemas de comercio electrónico (e-commerce), como otras empresas y/o softwares han abarcado esta característica y explorar otras investigaciones: sus aciertos, fallos, desafíos enfrentados y conclusiones al momento de proponer metodologías de recomendación. Técnicas como: *Vector TF-IDF*, *SVD*, *Stemming*, *Palabras Claves*, *Umbral*, serán mencionadas y analizadas en este capítulo, con el fin de determinar cual es el mejor camino para resolver el problema definido. Además, se abordará como es el proceso de realizar una recomendación y que tan vital resulta para una empresa obtener mayor captación de clientes en un contexto e-commerce.

2.1. Modelos de negocio

Un modelo de negocio es una estructura conceptual que detalla la forma que opera una empresa y como alcanza sus objetivos; además de detallar como ofrece valor al cliente [CERTUS, 2020]. Los modelos de negocios han sido tema de discusión e investigación de acuerdo a su definición y objetivos que persiguen, existiendo disímiles investigaciones que han aportado en uno o más subdominios de un modelo de negocio [Pateli and Giaglis, 2008]. Entre estas definiciones, se determina la esencia del modelo de negocio el actuar como nexo entre una estrategia de negocio, un proceso de negocio, y culminando en un sistema de información. Un sistema de información es un conjunto de componentes interrelacionados que recopila (o recupera), procesa, guarda y distribuye información para dar soporte a la toma de decisiones y control en una organización [López, 2016]. La estrategia de negocio apunta a los objetivos de la organización (crecimiento de la marca, solidez financiera, etc.) y de qué forma se realizará. Además, el modelo de negocio se encarga de definir el marco conceptual, por ejemplo, componentes, taxonomía, metodologías, etc. Todo esto converge en procesos de negocio que detallarán las actividades, controles y métodos a utilizar en ciertas secciones de su ejecución, aspecto que en cada sistema se diferencia y se aprecia en la siguiente figura.

Un aspecto fundamental dentro de un modelo de negocio son los componentes que lo conforman:

- **Cliente:** Principal actor dentro del modelo, dado su rol de generar datos y ser el objetivo de captación de las empresas.

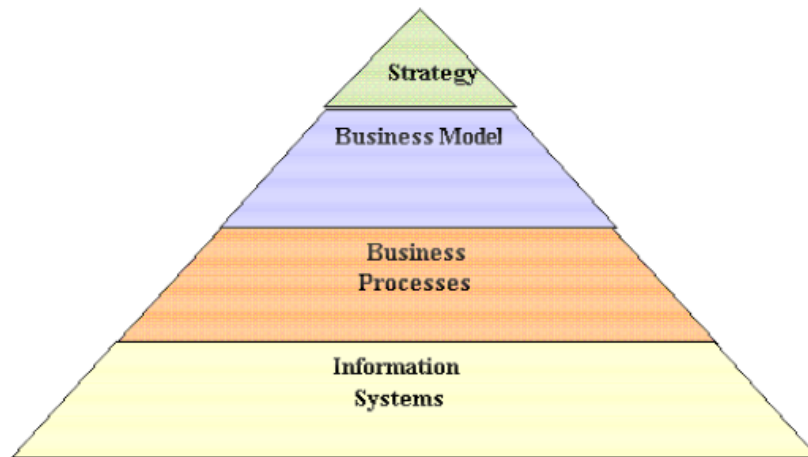


Figura 2.1 Ilustración relacionada a Business Model Definition Framework donde se demarca entre que conceptos se define un *Business Model*. Extraído de [Pateli and Giaglis, 2008]

- **Producto y/o servicio:** Corresponde a lo que entrega la empresa como tal, que ofrece a sus clientes, que valor entrega, etc.
- **Infraestructura:** Este aspecto detalla el cómo realiza sus procesos la empresa, en que contexto se enfocará, que tipo de tecnologías ocupa, que profesionales requiere en sus cargos, su tipo de estructura, etc.
- **Finanzas:** Concepto que aborda todo lo relacionado a los ingresos, gastos, inversión, capital, etc.

Estos componentes y sus diferencias entre modelos, permite definir 8 tipos de modelos de negocio que pueden optar las empresas [CERTUS, 2020], estos son:

- **Fabricante:** Una empresa fabricante se encarga de manipular materias primas y generar productos en gran masa, para luego venderlos a distribuidores o directamente a los clientes.
- **Distribuidor:** Siguiendo el punto anterior, una empresa distribuidora le compra los productos a una empresa fabricante con el fin de distribuirlos a los clientes o los negocios que lo vendan (retailers).
- **Retailer:** Como la anterior descripción indica, un *retailer* corresponde a un comercio minorista que adquiere los productos de distribuidores para luego venderlos directamente al público.
- **Franquicias:** Este modelo de negocio aborda los eslabones anteriores de manera total, es decir, pueden actuar como fabricante, distribuidor y retailer.
- **Ecommerce:** Son conocidos como comercios electrónicos, se enfocan en comprar y/o vender productos mediante plataformas o aplicaciones web en línea.

- **Freemium:** Uno de los modelos más comunes en internet. La empresa ofrece sus servicios más básicos con limitaciones de forma gratuita, para ofrecer un plan *premium* con todas las características complementarias.
- **Agregador:** Se trata de un modelo de negocio bastante reciente, donde la empresa agrega proveedores de servicio en un nicho y actúa como intermediario para venderlo con el uso de su propia marca.
- **Publicidad:** Este modelo de negocio se basa en el uso de publicada paga por patrocinadores en sus servicios para lograr ingresos. Actualmente se encuentra enfocado en la demanda de productos y servicios en internet, dado su exponencial crecimiento de usuarios en este medio.

2.1.1. Comercio electrónico (e-commerce)

La evolución del comercio electrónico desde sus inicios en Estados Unidos en los años 1920 surge debido a la venta por catálogo lo cuál experimentó grandes cambios y avances, incluyendo tecnologías de información y comunicación para el apoyo de procesos. En el año 1960, el nacimiento de una herramienta que permitiría la transacción de datos financieros de manera electrónica, el Electronic Data Interchange (EDI) [Cárdenas, 2020]. Entre sus numerosas ventajas se encuentra la capacidad de realizar compras y ventas en medios no físicos, mostrar catálogos a una mayor cantidad de personas y generar transacciones a nivel mundial y fundando mejoras a los modelos de negocio convencionales.

Dentro de este modelo de negocio, existen varias clasificaciones según los actores en la comunicación, los más reconocidos son: Business To Business (B2B), Business to Client (B2C) y Business to Employee (B2E), entre otros [L, 2021]. Estos modelos consideran el contexto, características y participantes en una relación comercial. Dado esto, los principales beneficiados son las empresas que realizan compra y venta de productos en el mercado, aumentando considerablemente su distribución, producción y venta de artículos a otras empresas o personas naturales.

2.1.2. Business to business (B2B)

Los orígenes del mercado *Business To Business* (B2B) empieza en los 1800 con la industria de revistas como “The furrow” por John Deere, el cual enseñaba técnicas de de agricultura a agricultores. Antes del internet, el mercado B2B consistía mayoritariamente en:

- Catálogos para realizar intercambios de productos
- Convenciones o exhibiciones, donde las empresas mostraban sus productos a otras empresas.
- Documentación escrita para realizar ventas
- Publicidad con el consumidor final.

Con la llegada de la plataformas digitales, el mercado B2B ha sufrido inmensos cambios, gracias a la rapidez de distribución de información, flexibilidad en métodos de compartir los productos y ampliación de información a ofrecer a posibles clientes. Principalmente se enfocó en:

- El principal medio para procesos formales, el email.
- Páginas web: medio fundamental para compartir información y catalogo de productos
- Redes sociales
- Publicidad digital
- Foros y otros medios digitales.

Con esto, varias empresas optaron a implementar un modelo de negocio B2B e-commerce, siendo la pionera en integrar este mercado la empresa *Wer Liefert Was* en el año 1996 [2.2](#). Luego de unos años, se funda *Alibaba.com*, el gigante chino que lidera en ventas principal plataforma web que actualmente es líder de ventas en las zonas oriente y mediterránea.

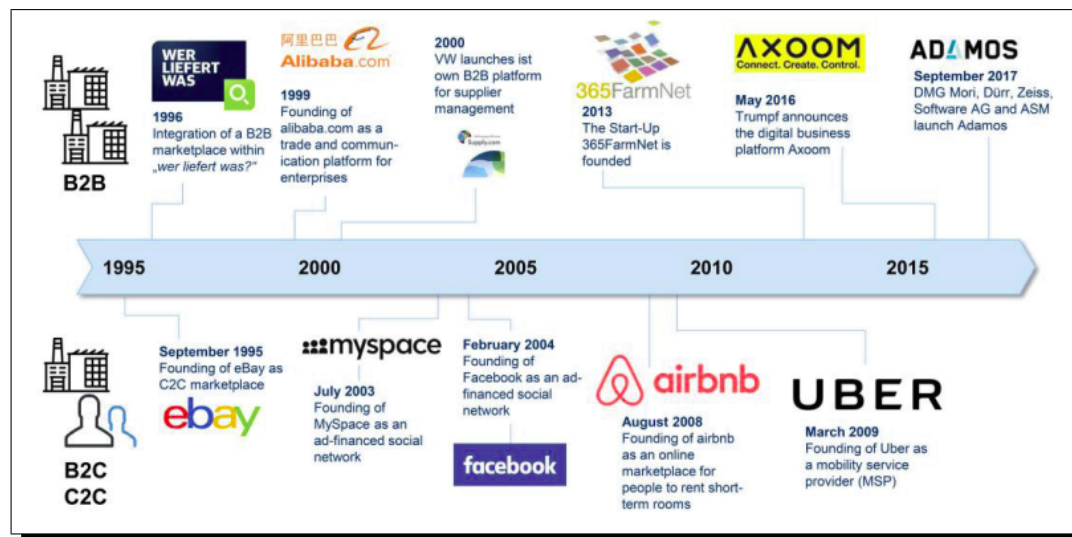


Figura 2.2 Visualización del recorrido histórico de plataformas e-commerce principales relacionadas al modelo de negocio *Business to Business* (B2B). Extraída de [\[Drewel and Gausemeier, 2018\]](#).

Desafíos en sistemas B2B

En el artículo [\[Fensel et al., 2001\]](#) hace referencias diferentes desafíos y tareas a tener en cuenta en el contexto de un comercio electrónico B2B (Business to Business), donde las empresas y su registro de datos no estandarizados, generan una fuente de información demasiado tosca y difícil de analizar, dado que, por lo general el catálogo de productos y semántica que descripción de cada producto va variando según cada empresa; además, que los proveedores que tienen su listado de productos en la plataforma B2B, son diseñados para lectura humana, es decir, imágenes, especificaciones u otra estructura de información dificultará en gran medida el análisis del producto y sus atributos. En este mismo documento, se detalla el uso de un motor para extracción de información de diferentes catálogos no estructurados, ocupando métodos no supervisados. Además, entra en uso de estándares validados, como lo es el UNSPSC (The Universal Standard Products and Services Classification) que es una clasificación jerárquica con 4 niveles y cada uno con valores numéricos y textuales que identifican entre sí;

siendo aun así, un proceso costoso y demandante, dado su necesidad de alta comprensión del contexto del problema.

2.2. Sistemas de información

Los sistemas de información son el resultado de la estrategia empresarial que se quiere implementar, la cual se aplica dentro de un contexto de modelo de negocio, que detalla como operará una empresa; con esto definido, las empresas van definiendo como serán sus procesos para obtener sus objetivos respectivos la cual confluye en procesos de negocio. Los sistemas de información corresponde al conjunto de estos conceptos, que recopila (o recupera), procesa, guarda y distribuye información para dar soporte a la toma de decisiones y control en una organización. Dentro de esta sección se analizará que tipos de sistemas de información se conocen y en cual de estos se clasificaría este trabajo.

Existen tres tipos de sistemas de información en base a los objetivos que cumplen, desde aspectos estratégicos, la toma de decisiones y sistemas que entregan apoyo de manera operativa. [OSI, 2018] Es una página detalla los tipos de sistemas de información generales, mencionando sus características, ventajas, desventajas, aplicaciones y ejemplos. Esta fue creada para ser usada en el curso *Organización y Sistemas de Información*, siendo los propios ayudantes quienes completaron la información y complementaron la información respectiva. El listado es gigante, por lo que a continuación se analizarán los Sistemas de información (SI) más populares.

- **Supply Chain Management (SCM)** corresponden a las herramientas y métodos que tienen por objetivos el mejoramiento y la automatización del abastecimiento, reduciendo los tiempos de entrega; se definen como la unificación de los procesos de negocio desde los proveedores hasta un usuario final.
- **Supplier Relationship Management (SRM)**, que es un enfoque sistemático para evaluar las contribuciones de los proveedores a tu negocio [Electrónica, 2021]. Realiza una segmentación de proveedores, clasificando por importancia, desempeño y consistencia, con el fin de darles prioridad; también un desarrollo de una estrategia de proveedores, donde se define un plan para trabajar con cada proveedor, garantizando relaciones exitosas, para aplicarlas en una futura relación con la otra empresa.
- **El sistema de soporte a decisiones (DSS)** son un concepto muy amplio de definir de manera concreta, pero una definición acertada es: "un sistema de información basado en un computador interactivo, flexible y adaptable, especialmente desarrollado para apoyar la solución de un problema de gestión no estructurado para mejorar la toma de decisiones"[OSI, 2018]. Las funciones que realiza este sistema nacen al ir compilando la información útil para apoyar al proceso de toma de decisión, se puede ocupar datos como el inventario de todos los bienes que se dispone, comparativa de ventas entre semanas, proyección de beneficios a futuro y evaluar consecuencias de diferentes decisiones. Existen varias clasificaciones de los DSS en base a diferentes criterios, uno de estos es la relación con el usuario, siendo DSS pasivo (ayuda en el proceso pero no lleva a cabo una decisión explícita), DSS activo (puede llevar a cabo una decisión explícita) y DSS cooperativo (permite la retroalimentación entre el encargado de la toma de decisiones y el

sistema). Este trabajo catalogaría de forma completa como un DSS, específicamente uno cooperativo, dado que posee procesos que brindan apoyo en la toma de decisiones (para elegir un proveedor) pero permite al usuario registrar una posterior retroalimentación al sistema como calificación de desempeño y en otras características.

2.3. Aplicaciones y módulos Web

Dentro del mercado actual existen varios sistemas informáticos basados en B2B que proponen abordar características tanto a proveedores como clientes. Algunos poseen módulos basados en la gestión de compras al proveedor, administración de stock, historial de compras y facturas, etc.

En este contexto, el software *Autoventa* [Autoventa, 2022] permite al vendedor en terreno tomar el pedido y enviarlo a la oficina para ser gestionado, mediante la creación de notas de venta, facturación y despacho de productos. Además, permite la integración de un portal B2B donde los clientes gestionan sus pedidos y hacen seguimiento a sus entregas; junto con un listado de clientes y fechas de entrega. En esta memoria, nosotros abordamos la forma en que el cliente realiza sus pedidos a varios proveedores, donde tendrá su propio historial de cuales han entregado un buen servicio; además el catálogo de proveedores estará disponible para todo cliente.

En un contexto de una empresa con mayor flujo de ventas, se encuentra *Mercado Libre*, la cual constituye una de las principales plataformas en el mercado, catalogadas *marketplace*, basadas en B2C actuales del mercado, y con presencia operativa en más de 15 países de América Latina fomentando la industria del comercio electrónico en Argentina, Brasil, Chile, Colombia, entre otros [Libre,]. Una de las características principales es la inmensa cantidad de vendedores que suben productos día a día, y su capacidad de actuar como mediador en casos de fraude o problemas. También, tiene la capacidad de calificar a todo vendedor de la plataforma y recomendar productos en base a sus preferencias. Otras empresas proveedoras también suben sus productos en esta plataforma para vender al por mayor. En el desarrollo de esta memoria nosotros proponemos el desarrollo de catálogo de productos y realizar recomendaciones al cliente usando una serie de atributos de calidad sobre estos (e.j., precio, eficiencia de entrega, etc.), los cuales el cliente definirá como prioritarias sobre otras, permitiendo una información más personalizada hacia el usuario y recomendaciones más satisfactorias.

Por otro lado, se encuentra el software *Gestioo* [Gestioo, 2022], que es una plataforma de gestión online para talleres, enfocado en empresas que realizan reparación y servicio técnico. Esta distribuido en base a módulos, cuales interactúan entre si y pueden obtener información de cada uno, algunos son de pago, por lo que hay suficiente independencia para funcionar de manera autónoma. Uno de sus módulos se llama: Administración de stock que permite actualizar el stock en función de facturas emitidas, gestionarlo, mover productos entre sucursales, seguimiento de stock y historial de cargas y de facturas.

Por supuesto, si se habla de plataformas o aplicaciones web B2B no se puede dejar de lado al consorcio *Alibaba*, líder mundial en ventas entre empresas en todo el mundo (ver Figura 2.2); que se distribuye en tres principales sitios: Taobao (C2C plataforma), Tmall (B2C plataforma) y Alibaba.com (B2B plataforma), cada una con millones de usuarios. La popularidad de Alibaba radica en la flexibilidad que entrega en subir catálogos de las empresas, teniendo un plan gra-

tuito que permite subir cierta cantidad de productos, pero con una comisión de recaudación por cada venta realizada; la seguridad que entrega en sus procesos, la capacidad de poder cerciorarse de los productos vía email con el proveedor, detalles de requisitos de compra (cantidad mínima, precio por ventas al mayor, etc.) [Zeibak, 2017].

Otra característica es la confidencialidad y seguridad que entrega a los vendedores de la plataforma, dado la gran masa de productos que se despachan por venta, resulta vital para los usuarios de Alibaba garantizar que la compra/venta se realice de manera adecuada, por ello, Alibaba cuenta con verificación de las actividades y datos de cada empresa, desde la verificación de las licencias que posean hasta el chequeo de las premisas de estas, la capacidad de pagar a terceros para que inspeccionen la empresa proveedora (Alibaba recolectaría los recibos y solo los hará efectivos si el resultado de la inspección es positivo; todo esto conlleva a clasificaciones o estatus de proveedor dentro de la plataforma, que se ve reflejado como gold member [Alibaba, 2022].

[Cardemil, 2022] En sus tesis *E-Business: Definición de la estrategia de internet, business, para una empresa editorial* detalla como es la adopción a un modelo de negocio e-commerce, específicamente *Business to Business* para la editorial *Grijalbo Mondadori*, la cual tiene comunicación frecuente con sus distribuidores, principales clientes de esta empresa. El trabajo se divide en 4 etapas: Etapa 0: Organización interna, Etapa 1: Presencia en la Web, Etapa 2: Orden de compra online, Etapa 3: E-commerce. Cabe destacar, que el trabajo especifica la planificación, trabajadores, perfiles del personal, etapas y requerimientos necesarios para la realización de la solución, que se especifica en el plan de acción.

En el capítulo 4 ahonda sobre las consideraciones a tener en cuenta antes de realizar el negocio online y como segunda parte, especifica los requerimientos de información y procesos para cada **componentes del negocio: Clientes (distribuidores), Proveedores y Autores**. Se analizará principalmente lo estudiado en esta última sección.

Dentro de los **requisitos de información** que se mencionan, estos son clasificados y distribuidos según la perspectiva de cada componente, siendo uno de los más importantes a tener en cuenta en la creación de un sistema o un módulo web para una empresa, es la información que maneja un proveedor. Estos son:

- Estadísticas de Ventas mensuales de los productos de su origen.
- Registro de pedidos pendientes: fecha de registro; detalle de los libros solicitados.
- Información de las devoluciones, indicando motivos, detalle de los productos devueltos y costos asociados.
- Información de libros autores nacionales y sus estadísticas de ventas.

En el contexto de este trabajo (módulo web) resulta vital entregar esta información en cada proceso de compra y venta que se realice, uno como vendedor de productos necesito saber cuantas ventas he realizado durante el mes, que productos manejan un mayor índice de ventas y que se me notifique algún problema que suceda con la entrega de productos (como devoluciones).

Los requerimientos de procesos actúan como base para un posterior desarrollo de la página web. Estos se indican en el siguiente listado:

- El distribuidor ingresa al sitio web mediante una clave que brinda la editorial, con el fin de garantizar quienes compren estén autorizados
- En el mismo sitio el distribuidor realiza la toma de pedidos, el cual es incorporado en los sistemas de la editorial.
- Cada orden de compra posee un número único que lo identifica.
- Definir diferentes estados posibles de la orden, con el fin de que el distribuidor pueda ocupar el número único y conocer en que etapa se encuentra su orden; saber si fue aprobado o despachado resultan fundamental para todo comprador.
- Dentro de la orden inicial, generalmente, pueden que ciertos artículos queden pendientes por falta de stock. En estos casos, se crea un pedido pendiente asociado a la orden de compra, con tal de notificar al distribuidor y despacharlo lo antes posible.
- Esta la opción de que los distribuidores ingresen en el sitio web la liquidación mensual de títulos, para realizar la facturación correspondiente.
- En este trabajo se define como la forma de pago, mediante facturas impresas, dado que la empresa ya trabajaba con este método, entonces la resistencia al cambio resultaría costosa.
- Los distribuidores tienen acceso a la información del inventario, es decir, el catálogo de la editorial, como: títulos existentes, stock por título, entre otros datos de interés (como los títulos próximos a tener en stock).

Dentro de los puntos anteriormente detallados, es importante recalcar los problemas que se presentaron al presentarle a la empresa métodos más actualizados y tecnologías digitales para sus procesos empresariales, existiendo mucha resistencia al cambio dado que, estos ya llevan años trabajando con el mismo sistema. Este mismo aspecto se aborda en este trabajo, estando 10 años más tarde, aún existen Pymes que sienten miedo y frustración de adoptar nuevas metodologías. Esto se menciona en la conclusión de [Cardemil, 2022], que la editorial ha logrado vislumbrar que en el futuro será necesario ir actualizándose hacia los beneficios que entrega internet y proyectar un desarrollo real del negocio en internet es de valorar, ya que la tendencia es mantenerse en los mismos esquemas. Además características como que las ordenes de compra resultan fundamentales en el proceso de compra y venta (para seguimiento, registro en base de datos, unicidad de elementos, etc).

En el capítulo 5 se define como será el proceso para la **construcción del sitio web**, donde se especifica los roles que se encargarán de la creación de la plataforma, en este caso es el *Proveedor de Servicio*.

2.4. Interfaz de programación de aplicaciones (API)

Las *Application Programming Interfaces (API)* son mecanismos que permiten a dos componentes de software comunicarse entre sí mediante un conjunto de definiciones y protocolos, actuando como un medio para una transferencia de datos. Siendo la definición etimológica, la palabra *application* se refiere a cualquier software con aplicación distinta, *interfaces* es un

contrato de servicio entre dos aplicaciones y *programming* al hecho que debe ser diseñada y desarrollada usando algún lenguaje de programación. La ventaja de ocupar este mecanismo, es que las entidades no necesitan profundizar el funcionamiento interno del servicio ni como esta implementada, ahorrando tiempos de desarrollo, simplifica el diseño y otorga flexibilidad de uso por otras aplicaciones, manteniendo la seguridad y el control [Hat, 2022].

Existen diferentes tipos de *API* que se explica en términos de cliente y servidor, siendo la aplicación que envía la solicitud llamado cliente y la que envía la respuesta se llama servidor, existiendo así 4 tipos de APIs:

- **API de SOAP:** Estas API utilizan el protocolo de simple acceso a objetos, intercambiando mensajes mediante XML entre el cliente y el servidor. Se trata de una API menos flexible con más popularidad en el pasado.
- **API de RPC:** Estas API se denominan llamadas a procedimientos remotos. El cliente completa una función (o procedimiento) en el servidor, y el servidor devuelve el resultado del cliente.
- **API de WebSocket:** La API WebSocket es otro desarrollo moderno de la API web que utiliza objetos JSON para transmitir datos. La API de WebSocket admite la comunicación bidireccional entre las aplicaciones cliente y el servidor, siendo capaz este último de enviar mensajes de devolución llamadas a los clientes conectados, siendo más eficiente que una API Rest.
- **API de Rest:** Estas son las API más populares y flexibles que se encuentra en la web actualmente. El cliente envía las solicitudes al servidor como datos, iniciando funciones internas dentro del servidor, devolviendo los datos de salida al cliente. Este tipo de API es la utilizada dentro de este trabajo.

2.4.1. API Rest

La arquitectura API REST empezó a desarrollarse en 1993, cuando surgieron los primeros sitios web de uso general. Tras seis años trabajando en la estandarización de HTTP y URI, el investigador Roy Fielding definió REST en 2000, estableciendo el estilo arquitectónico de los protocolos web. Con el tiempo, los conceptos REST se han adaptado al desarrollo de la informática, siendo desde el 2015 la iniciativa OpenAPI, bajo el patrocinio de la Fundación Linux y en colaboración con Google, IBM, Microsoft han creado la especificación OpenAPI basada en reglas Restful, que decreta como producir, consumir y visualizar servicios web RESTful [OVHcloud, 2024], siendo definido 6 reglas de diseño:

1. **Interfaz uniforme:** Todas las solicitudes deben respetar la norma de formato de dicha API, independiente del cliente que realice la solicitud, cada elemento debe cumplir dicho formato.
2. **Separación Cliente:** Las API REST requieren que las aplicaciones cliente y servidor sean independiente entre sí. El cliente solo tiene que conocer el nombre completo del recurso, siendo la única información que comparten es la información que se intercambia a través de las transacciones API.

3. **Sin estado:** Cada solicitud del cliente debe contener toda la información necesaria para procesarla, la cual el servidor no necesita conservar ninguna información sobre dicha solicitud; es decir, no existe el concepto de sesión y no tiene ningún estado ("stateless") con respecto a un cliente en particular.
4. **Puesta en caché:** Los recursos deben poder almacenarse en uno o más puntos en el cliente y el servidor. En el caso de este último, si se ha solicitado un recurso y es probable que esta vuelva a solicitarse, deberá almacenarse en caché para obtener una respuesta más rápida. A pesar de no ser necesario o especificado en el diseño de una API, representa una mejora destacable en rendimiento.
5. **Arquitectura de sistemas superpuestos:** Las consecuencias de la separación entre cliente y servidor, el diseño sin estado y el almacenamiento en caché no se puede saber si se está comunicando directamente con un servidor, por lo que aplicar agentes intermedios como un agente de servicios, load balancer, sistema de entrega de contenidos, proporciona una mayor flexibilidad y eficacia en las solicitudes.
6. **Código bajo demanda:** Cada vez es más común que el código que se entregue para ejecutarse con el cliente, como objetos Java o aplicaciones web JavaScript, siendo la restricción, que solo pueda ejecutarse previa petición del cliente.

2.5. Sistemas de recomendación

Los sistemas de recomendación aplican técnicas de análisis de datos con el objetivo de apoyar a los usuarios y clientes en encontrar que productos resultan mayormente favorables para ellos, en base a sus preferencias y de usuarios que comparten cercanos al objetivo [Sarwar et al., 2000]. Siendo el *análisis de decisiones* el estudio de decisiones que se toman en un ambiente de gran incertidumbre.

El libro [Aggarwal, 2016] funciona como gran apoyo para la creación de este sistema de recomendación, dada su completa información sobre los métodos, desafíos, ejemplos, ejercicios, consideraciones y más información que se debe tener en cuenta. La importancia de los sistemas de recomendación nace a la creciente del internet como el medio para transacciones, ya sean electrónicas o de empresas, el cual da espacio para que los usuarios puedan brindar su retroalimentación a los mismos procesos u transacciones en cuales participan; casos famosos como Netflix, actúan como punto inicial de referencia, dada su gran capacidad de poder recomendar productos (en este caso películas) a quienes ocupen la plataforma.

La calidad de la recomendación es fundamental para la satisfacción del usuario, dado que una mala recomendación genera frustración al cliente, por lo que, tener en cuenta las preferencias del usuario, evita una falsa sensación de ganancia o **falso negativo**; siendo el caso contrario, que se entregue una buena recomendación pero no es de preferencia del usuario. El principal caso a evitar, es el **falso positivo**, dado que la insatisfacción resulta mayor, debido a su propia preferencia y conocimiento. Estos casos se aprecia en el siguiente diagrama:



Figura 2.3 Diagrama Preferencia v/s Recomendación, donde se denota los 4 casos posibles resultados al entregar una recomendación. Creación propia utilizando información obtenida de [Sarwar et al., 2000].

2.5.1. Desafíos en sistemas de recomendación

Dentro del proceso de creación de un sistema de recomendación, existen varios procesos u etapas que representan un desafío que denota la eficacia y precisión del modelo. El principal problema es el *cold start*, que significa el lento inicio de estos servicios, debido a la poca cantidad de información adicional, generando que los cálculos de similitud y estimaciones se vean afectadas en su precisión. Otro aspecto es la naturaleza de la información, como la dispersidad de los datos, la dimensionalidad, el tipo de dato, la cantidad, etc. Por ello, resulta importante medir el porcentaje de densidad de una matriz, con esta fórmula:

$$Densidad = \frac{Nmerodeelementosnulos}{Totaldeelementosenlamatriz} \cdot 100 \quad (2.1)$$

El problema de la densidad o dispersión de la matriz es una tarea a tener en cuenta para el tratamiento de los datos resultantes, cuando los datos están mas compactados resulta mucho más fácil y rápido aplicar mejores algoritmos; sin embargo, hay que tener cuidado como se realiza esta compactación o reducción de dimensiones, dado que inherentemente se perderá información, siendo vital analizar y comprender cual es la información que más aporta al contexto del problema. Para esto, existen diferentes métodos que pueden ayudar a cumplir el objetivo:

- **Muestreo de datos:** A través del análisis de los datos para seleccionar un sub conjunto que sea más compacto en proporción al original, siendo muy difícil de aplicar en sistemas dinámicos o de actualización continua.

- **Selección de características:** Solamente ciertas características es usado para el análisis de datos, reduciendo su dimensionalidad pero perdiendo información en el proceso.
- **Reducción de datos a través de rotación de ejes:** En este método se analiza la correlación entre los atributos para representar un número menor de dimensiones. Ejemplos de estas técnicas son: *principal component analysis* (PCA), *singular value decomposition* (SVD) o *latent semantic analysis* (LSA) para contexto de textos.
- **Reducción de datos a través de transformación de tipo:** Este tipo de reducción hace referencia a la portabilidad de los datos, representando la información con otra estructura. Por ejemplo, los grafos se pueden representar en representaciones multidimensionales.
- **Trade-off o balance entre cobertura y precisión:** Dentro de los sistemas de recomendación, se evalúa la precisión que entrega la recomendación para el usuario, siendo en primera instancia el objetivo de buscar la mayor precisión posible, sin embargo, el obtener una mayor precisión disminuye la cantidad de cobertura o diversificación que puede entregar el modelo, significando una mayor probabilidad de sobre ajuste (overfitting).

Principal Components Analysis

Principal Components Analysis o PCA es un método de análisis multivariante que transforma un conjunto de datos con múltiples variables correlacionadas en un nuevo conjunto de variables no correlacionadas llamadas componentes principales, siendo ordenados de tal manera que el primer componente principal es el que captura la mayor variabilidad de datos, el segundo captura la siguiente mayor y así sucesivamente. El objetivo de aplicar un PCA es la reducción de dimensionalidad de ciertos atributos del conjunto de datos, con el fin de priorizar y analizar solamente aquellas dimensiones o características que tengan mayor relevancia en el cálculo, con el fin de mejorar el desempeño del modelo por una leve disminución en la precisión de los resultados. Se suele ocupar para un análisis exploratorio, reducción de ruido, compresión de datos y selección de características [Aggarwal, 2015].

Singular Value Decomposition

Singular value decomposition (SVD) es un método de factorización basado en matrices, que separa la información en 3 matrices, el que no necesita calcular la covarianza de las matrices, permitiendo trabajar con valores nulos o perdidos, sin embargo, por lo general resulta más costosa en términos de cómputo para los sistemas [Aggarwal, 2015].

Expresiones de similitud

El uso de diferentes expresiones de similitud permite la cuantificación de que tan diferente o alejado se encuentra un ítem con otro, usando sus propias características o información anexa a cada uno de estos. Dentro de un sistema de recomendación la opción más común es usar una similitud basa en coseno, que calcula el ángulo entre dos documentos o productos, siendo $\bar{X}$ y $\bar{Y}$ los dos productos, siendo x_1, y_1 las características asociadas a cada uno y d la cantidad de columnas o largo de información de ambos productos. Entonces, la similitud de coseno entre $\bar{X}$ y $\bar{Y}$ se define como [Aggarwal, 2015]:

$$\cos(\bar{X}, \bar{Y}) = \frac{\sum_{i=1}^d x_i \cdot y_i}{\sqrt{\sum_{i=1}^d x_i^2} \cdot \sqrt{\sum_{i=1}^d y_i^2}} \quad (2.2)$$

Normalización de datos

La normalización de datos es una técnica de pre-procesamiento de datos común en trabajos de estadísticas, que se utiliza para ajustar características o atributos de los datos dentro de un rango determinado, permitiendo así una mejor interpretación, además de mejorar la eficiencia y precisión del algoritmo o sistema integrado. Dentro de las técnicas mayormente usadas, se encuentra la normalización L1 y L2, que tienen diferentes aplicaciones y alcances (obtenida desde [Aggarwal, 2015]).

La normalización L1 o regularización de Lasso, permite eliminar la influencia de valores atípicos (outliers), agregando una penalización equivalente a la suma de los valores absolutos de los parámetros (distancia Manhattan), siendo x_i la característica individual del vector, el cual tiene hasta n características el vector:

$$||x||_1 = \sum_{i=1}^n |x_i|, \quad \forall i = 1, \dots, n \quad (2.3)$$

Por otra parte, la normalización L2 o regularización de Ridge, provoca que la influencia de los valores atípicos (outliers) sea más notable, aplicando una penalización de los cuadrados de los parámetros (distancia euclideana):

$$||x||_2 = \sqrt{\sum_{i=1}^n |x_i|^2}, \quad \forall i = 1, \dots, n \quad (2.4)$$

Además, existe una normalización comúnmente utilizada para llevar el conjunto a un rango entre 0 y 1, tomando en cuenta el máximo y mínimo valor dentro del listado de resultados. Esta es la normalización: Max-Min, que aplica la siguiente Ecuación:

$$z_i = \frac{x_i - \min(x)}{\max(x) - \min(x)} \quad (2.5)$$

2.5.2. Tipos de datos

En la investigación presentada por [Wei et al., 2007] se realiza en un estudio de múltiples técnicas en sistemas de recomendación en el contexto de e-commerce. Los autores mencionan lo vital y difícil de definir las preferencias que tiene el cliente y por ende su criterio, siendo un desafío importante precisar que aspectos del usuario serán analizados y factores a la hora de decidir y recomendar. Sistemas recomendadores como *MovieLens*, *FAB* y *The Entree Restaurant* toma varios datos del cliente, confluyendo una clasificación en varias clases:

- **Información del cliente:** Estos datos corresponden a atributos y/o características que posea el cliente u empresa; por ejemplo: Información demográfica.

- **Calificación a productos:** Calificación que le da los usuarios a los productos, ya sea cuantitativo o cualitativo. Este último, actualmente, mediante web mining se puede obtener una calificación implícita según la navegación en el internet.
- **Información de producción:** Hace referencia a la información de los productos pertenecientes al contexto, dando como ejemplo *MovieLens* (que se basa en la recomendación de películas) ocupa la información de los actores, productor, director/a, duración, género, etc.

En [Aggarwal, 2016] se detallan dos tipos de información o datos que trabaja un sistema de información, los cuales definen el tipo de modelo u contexto que se desarrolla el sistema de recomendación:

- **La interacciones usuarios y productos**, tal como la calificación o el registro de compras. Siendo esto ocupados por modelos basados en filtrado colaborativo
- **La información de las características o atributos sobre los usuarios o los productos**, tales como su descripción, etiquetas o palabras claves. Siendo ocupados por modelos basado en el contenido.

2.5.3. Modelos en sistemas de recomendación

Dentro de los enfoques posibles que se pueden ocupar dentro de un sistema de recomendación, hay distintos modelos que son populares y que definen como será la estructura, la naturaleza de la información a usar para entregar recomendación, la interacción entre los componentes y, la precisión y como afronta los desafíos que demanda crear un sistema de recomendación. Esto conlleva a la definición de varios tipos de modelos que pueden conformar un sistema de recomendación, los cuales serían:

- Modelos de filtrado colaborativo
- Sistemas de recomendación basados en contenido
- Basado en el conocimiento
- Basado en la demografía
- Modelos híbridos

Modelos de filtrado colaborativo

Los modelos de filtrado colaborativo ocupan los datos de calificaciones o *ratings* que le brindan los usuarios a los ítem para así poder realizar una recomendación, representándolos a través de una matriz, la cual, las filas representarían los usuarios y las columnas a los ítem, siendo esta compuesta por las calificaciones anteriormente mencionadas. El principal desafío como se afirma en [Aggarwal, 2016] es que tan dispersa esta la matriz, es decir, que tanta información o datos faltan dentro de esta misma; esto genera que las recomendaciones a través de un filtrado colaborativo se sobre ajusten a recomendar solamente productos que si tengan información, dado que, los ítem que posean poca información, no habrá forma de realizar una conexión significativa a otro usuario u otro productos.

Hay dos tipos de métodos comúnmente ocupados en filtrado colaborativo:

- **Métodos basados en memoria o heurística:** Estos algoritmos, también referidos como algoritmos de filtrado según los vecinos, en cuales los ratings entre los usuarios y los ítems son predichas basándose en el ratings de los vecinos mas cercanos a ellos. En esta sección, se encuentra los dos enfoques: **User-based collaborative filtering** y **Item-based collaborative filtering**. Las ventajas de este enfoque es que resultan fácil de aplicar y de explicar, sin embargo, no funcionan del todo bien con matrices dispersas, dado que no se podrían encontrar suficientes usuarios similares entre si o ítem
- **Métodos basados en modelos:** Hace referencia a los métodos de predicción, usando técnicas de machine learning y minería de datos, el cual se define un modelo separando el conjunto de datos en fracciones, para así realizar un proceso de entrenamiento y validación. Estos se pueden apreciar en la Tabla 2.1.

Colaborativo	Técnicas	Artículos típicos
Basado en la heurística método	Algoritmo KNN y el mejorado	Resnick, 1994; Mohan, et al 2006
	Minería Web	Choa, et al, 2004; Cho, et al, 2002
	Árbol de decisión	Cho, et al 2002
	Teoría de grafos	Aggarwal, 1999
	Vector de apoyo máquina	<i>Min & Han, 2005</i>
Basado en modelos método	Modelo Bayesiano	Breese, et al, 1998; Chien, et al, 1999
	Agrupación	Breese, et al, 1998; Goldherg, et al, 2001
	Reglas de asociación minería	Kim, et al, 2004
	Redes neuronales artificiales	Pazzani, et al, 1997
	Regresión lineal	Vucetic, et al, 2005
	Entropía máxima	Pavkov, et al, 2002
	Análisis semántico latente	Hoffman, 2004; Cheung, et al, 2004
	Proceso Markov	Shani, et al, 2002

Tabla 2.1 Una taxonomía de enfoques de filtrados colaborativo. Extraída de: [Wei et al., 2007]

Tipos de calificación o ratings

En [Aggarwal, 2016] hace referencia a los tipos de calificación o ratings y como se representan, siendo lo común que sean basado en intervalos, en cual se destaca el rango o jerarquía del rating sobre otro; también se pueden ocupar variables categóricas de tipo ordinal, cuales comprenden a un texto que denota la calificación brindada por los usuarios pero que su in-

interpretación jerárquica no se acerca a un número claro, es decir, la semántica definida por el sistema define su valor, como se explica a continuación:

- **Calificación explícita (explicit ratings):** Se refiere a aquellas calificaciones en donde el usuario brinda una calificación directa a un producto, demostrando su contento o agrado. Para sistemas de calificación unitaria, la calificación explícita resulta difícil de obtener, dado que al solamente poder expresar su contento a ciertos productos, esto no significa que los demás le desagraden, generando una matriz mas dispersa y menos precisión en similitud
- **Calificación implícita (implicit ratings):** Se refiere a aquellas calificaciones en donde el usuario no brinda una calificación directa a ciertos productos, más bien denota interés con otros métodos, ya sea con su registro de compras de ciertos productos, el tiempo que estuvo en la pantalla del producto, si lo agrego al carrito, cuantas veces apretó el link que llevaba al item, etc. Siendo interacciones en donde el usuario no brinda de manera consciente su nivel calificación al producto.

[Zhang et al., 2008] plantea un modelo basado en el rating en un contexto de comercio electrónico B2B, ocupando varios atributos para determinar la reputación de la empresa proveedora, tales como: calidad del producto y servicio, seguridad, compatibilidad del producto con la descripción, precio, desempeño y eficiencia en la entrega, rapidez y conveniencia en retorno de productos o dinero en caso de problemas, protección de la privacidad del cliente y seguridad del pago de este último.

Modelos basados en contenido

En este modelo de sistema de recomendación, la descripción de los productos o de los usuarios son usados para realizar una recomendación, sin embargo, también se ocupa la información definida en el anterior modelo. Esto sucede para casos en cuales, el sistema no cuenta con la información de otros usuarios pero si con mucha información precisa de los productos a ofrecer, para estos casos, el filtrado colaborativo no se podrá aplicar (dada su necesidad de contar con información de otros usuarios), sin embargo, si el usuario cuenta con un registro de compras o de calificaciones a ciertos productos, se puede ocupar su descripción o palabras claves para determinar la similitud a otros productos. Estas calificaciones textuales o de documentos, son usadas para crear modelos de clasificación o regresión, la cual es **específica para el usuario**. Sin embargo, este modelo cuenta con sus desventajas:

- En muchos casos se entregarán recomendaciones obvias al usuario dado el uso de palabras claves y/o contenido, esto debido a que el modelo es construido específicamente para el usuario, disminuyendo la diversidad de productos a recomendar.
- Efectiva para entregar recomendaciones para nuevos productos que se registren, dado que si cuentan con suficiente información asociada, se podrá calcular de manera eficiente la similitud con otros productos; sin embargo, para nuevos usuarios, el modelo palidece por la misma razón brindada en anterior punto.

Para aplicar este modelo en un sistema de recomendación, es importante tener en cuenta los dos tipos de información necesarias: Información detallada sobre varios productos, esta

puede ser la descripción asociada a cada uno, su valorización, descripción de quien lo provee, etc. Por otro lado, es tener información suficiente del usuario, para crear un **perfil de usuario**, esto puede ser usando datos implícitos o explícitos. Para ello, se definen varias etapas para entrega una recomendación:

- **Preprocesamiento e ingeniería de características:** Dado el uso de información textual, este modelo puede enfrentarse varios contextos de problemas e información, variando entre semánticas y sintaxis de las mismas palabras. Para esto, se necesita definir que tipo de variable corresponde cada característica (numérica, de texto, categorica y ordinal), con el fin de aplicarle técnicas de pre procesamiento (como se detallan en [Aggarwal, 2015]).
- **Definición del perfil del usuario:** En este punto, resulta importante definir cual será la información del usuario que será usada para su recomendación, si será su registro histórico de compras o su calificación a ciertos productos, el modelo resultado representará las tendencias del usuario.
- **Filtrar y recomendar:** Ya en este punto y con los anteriores ya realizados, el modelo ya es capaz de poder entregar productos al usuario basándose en su perfil y la similitud de este a los demás productos.

Al trabajar con información y datos que no cuenta con una estructura definida (como lo son las numéricas), resulta importante definir como se realizará la fase de extracción de características de los datos, es decir, como determinar que un producto es similar a otro en base a sus atributos textuales. Para ello, en [Aggarwal, 2016] se define los pasos necesarios para aquello:

- **Remover palabras vacías (Stop words):** Muchos textos que son obtenidas de descripciones o fuentes similares, traerá palabras que no aportan en la descripción del producto o su caracterización, aquellas que se ocupan para unir palabras, como: “”, “.además”, etc. No se ocupan y solamente sirven como ruido para el análisis, por ello, resulta importante primero el filtrar estas palabras de los datos.
- **Eliminar los aditivos o extensiones de la palabra (Stemming):** Esto significa que variaciones de la misma palabra son consideradas como iguales, por ejemplo, palabras que aparecen de forma singular o plural en diferentes textos se deberán tener en cuenta como iguales.
- **Extracción de frases:** En este punto, la idea es detectar palabras que ocurran o aparezcan de manera simultánea en diferentes documentos.

Luego de haber aplicado todo estos pasos, las palabras pueden ser representadas como un vector en un espacio, siendo cada palabra referida como *term*. Resulta común pensar en primera instancia en solamente calcular la frecuencia total de apariciones de las palabras en todos los productos y desde ahí estimar similitud, sin embargo, esto generaría un sesgo no deseado. Por ejemplo, si ‘contenedor’ aparece 1000 veces en un solo producto y no aparece en otros, esto generará que productos que no contengan esta palabra no sean similares. Para esto, se ocupa la métrica de: **Frecuencia inversa por documento** o en inglés *inverse document frequency*, que mide la frecuencia de apariciones de la misma palabra pero en cuantos documentos o productos diferentes.

$$id_i = \log\left(\frac{n}{n_i}\right) \quad (2.6)$$

Siendo id_i la frecuencia inversa del i término y n_i la cantidad de documento que contienen el i término; el valor de n representa al valor total de documentos o productos. Sin embargo, de manera opcional, [Aggarwal, 2016] detalla que puede existir una excesiva frecuencia de una sola palabra en todos los documentos, lo que genera una menor importancia como representación del producto, por lo que se puede aplicar funciones de normalización o estandarización para estas palabras, como de logaritmo o de raíz cuadrada, siendo x_i la frecuencia total que aparece el término i :

$$f(x_i) = \sqrt{x_i} \quad (2.7)$$

$$f(x_i) = \log(x_i) \quad (2.8)$$

La unión de estas dos funciones nos entrega una frecuencia normalizada $h(x_i)$ para cada i palabra, siendo un modelo muy popular llamado: **tf-idf model** (*term frequency inverse document frequency model*):

$$h(x_i) = f(x_i) \cdot id_i \quad (2.9)$$

Esta técnica se puede apreciar en la siguiente figura extraída de [Villalonga, 2022], artículo en donde se ejemplifica como es el proceso de obtención de palabras, caracterización y asignación de pesos dentro del documento según la cantidad de apariciones y cantidad de palabras del documento.

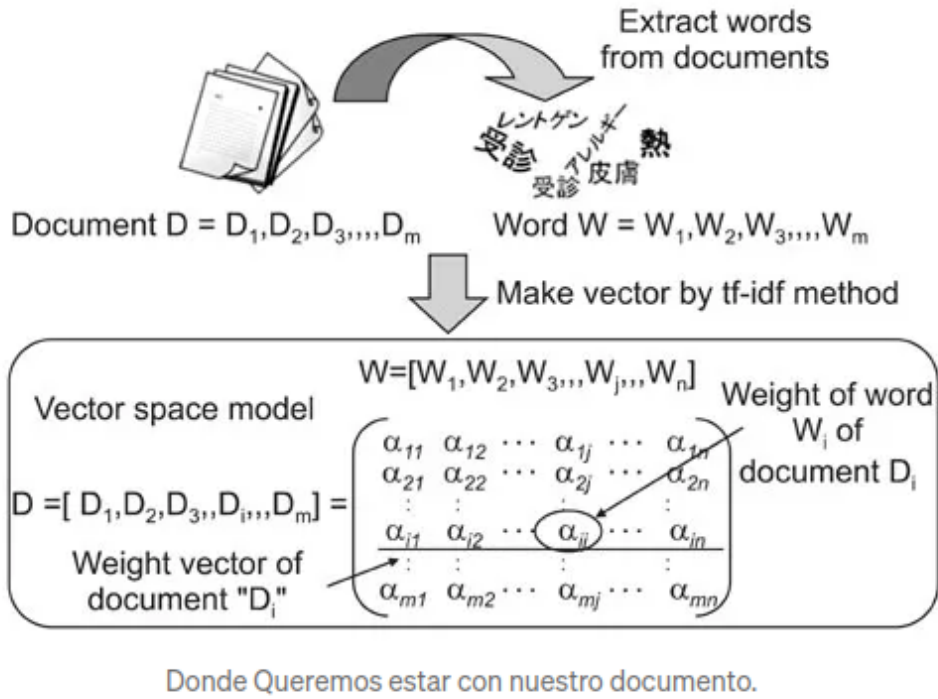


Figura 2.4 Flujo Vector TF-IDF extraída de [Villalonga, 2022]

Un problema que puede surgir al momento de analizar texto, como se nombra en [Honrado et al., 2000] es la inflección o la cantidad de significados que puede tener una misma palabra pero en diferentes contexto. Esto sucede especialmente en los lenguajes románticos, siendo el español parte de ellos que dificulta en identificar palabras que tienen el mismo origen pero su significado en la oración dificulta el análisis. Para este caso, se ocupara el algoritmo brindado en *snowball*, un paquete de python con soporte a español.

Por otro lado, en muchos escenarios los atributos numéricos pueden estar representados en diferentes escalas de referencia, lo que genera que la comparación entre estos atributos resulte difícil y sesgada, por ejemplo, la escala de la edad es diferente a la escala que se maneja el sueldo, resultando que al ocupar funciones de comparación como la distancia euclidiana, dominará los atributos que tengan un mayor valor. Para solucionar este problema, comúnmente se ocupa la estandarización. Considerando el caso donde el atributo j tiene un promedio de u_j y una desviación estándar σ_j ; el valor del atributo j del producto i x_i^j será normalizado como:

$$z_i^j = \frac{x_i^j - u_j}{\sigma_j} \quad (2.10)$$

Modelos basado en conocimiento

Un sistema de recomendación basándose en el conocimiento, hace referencia a contextos donde los productos no suelen comprarse o no existe un registro formal del todo detallado entre interacciones entre usuarios y productos, siendo difícil tener un número suficiente de información para entregar una recomendación detallada. Para este caso, el usuario detalla sus

requerimientos específicos sobre sus preferencias (ya sea con palabras claves o descripciones) lo que permite identificar reglas o patrones del perfil del usuario (especificar explícitamente que es lo que quiere), con lo que, se podrá estimar la similitud de este perfil con los demás productos.

2.5.4. Evaluación del sistema de recomendación

La evaluación de un sistema de recomendación resulta un proceso vital para identificar el desempeño y la veracidad de las recomendaciones obtenidas, este concepto se denomina como utilidad.^{en} [Silveira et al., 2019], que hace referencia al grado de utilidad o satisfacción que muestra el usuario sobre las recomendaciones entregadas, que depende del orden de la preferencia de este. En [Silveira et al., 2019] basan su definición en como el usuario reacciona a las recomendaciones y si le genera valor. Para ello, se detallan diferentes métricas que sirve para determinar esta utilidad:

- Métricas de error: Estas métricas apuntan a valores comúnmente usados en estadísticas, como *Mean Absolute Error* (Error Absoluto Medio) que calcula la diferencia entre las calificaciones predichas por el sistema y la brindada por el usuario. Además de la métricas *Root Mean Squared Error* (Distancia cuadrática mínima media), que calcula una mayor diferencia del error obtenida.
- Precision: Esta métrica hace referencia a la relevancia que le da al usuario al listado de recomendaciones que entrega el sistema, ya sea mediante una calificación, una acción o mostrar un interés directo al ítem presentado. Siendo C_u el conjunto de productos que el usuario considera relevante y R_u el conjunto de productos que el sistema entrega como recomendación.

$$precision = \frac{|C_u \cap R_u|}{|R_u|} \quad (2.11)$$

- Recall: Métrica que calcula el número de productos o ítem de la lista de recomendados, que son parte del conjunto de productos que el usuario considera relevantes, ya sea por un registro inicial de calificaciones o un registro de compras.

$$recall = \frac{|C_u \cap R_u|}{|C_u|} \quad (2.12)$$

- Diversidad: La diversidad es un concepto objeto de varios estudios de definición, sin embargo, en base al trabajo ya mencionado, se define como una métrica que tiene un efecto contrario a la similitud, en donde, recomendaciones con una baja diversidad no serán de interés para el usuario. Como resultado, la diversidad se obtiene calculando la diferencia de que tan distintos resultan los productos a través de la distancia $d(i, j)$, siendo un producto i y un producto j , ambas del conjunto de recomendaciones R_u .

$$diversidad = \sum_{i \in R_u} \sum_{j \in R_u, i \neq j} d(i, j) \quad (2.13)$$

2.5.5. Toma de decisiones en un e-commerce

Dentro del contexto de una empresa, que implementa un modelo de negocio de comercio electrónico (e-commerce), la toma de decisiones resulta fundamental para obtener los beneficios que ellos estimen más importantes, ya sea: ganancias, confianza de sus clientes, rapidez de entrega o calidad del producto. Todos estos aspectos, cumplen un rol determinante al momento de seleccionar el criterio para la acción a realizar y de que forma estimarlo. Por ello, mediante la evolución histórica de estos modelos de negocio, se han realizado experimentos, sistemas y aplicaciones que brindan un apoyo en la toma de decisiones a los empresarios.

En todo este proceso de compra y venta, intervienen una serie de etapas lineales, con contexto temporal, que comienza cuando la persona/empresa (cliente) detecta una necesidad hasta el momento que se adquiere el producto o servicio, siendo conocido como *customer journey* según la investigación propuesta por [Richardson and Innovation, 2010]. Las etapas (ver Figura 2.5) se clasifican en los siguiente puntos:

- **Descubrimiento:** En esta etapa el cliente reconoce la necesidad debido a un problema que se encuentra, siendo su opción más segura, buscar una solución en internet.
- **Consideración:** En esta etapa, el cliente contempla las distintas opciones, lo que lleva a una búsqueda por saber cual es la mejor solución para el problema.
- **Conversión:** La tercera etapa corresponde a la búsqueda de alternativas que den el mismo tipo de solución, para llevar una comparativa que ayude a su decisión.
- **Retención:** Esta etapa corresponde al proceso de obtención del producto y/o servicio, percibiendo si brindó un resultado positivo o no para el.
- **Recomendación:** La quinta y última etapa se basa en la post venta. Obteniendo feedbacks del cliente y realizar seguimiento, se vuelve fundamental para mantener al cliente. Si es positivo, se puede convertir en un cliente habitual y recomendar el producto y/o servicio a otros posibles clientes.

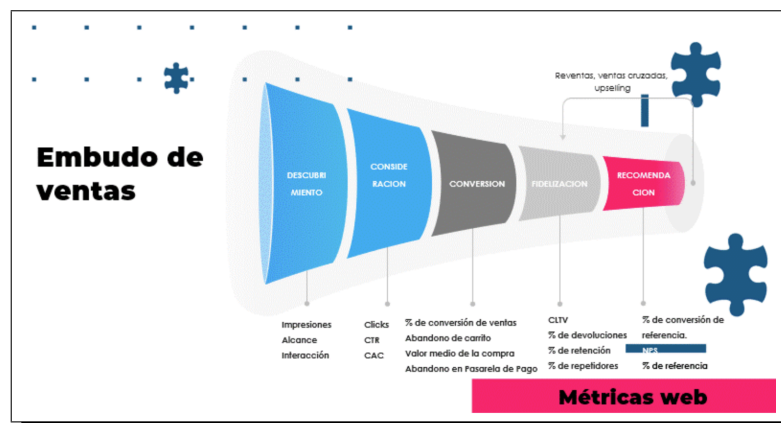


Figura 2.5 Pasos por el cual pasa un comprador para la selección de un producto, visualizado como un embudo, dado que el primer paso es captar la mayor cantidad de personas. Imagen extraída de [Conecta, 2022b].

Para poder definir un buen customer journey y por lo tanto, una mejora de los procesos de cada etapa, es necesario analizar y estudiar los datos que nos aporta el comportamiento de otros usuarios, que serán de utilidad para tomar una mejor decisión. Las métricas que nos permite cuantificar el desempeño de la estrategia utilizada, se denomina *Key Performance Indicator* (KPI), que pueden ser datos como: la satisfacción del cliente, tiempos de entrega, ventas efectivas, productos en mal estado, etc. Determinar el avance del sistema ecommerce ocupando estas métricas resulta fundamental para visualizar los beneficios de integrar un nuevo módulo o sistema a la estructura actual. Los compradores online tienen en cuenta además más información que influye en su criterio de decisión, como lo son [Conecta, 2022a]:

- . **Gastos de envío y devoluciones:** En este contexto, cada comprador (ya sea empresa o persona natural) tendrá en mente el costo de envío asociado a la compra de un producto y/o servicio. Además, el proceso de devolución en caso de falla resulta fundamental para conservar a los clientes. Mientras más simple sea el proceso, menos errores hará el usuario.
- . **Opiniones y reviews de otros clientes:** La información brindada por otros usuarios resulta fundamental para influir en la decisión de un comprador. A mayor cantidad de reseñas entonces existirán más criterios y perspectivas para poder tomar una adecuada decisión.
- . **Sencillez en el proceso de compra:** Los usuarios suelen optar por procesos simples y entendibles, con el fin de evitar confusiones y frustración. Un ejemplo de este tópico, es la técnica *one-click buying*, que permite a los compradores realizar una orden de compra sin ocupar el carrito de compra.
- . **Calidad del producto:** Este aspecto resulta fundamental para el comprador, en dependencia de la calidad del producto entonces se determinará si será comprado o no, y además si el cliente se convierte habitual.
- . **Opciones de pago:** Múltiples opciones de pago siempre son bienvenidos, cada cliente tiene una seleccionada en base a sus intereses y/o necesidades. En este trabajo, este aspecto fue abordado en la Sección 2.4.

Capítulo 3

Desarrollo del módulo

La creación de un sistema o un módulo comienza con la definición y decisión de la arquitectura que tendrá el software, en este caso, uno basado en una aplicación web de punto y venta desarrollado con anterioridad, el cual cuenta con un sistema de recomendación integrado. Para ello, resulta fundamental identificar los requisitos, datos, información previa, modelo de negocio, desafíos, tecnologías, tipos de API, diseño del modelo de dominio, tipo de base de datos, infraestructura y el cómo se comunicará cada uno de estos componentes para que el usuario pueda visualizar sus productos y los que el sistema le ofrece. Para ello, resulta necesario la creación de una interfaz web que pueda consumir la información alejada en el servidor con el objetivo de mostrarla de manera intuitiva y que el usuario pueda comprenderla (así usar efectivamente el módulo), esto se conecta con la API del servidor (o backend), el cual tendrá las conexiones a la base de datos con toda la información respectiva; todo esto acompañado con el sistema de recomendación que entregará la información personalizada al usuario.

Para ello, antes de comenzar con el desarrollo, es importante tener en cuentas consideraciones previas, tales como:

- **Modelo de negocio de la empresa:** Resulta fundamental definir y comprender el modelo de negocio de CGTEC, para justificar el uso de ciertas plataformas y como fue estructurada la información de los productos, pymes y proveedores.
- **Estructura de almacenamiento de datos:** Esto se acompaña a la definición del modelo de dominio y el diagrama de entidades necesaria para definir el modelo de base de datos.
- **Obtención de información a través de una API:** Se obtendrá la información de los productos del catálogo a través de los servicios disponibles de Mercado Libre, dada su gran disponibilidad y flexibilidad información que acompaña a cada producto registrado.
- **Ambiente de tecnologías:** Se justificará el uso de ciertas tecnologías y la razón de aplicación para este contexto, específicamente en los lenguajes de programación usados y en como fue desplegado en diferentes servicios de producción.

3.1. Consideraciones previas

Dentro del desarrollo de la interfaz, se enfrentaron varios obstáculos, aspectos como la coordinación de la empresa/cliente que usará el sistema que lleva a una adaptación de las tecno-

logías a utilizar, representando así, nuevos desafíos en el desarrollo del submódulo. Con esta sección se busca detallar y argumentar las razones de diseño del sistema, con el fin de resulte entendible la arquitectura.

3.1.1. Modelo de negocio de CGTEC

CGTEC es una empresa de soluciones tecnológicas con la misión de tecnologizar el la provincia del Choapa (cuarta región), entregando soporte técnico y asesorías informáticas para negocios o el hogar. Sus servicios corresponden desde soporte técnico en productos como : celulares, impresoras, computadores, parlantes y tablets, entre otros, por lo que, su gama de registros de compras son los repuestos para estos equipos, los cuales obtienen y compran según la demanda, disminuyendo el tiempo de stock. Esto resulta de mucha importancia para poder determinar cuando se necesitaría obtener un nuevo repuesto, dado que esto dependerá netamente del usuario o cliente (y el equipo).

Otro aspecto a considerar es la forma de como guardan los registros de compras. Este se realiza a través de una única plataforma que les entrega un catalogo actualizado, en este caso Mercado Libre, en donde compran los repuestos que necesitan. Cabe destacar, que la empresa no maneja un registro formal ocupando una herramienta como excel u otro sistema de registro, sino que, sencillamente, sus compras están guardadas en las ordenes de compras a la cuenta asociada de la empresa en Mercado Libre, lo que representa un problema en la obtención de la información asociada.; pero también muchas ventajas en el momento de obtener información de los proveedores y los productos que son parte de la compra. Para ello, resultó necesario ocupar los servicios de desarrollador que entrega mercado libre, que se explicarán en la siguiente sección.

3.1.2. Servicios de Mercado Libre

Para obtener la información adjunta de cada producto, resulta necesario ocupar los servicios de API que ofrecen Mercado Libre, por lo que resulta necesario analizar qué es esta empresa y de que forma consumir y obtener su información. Mercado libre es una empresa multinacional dedicada al comercio electrónico, en cuales los usuarios pueden comprar, vender productos tanto nuevo como usados, la cual se caracteriza por su capacidad de recibir calificación por vendedor, comprador y producto de manera individual; esto se acompaña con un excelente acompañamiento en envío, recepción y post compra que brinda a cada uno de sus usuarios; si bien esta plataforma se diseño para ser de carácter B2C (Businnes to Client), muchas empresas ocupan los servicios para encontrar productos que favorezcan su negocio (como es el caso de CGTEC). Esta plataforma cuenta con un soporte para desarrolladores mediante una API (Application Programming Interface) para que aquellos usuarios experimentados puedan crear aplicaciones o plataformas ocupando los servicios e información pública disponible [[MercadoLibre, 2023](#)]. Para ello, se creó una aplicación en esta plataforma que permitió obtener información variada de los productos, vendedores y usuarios, como se puede apreciar en la siguiente lista y en la Figura 3.1:

- **Información personal del usuario:** Nombre, dirección, rut, etc.

- **Registro de compras del usuario:** Se puede obtener el registro o movimientos históricos desde hace un año, el cual entrega el vendedor, producto, retroalimentación y otra información asociada a esta compra.
- **Información acerca del vendedor:** Para este contexto, resulta vital obtener información asociada al vendedor (proveedor), como su reputación, cantidad de transacciones, catálogo completo de sus productos, dirección, etc.
- **Categorías disponibles en mercado libre:** Este punto resulta importante a tener en cuenta, en mercado libre cada producto pertenece a una categoría en específico, la cual puede tener categorías padres (la sub categoría adaptador de computadores su categoría padre sería computadores) la cual cada una tiene su propio conjunto de atributos específicos.
- **Catálogo total disponible por categoría:** Se puede obtener el catálogo completo de cada sub categoría disponible.

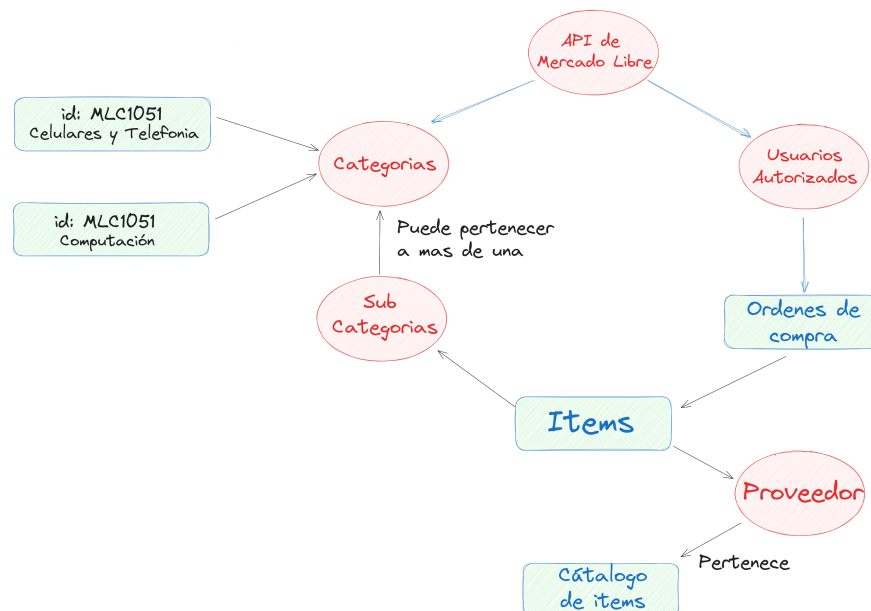


Figura 3.1 Diagrama de organización de información en Mercado Libre

3.2. Sistema de Recomendación

Dentro del proceso de sistema de recomendación, resulta importante tener en cuenta los modelos disponibles que fueron detallados en la Sub Sección 2.5.3. Para ello, explora la información que pudo recopilar de la empresa respectiva, el análisis y exploración de esta a través de gráficos para posteriormente definir cuál será el modelo de recomendación a ocupar y como se construirá usando los recursos anteriormente mencionados.

1. Preprocesamiento de datos
2. Exploración de datos

3. Construcción del Modelo de recomendación
4. Interpretación y recomendaciones
5. Iteración y mejora

Dentro del desarrollo del sistema , se debe tener en cuenta que servicios fueron ocupados para obtener la información, tales como los de Mercado Libre y fue implementada en el lenguaje de programación *Python*. Además, como se verá en los siguiente puntos, la única información disponible es sobre los productos de Mercado Libre y el registro de compras del usuario (en este caso la empresa).

3.2.1. Preprocesamiento de datos

En el contexto de comercio electrónico se cuenta con mucha información relacionada a la compra y venta de los productos en la plataforma, tales como precio y cantidad disponible son información básica en estos sistemas. Dentro de toda la información disponible para cada producto, se pudo recopilar la siguiente para los productos:

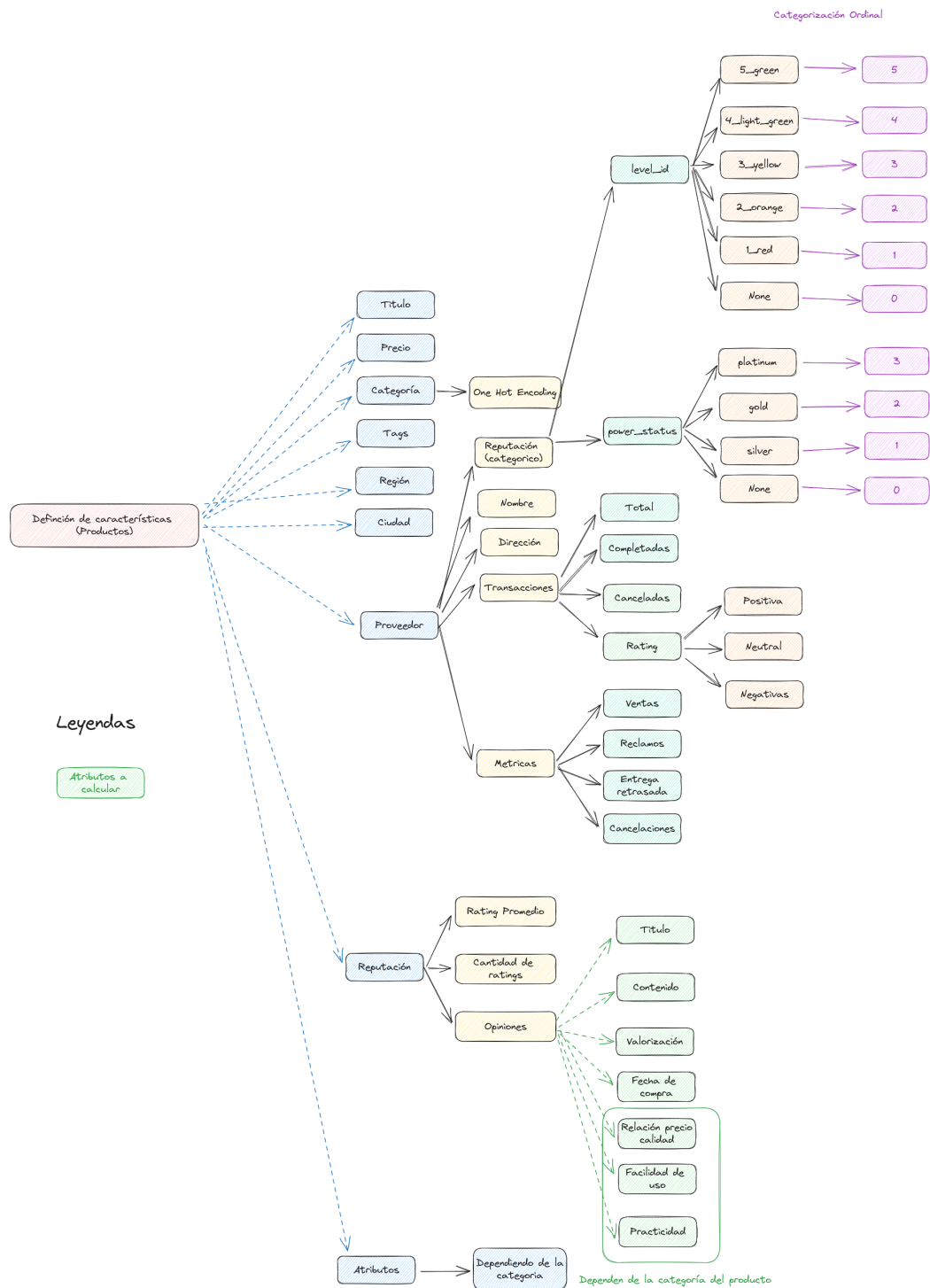


Figura 3.2 Diagrama de características de productos

Mediante estas características se pudo determinar 3 tipos de datos dentro de las características de los productos:

- **Características numéricas:** Dentro de esta categoría corresponderá todos los atributos que sean de este tipo, los cuales serán parte dentro del cálculo de similitud posterior, las cuales serán: precio, cantidad de calificaciones que cuenta el producto, cantidad de una estrella, cantidad de dos estrellas, cantidad de tres estrellas, cantidad de cuatro estrellas, cantidad de cinco estrellas, calificación promedio, cantidad de likes que cuenta las opiniones, cantidad de dislikes que cuenta las opiniones, cantidad de ventas completadas, cantidad de ventas canceladas, cantidad de ventas totales, porcentaje de rating positivo, porcentaje de rating neutral, porcentaje de rating negativo, cantidad de ventas, de reclamos, entregas retrasadas y canceladas por mes.
- **Características textuales:** Descripción del producto, título del producto, etiqueta o tags asociados a este, opiniones, nombre del vendedor, etiquetas, región y ciudad del vendedor.
- **Características categóricas ordinal:** Reputación (5_gold, 5_platinum, etc) y reputación estrellas ("5 estrellas", "4.5 estrellas", etc).
- **Características categóricas nominales:** Estas características hace referencia a los atributos de cada categoría, por lo general se representan como textos si el producto tiene información del atributo respectivo.

Para esto, se aplicaron técnicas de preprocesamiento para mejorar el desempeño del modelo a usar, lo que se traduce en las características determinadas en la Tabla 3.2.1 en donde se detallan cuales serán ocupadas dentro del cálculo de similitud por el sistema, siendo las únicas que no se ocuparan los ID del producto y vendedor (dado que obtener una similitud entre ID no aportaría una buena métrica de comparación) y la categoría, la cual será ocupada solo para organización de las recomendaciones. Para las variables numérica se pueden aplicar técnicas de normalización o estandarización y para la información de carácter de texto, se ocupa el vectorizador de TF-IDF que vimos en su Ecuación 2.5.3, que permite visualizar la frecuencia de ocurrencia para cada característica textual dentro de los productos y así obtener una similitud entre productos.

Usada para Cálculo	No Usada para Cálculo
Título, Descripción, Precio, Tags, Región, Ciudad, Cantidad Reviews, Cantidad Estrellas, Review Promedio, Likes y Dislikes, Información del Vendedor, Reputación Vendedor, Registros de transacciones (completadas, canceladas, total), Ratings, Ventas, Reclamos, Entregas, Cancelaciones	IDProducto, Categoría, IdVendedor

Tabla 3.1 Tablas con las características de la matriz resultante

3.2.2. Exploración de datos

En esta sección se irá explorando y detallando cual es la información y datos con que se cuentan que estén asociados en la empresa. De primera instancia se pudo obtener el registro

de las compras realizadas en el año a diferentes proveedores y diferentes productos, los cuales, se pueden visualizar en la Figuras 3.3 y 3.4.

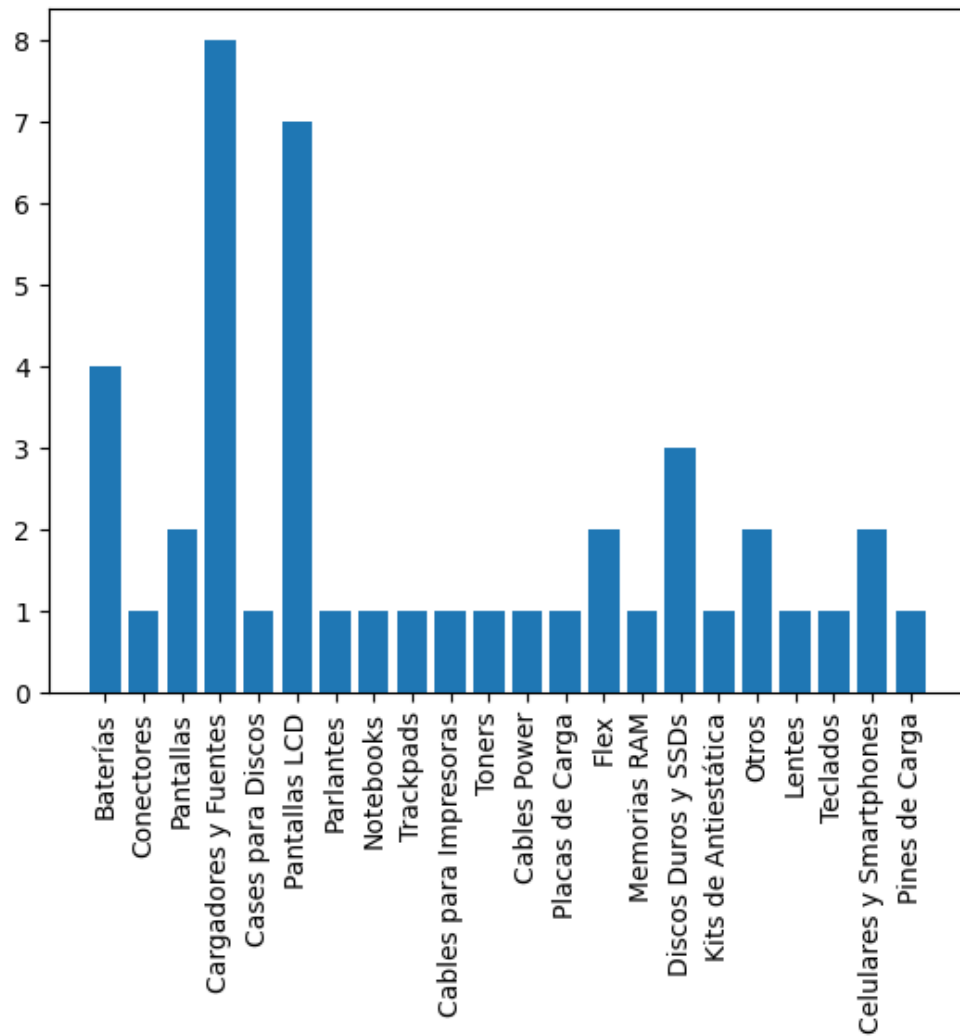


Figura 3.3 Gráfico de barras con las categorías de productos de interés de la empresa

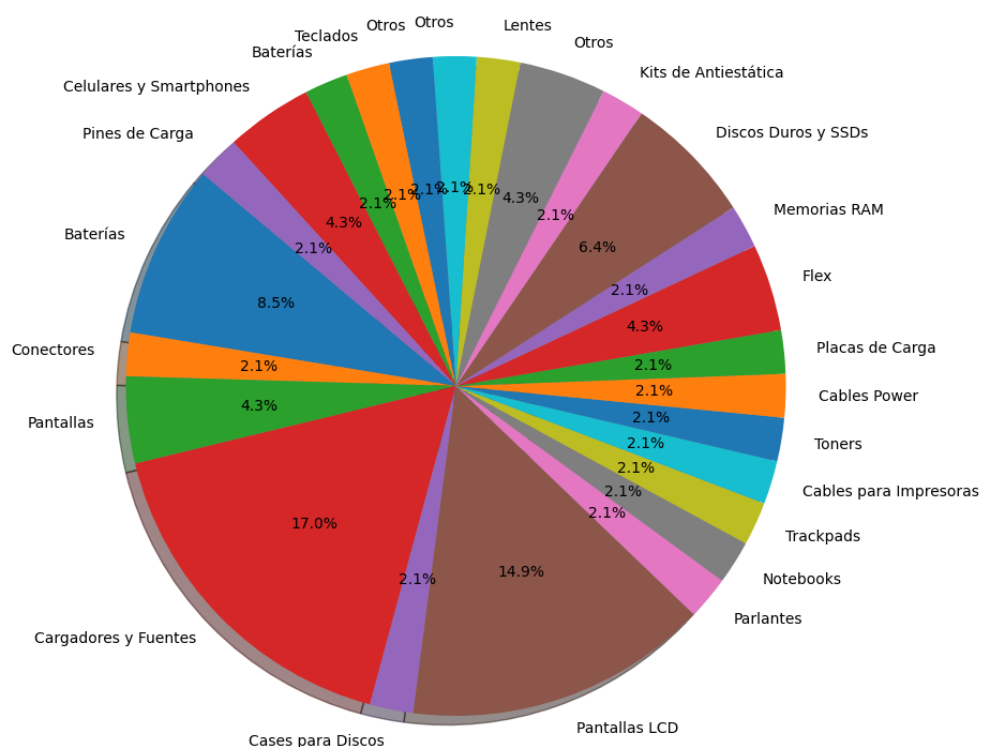


Figura 3.4 Gráfico de pie sobre las categorías y su porcentaje de productos en el catálogo

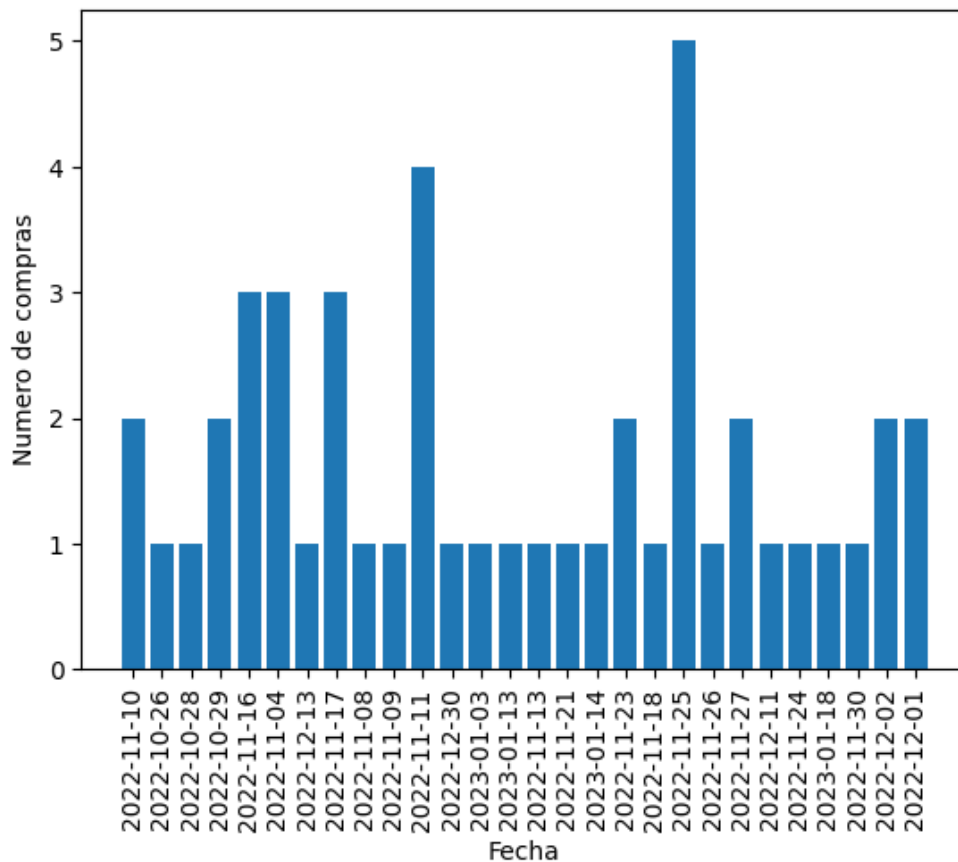


Figura 3.5 Gráfico de barras donde se muestra el registro de compras de la empresa por fecha

Dentro de este análisis, se pudo obtener el registro de las compras en el tiempo que se realizó cada una Figura 3.5 y la cantidad de compras según la categoría definidas por el sistema de mercado libre. Podemos apreciar que la categoría de 'Cargadores y Fuentes', 'Baterías', 'Pantallas' y 'Discos Duros' son los productos que tienen mayores registros dentro de este conjunto de datos inicial. El resultado obtenido de las fechas no denota un patrón claro en la compra de ciertos productos, que se acompaña con el desafío de atributo temporal en los registros del usuario; para este caso, los clientes de la empresa varían según temporada (agregando que el registro no goza de gran cantidad de datos para determinar una tendencia clara).

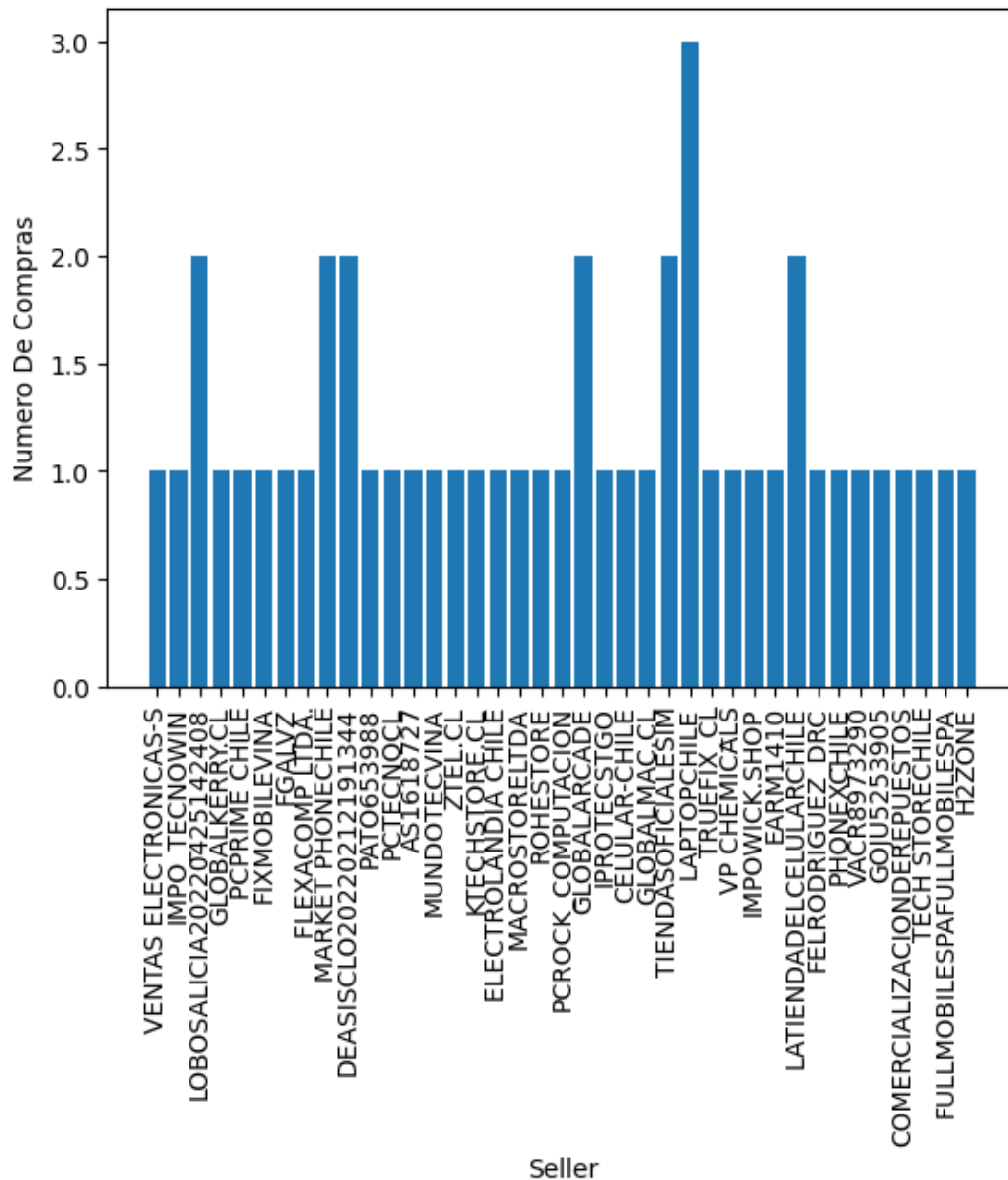


Figura 3.6 Cantidad de compras de la empresa a cada proveedor registrado

Estas figuras hacen correspondencia a los datos e información sustraída del conjunto de datos de la misma empresa, y dada la naturaleza del modelo que se ocupará, se obtuvo a través de los servicios de mercado libre la información de todos los productos asociados a cada categoría, las cuales se tomaron las 4 principales con más compras de la empresa. Estas son: Cargadores y Fuentes, Pantallas LCD, Baterías y Discos duros y SSD, cada una de estas categorías tienen al menos 1000 filas (1000 ítems) cada una, imposibilitando la obtención de más dada la limitación de los servicios de mercado libre. Para estas, se realizó un análisis exploratorio en diferentes ámbitos (lo que se puede apreciar en las Figuras 3.11).

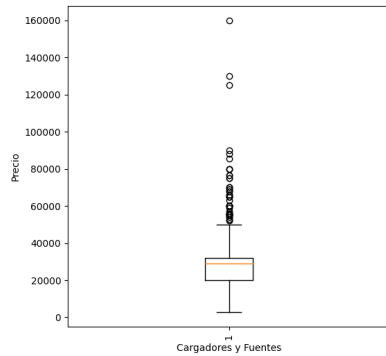


Figura 3.7 Cargadores y Fuentes

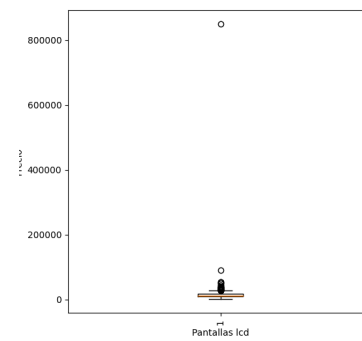


Figura 3.8 Baterías

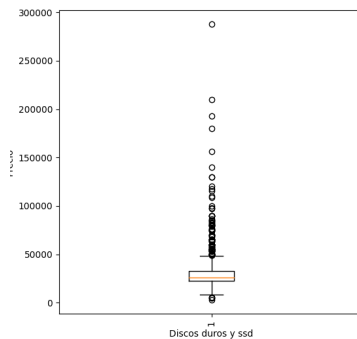


Figura 3.9 Pantallas

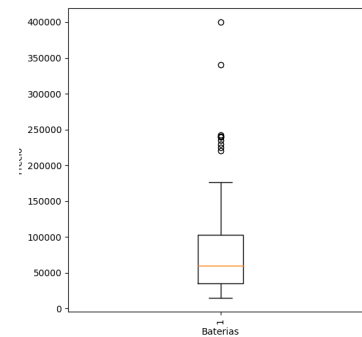


Figura 3.10 Discos Duros y SSD

Figura 3.11 Gráficos box plots que muestra el rango de precios según categoría

Dentro del análisis de estas cuatro categorías, se reconocen el título y código de producto del conjunto de datos inicial, por lo que se explorará las características de los registros obtenidos del usuario:

idProducto	Titulo
MLC518800936	Cargador Para Dell 19.5v 4.62a Inspiron Vostro Latitud
MLC540076822	Cargador Dell 19.5v 6.7a 130w
MLC518395062	Cargador Notebook Hp Pavilion Split 19.5v 2.31a 45w X360
MLC956085170	Cargador Notebook Acer Aspire 19v 3.42a 65w E3 E1 E15 5610
MLC958944207	Cargador Orig. Netbook Packard Bell Dot_s.cl/010 (Kav60)
MLC929084184	Cargador Notebook 19v 3.42a Toshiba Satellite Packard Bell
MLC929084184	Cargador Notebook 19v 3.42a Toshiba Satellite Packard Bell
MLC951082625	Cargador Notebook Acer Aspire 19v 3.42a 65w E3 E1 E15 5610

Tabla 3.2 Títulos de Cargadores y Fuentes

idProducto	Titulo
MLC982138741	Bateria Original Samsung Note 20 N980 4300 Mah Genuina
MLC981083215	Batería Compatible Con Samsung S8
MLC964320528	Bateria Para iPhone 6s + Adhesivo - Dcompras
MLC1002089245	Bateria Para iPhone 12 Y 12 Pro Originales.

Tabla 3.3 Títulos de Baterías

idProducto	Titulo
MLC614763390	Pantalla Completa Para Samsung Galaxy S8 / Envio Gratis
MLC1032855126	Pantalla Para Redmi Note 11/ 11s 4g
MLC642691446	Pantalla Original Huawei P30 Lite (Repuesto)
MLC967634030	Pantalla Para Huawei P30 Oled + Mica Regalo - Dcompras
MLC1037100291	Pantalla Para Samsung A32 4g Con Marco + Mica Regalo - Ltc
MLC1037100291	Pantalla Para Samsung A32 4g Con Marco + Mica Regalo - Ltc
MLC1012445866	Pantalla Compatible Samsung A32 4g Oled + Lamina Full Mobile

Tabla 3.4 Títulos de Pantallas

idProducto	Titulo
MLC988116866	Disco Sólido Ssd Interno Western Digital Wds200t2b0a 2tb Azul
MLC908985821	Disco Sólido Ssd Interno Western Digital Wd Green Wds240g2g0a 240gb Verde
MLC987952221	Disco Sólido Ssd Interno Western Digital Wds500g2b0a 500gb Azul

Tabla 3.5 Títulos de Discos

Dentro de estos productos del registro inicial, resulta importante calcular la cantidad de valores nulos y cuantos de estos realmente aparecen en este catálogo inicial de 1000 productos obtenidos desde mercado libre, dado que, para los resultados que se analizarán en los siguientes capítulos, será necesaria analizar.

Categoría	Porcentaje Nulos	Productos Catálogo
Cargadores	49 %	'MLC540076822' 'MLC518395062' 'MLC951082625'
Baterías	45 %	'MLC964320528'
Pantallas	48 %	'MLC1032855126' 'MLC1037100291' 'MLC1037100291' 'MLC1012445866'
Discos	56 %	'MLC987952221'

Tabla 3.6 Tabla de datos nulos y productos que pertenecen al catálogo por categoría

Las principales ciudades de los productos que fueron comprados por la empresa, se ubican en la región metropolitana, en vendedores que varían entre diferentes ciudades, siendo un solo registro de la ciudad de Viña del Mar de una batería asociada. También se aprecia el rango de valores de los productos registrados en las Figuras ??, particularmente 'Cargadores y Fuentes' presenta un valor atípico que se aleja de los demás registros, siendo un cargador con un precio muy elevado; por otro lado, el rango de precios de las otras categorías se visualizan precios con un rango bastante amplio, que su valor medio va entre 10.000 y 20.000 pesos chilenos.

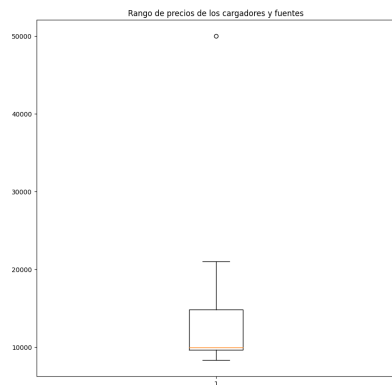


Figura 3.12 Cargadores y Fuentes

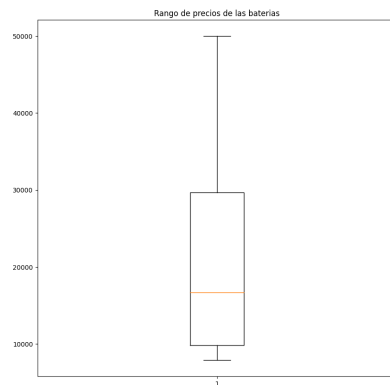


Figura 3.13 Baterías

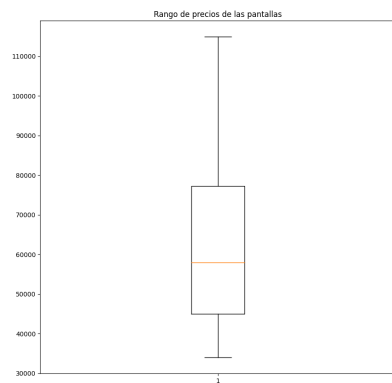


Figura 3.14 Pantallas

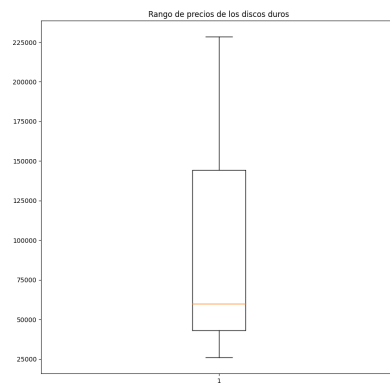


Figura 3.15 Discos Duros y SSD

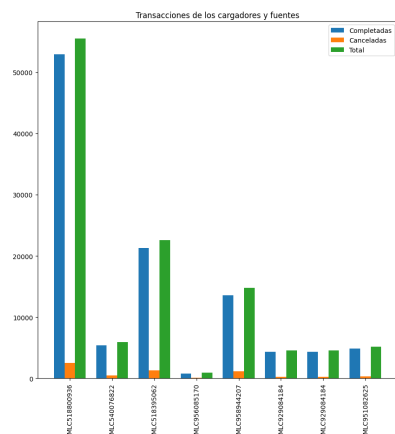


Figura 3.16 Cargadores y Fuentes

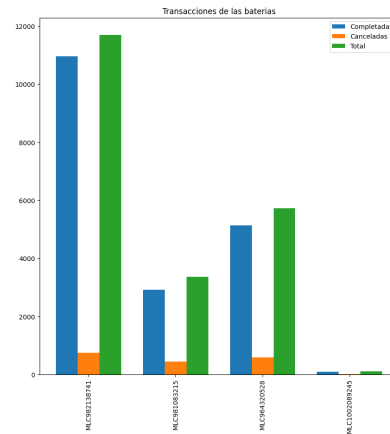


Figura 3.17 Baterías

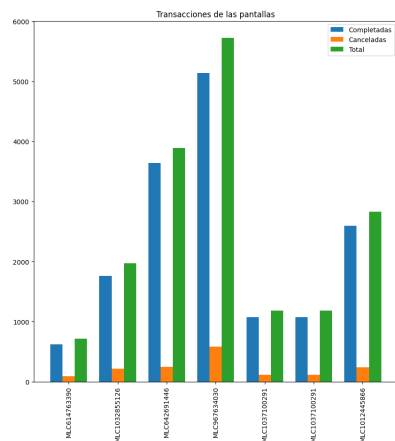


Figura 3.18 Pantallas

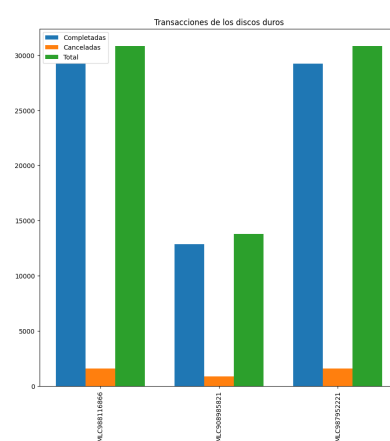


Figura 3.19 Discos Duros y SSD

En los siguientes gráficos se muestra las características de los vendedores asociados a los productos, desde el rating en estrellas que tiene el vendedor, donde un solo registro no obtiene el valor máximo dentro de los registros. El único resultado que se escapa de la norma apreciada, es el último vendedor registrado en las baterías, que se traduce en una menor reputación general y que la cantidad de transacciones es mucho menor a los otros proveedores vistos en esta categoría.

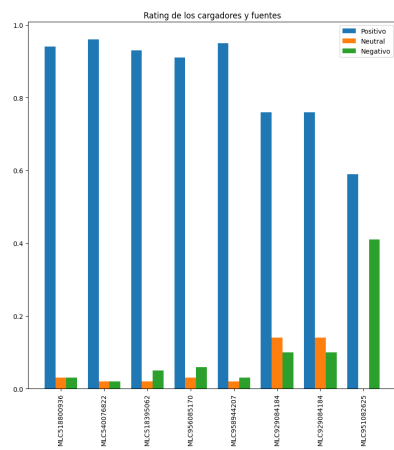


Figura 3.20 Cargadores y Fuentes

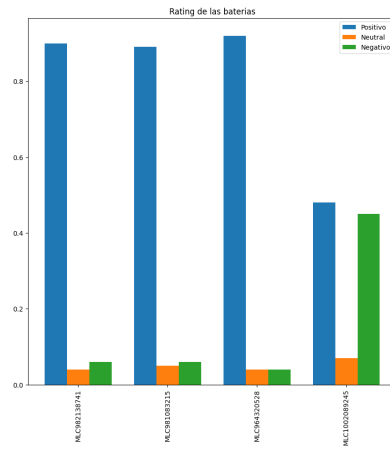


Figura 3.21 Baterías

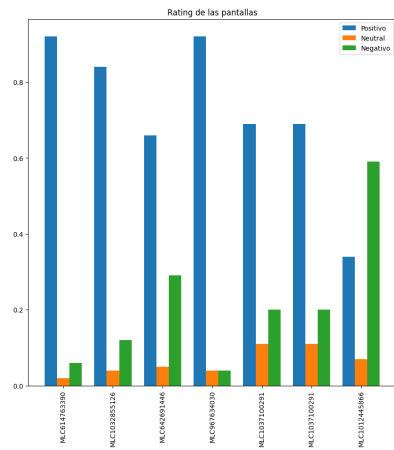


Figura 3.22 Pantallas

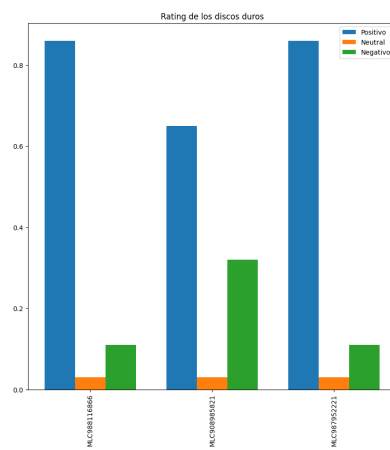


Figura 3.23 Discos Duros y SSD

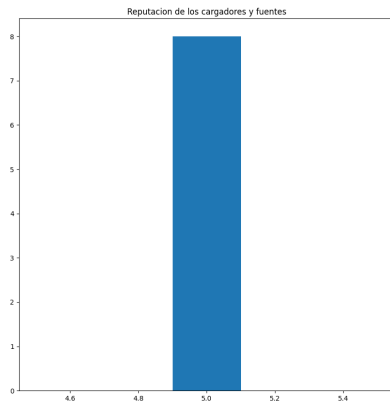


Figura 3.24 Cargadores y Fuentes

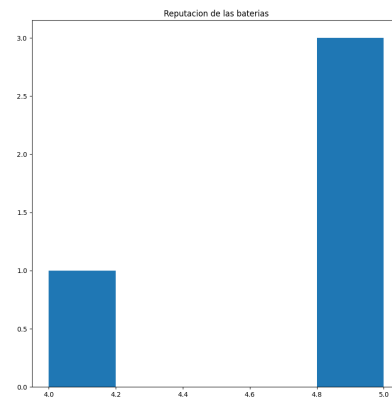


Figura 3.25 Baterías

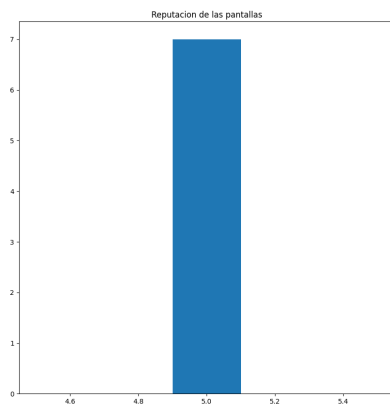


Figura 3.26 Pantallas

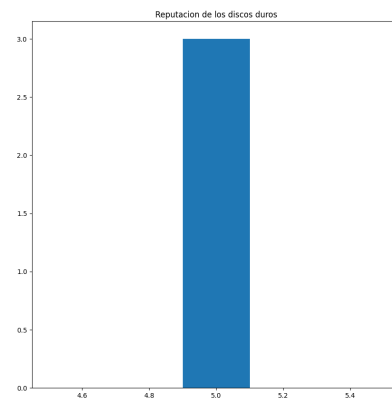


Figura 3.27 Discos Duros y SSD

3.2.3. Construcción del Modelo de recomendación

Como se comentó en los puntos anteriores, se puede visualizar que la información disponible denota un uso recomendado de un modelo basado en contenido, dada la falta de información sobre otros usuarios (solamente la empresa), restringiendo así el uso de un modelo con filtrado colaborativo, y por otro lado, ocupar reglas específicas como se dan en el modelo de conocimiento no resultan tan adecuadas para este contexto, esto debido a que las preferencias o tendencias de la empresa varían según las necesidades de sus clientes, los cuales, no se nota una clara tendencia temporal o favoritismo entre ciertos productos o marcas. Para realizar la construcción del modelo de recomendación basado en contenido, resulta importante el cual será el perfil del usuario y/o información que será clave para el proceso, como se definió en Sección 2.5.3 existen dos tipos de información que se puede ocupar: implícita o explícita. Para este caso, la información disponible es del primer tipo, dado que dentro del registro de compras obtenido de la empresa no cuenta con una calificación explícita de los productos (la calificación es opcional), por lo que, se construirá un perfil de usuario usando su registro de compras anteriormente analizado. Para ello, se realizó un análisis tomando en consideración puntos importantes como los siguientes.

Decisiones de diseño e implementación

Uno de los principales desafíos que se enfrenta un sistema de recomendación (como se menciona en la Sección 2.5.1 es el muestreo de datos, que determinará como se categorizará y ordenará el conjunto de datos disponible. En este caso, se definieron tres formas de como organizar la información según las cuatro categorías que se exploran:

- **Matriz única con toda la información:** La propuesta sugiere consolidar toda la información de productos en una sola matriz para simplificar la consulta y recomendación, aunque esto puede resultar en tiempos de cálculo excesivos debido a la gran cantidad de datos y valores nulos. Se plantea la necesidad de implementar un algoritmo de clasificación para organizar los productos en esta única matriz.
- **Matrices compartidas según categorías padre:** En Mercado Libre, las categorías y subcategorías forman un árbol jerárquico, reduciendo los datos nulos al agrupar productos bajo categorías padres como “computadores”. Sin embargo, esto limita las recomendaciones a productos dentro de estas categorías principales, lo que podría no abarcar todas las opciones disponibles. Además, la diversidad de subcategorías dentro de una categoría padre requiere el uso de algoritmos de clasificación para identificar productos específicos dentro de cada grupo.
- **Definir matrices por cada sub categoría:** La propuesta sugiere utilizar una matriz específica para cada categoría de productos (como en la Figura 3.3, lo que reduce los datos nulos y aumenta la precisión de las recomendaciones dentro de una categoría. Sin embargo, esto limita la cobertura del modelo a productos dentro de la misma categoría, dependiendo de las preferencias del usuario. Aunque el uso de múltiples matrices acelera el cálculo, también aumenta la complejidad del preprocesamiento debido al mayor número de iteraciones requeridas.

Con la finalidad de entregar una mayor precisión en los resultados de recomendación, el diseño a construir será la tercera propuesta, dado que, la primera y segunda propuesta entra la necesidad de un uso de algoritmos de clasificación para identificar grupos de productos entre todo los catálogos disponibles. Esto en cierta forma, será desaprovechar la característica que entrega Mercado Libre, en cual ya los productos están clasificados y permite una mejor diferenciación entre estos.

Con esto ya definido, se define los conjuntos de las subcategorías: $C_1, C_2, \dots, C_k$, la cual cada una tendrá asociada una matriz M_{C_i} que contiene la información relevante de los productos pertenecientes a esta subcategoría. Por lo que se puede definir la matriz como:

$$M_{C_i} = \begin{bmatrix} \text{tf-idf}_{11}, & \text{tf-idf}_{12}, & \dots & \text{tf-idf}_{1n} \\ \text{tf-idf}_{21}, & \text{tf-idf}_{22}, & \dots & \text{tf-idf}_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \text{tf-idf}_{m1}, & \text{tf-idf}_{m2}, & \dots & \text{tf-idf}_{mn} \end{bmatrix}$$

Donde:

- Cada fila de la matriz representa un producto específico $p \in C_i$.

- Cada columna representa un término t relevante para describir los productos de la categoría C_i .
- $tf - idf_{jk}$ es el valor TF-IDF del término t_k en el producto p_j

Para un producto p_j en la subcategoría C_i su vector de características basado en TF-IDF sería:

$$\mathbf{v}_{p_j} = \left[tf-idf(t_1, p_j, C_i), \quad tf-idf(t_2, p_j, C_i), \quad \dots \quad tf-idf(t_n, p_j, C_i) \right]$$

Por lo que, la forma final de la matriz, integrando los vectores de características de todos los productos en la subcategoría C_i , obtenemos la matriz M_{C_i} :

$$\mathbf{M}_{C_i} = \begin{bmatrix} \mathbf{v}_{p_1} \\ \mathbf{v}_{p_2} \\ \vdots \\ \mathbf{v}_{p_m} \end{bmatrix} = \begin{bmatrix} tf-idf(t_1, p_1, C_i), & tf-idf(t_2, p_1, C_i), & \dots & tf-idf(t_n, p_1, C_i) \\ tf-idf(t_1, p_2, C_i), & tf-idf(t_2, p_2, C_i), & \dots & tf-idf(t_n, p_2, C_i) \\ \vdots & \vdots & \ddots & \vdots \\ tf-idf(t_1, p_m, C_i), & tf-idf(t_2, p_m, C_i), & \dots & tf-idf(t_n, p_m, C_i) \end{bmatrix}$$

Implementación de TF-IDF, Stop Word y Stemming

Para implementar un sistema de recomendación basado en contenido, es esencial emplear el algoritmo TF-IDF, que calcula la similitud entre productos utilizando sus atributos de texto, como descripciones, etiquetas y comentarios. Este enfoque también considera las características de los vendedores, como sus etiquetas y descripciones, para mejorar la comparación. Se crea un diccionario de palabras clave a partir de los documentos, con especial atención a la identificación y omisión de palabras vacías o "Stop Words" que no son relevantes para el análisis (como se explora en el libro [Aggarwal, 2016]). Además, se aplica la técnica de stemming para identificar la raíz de las palabras y considerar sus variaciones gramaticales, lo que ayuda a mejorar la precisión de la comparación de textos. La biblioteca NLTK de Python proporciona funciones para implementar estas técnicas de procesamiento de lenguaje natural.

Para definir, la matriz M_{C_i} con stemming se define como:

$$\mathbf{M}_{C_i} = \begin{bmatrix} tf-idf(stem(t_1), p_1, C_i), & tf-idf(stem(t_2), p_1, C_i), & \dots & tf-idf(stem(t_n), p_1, C_i) \\ tf-idf(stem(t_1), p_2, C_i), & tf-idf(stem(t_2), p_2, C_i), & \dots & tf-idf(stem(t_n), p_2, C_i) \\ \vdots & \vdots & \ddots & \vdots \\ tf-idf(stem(t_1), p_m, C_i), & tf-idf(stem(t_2), p_m, C_i), & \dots & tf-idf(stem(t_n), p_m, C_i) \end{bmatrix}$$

La expresión matemática para con un producto p_j en la subcategoría C_i su vector de características basado en TF-IDF con stemming sería:

$$\mathbf{v}_{p_j} = \left[tf-idf(stem(t_1), p_j, C_i), \quad tf-idf(stem(t_2), p_j, C_i), \quad \dots \quad tf-idf(stem(t_n), p_j, C_i) \right]$$

Para calcular la similitud entre dos productos p_a y p_b dentro de la subcategoría C_i podemos utilizar la similitud del coseno. La similitud del coseno entre los vectores de características $\mathbf{v}_{p_a}$ y $\mathbf{v}_{p_b}$ se define como:

$$\text{sim}(p_a, p_b) = \frac{\mathbf{v}_{p_a} \cdot \mathbf{v}_{p_b}}{\|\mathbf{v}_{p_a}\| \|\mathbf{v}_{p_b}\|}$$

Donde:

- $v_{p_a} \cdot v_{p_b}$ es el producto punto de los vectores TF-IDF con stemming.
- $\|v_{p_a}\|$ y $\|v_{p_b}\|$ son las normas (magnitudes) de los vectores TF-IDF con stemming.

Integrando todo, la matriz M_{C_i} la expresión de similitud con stemming se define como:

$$M_{C_i} = \begin{bmatrix} \text{tf-idf}(\text{stem}(t_1), p_1, C_i), & \text{tf-idf}(\text{stem}(t_2), p_1, C_i), & \cdots & \text{tf-idf}(\text{stem}(t_n), p_1, C_i) \\ \text{tf-idf}(\text{stem}(t_1), p_2, C_i), & \text{tf-idf}(\text{stem}(t_2), p_2, C_i), & \cdots & \text{tf-idf}(\text{stem}(t_n), p_2, C_i) \\ \vdots & \vdots & \ddots & \vdots \\ \text{tf-idf}(\text{stem}(t_1), p_m, C_i), & \text{tf-idf}(\text{stem}(t_2), p_m, C_i), & \cdots & \text{tf-idf}(\text{stem}(t_n), p_m, C_i) \end{bmatrix}$$

Para calcular la similitud entre dos productos p_a y p_b dentro de la subcategoría C_i usando la matriz TF-IDF, utilizamos la similitud de coseno:

$$\text{sim}(p_a, p_b) = \frac{\sum_{k=1}^n \text{tf-idf}(\text{stem}(t_k), p_a, C_i) \cdot \text{tf-idf}(\text{stem}(t_k), p_b, C_i)}{\sqrt{\sum_{k=1}^n (\text{tf-idf}(\text{stem}(t_k), p_a, C_i))^2} \cdot \sqrt{\sum_{k=1}^n (\text{tf-idf}(\text{stem}(t_k), p_b, C_i))^2}}$$

Implementación de palabras claves del usuario

Dentro de las necesidades que se obtuvo de la empresa, esta muestra como valor agregado el que el mismo usuario pueda definir palabras claves en las recomendaciones obtenidas, con la finalidad de poder ver diferentes resultados en la obtención de recomendaciones por parte del sistema. Para esto, se implementó un sistema de pesos en las palabras identificadas en el vector TF-IDF, esto significa, para cualquier palabra t en el documento, si pertenece al conjunto palabras ingresadas por el usuario w_u , se le asigna un peso de 10 en vez de 1.

Para definir la matriz M_{C_i} , se necesita definir un factor de peso α_t para cada término t :

$$\alpha_t = \begin{cases} 10 & \text{si } t \in w_u \\ 1 & \text{si } t \notin w_u \end{cases}$$

La definición de la matriz M_{C_i} con los pesos ajustados se define como:

$$M_{C_i} = \begin{bmatrix} \alpha_{t_1} \cdot \text{tf-idf}(t_1, p_1, C_i), & \alpha_{t_2} \cdot \text{tf-idf}(t_2, p_1, C_i), & \cdots & \alpha_{t_n} \cdot \text{tf-idf}(t_n, p_1, C_i) \\ \alpha_{t_1} \cdot \text{tf-idf}(t_1, p_2, C_i), & \alpha_{t_2} \cdot \text{tf-idf}(t_2, p_2, C_i), & \cdots & \alpha_{t_n} \cdot \text{tf-idf}(t_n, p_2, C_i) \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_{t_1} \cdot \text{tf-idf}(t_1, p_m, C_i), & \alpha_{t_2} \cdot \text{tf-idf}(t_2, p_m, C_i), & \cdots & \alpha_{t_n} \cdot \text{tf-idf}(t_n, p_m, C_i) \end{bmatrix}$$

Para calcular la similitud entre dos productos p_a y p_b dentro de la subcategoría C_i usando la matriz TF-IDF con los pesos ajustados, utilizamos la similitud de coseno:

$$\text{sim}(p_a, p_b) = \frac{\sum_{k=1}^n (\alpha_{t_k} \cdot \text{tf-idf}(t_k, p_a, C_i) \cdot \alpha_{t_k} \cdot \text{tf-idf}(t_k, p_b, C_i))}{\sqrt{\sum_{k=1}^n (\alpha_{t_k} \cdot \text{tf-idf}(t_k, p_a, C_i))^2} \cdot \sqrt{\sum_{k=1}^n (\alpha_{t_k} \cdot \text{tf-idf}(t_k, p_b, C_i))^2}}$$

Reducción de dimensionalidad

Como se ve en los gráficos de columnas atributos y en la decisión de diseño que se definió anteriormente, dentro de los productos listados para cada categoría, hay muchas columnas que son totalmente nulas, lo que genera un tiempo de cálculo innecesario tomando en cuenta

esos datos. Para ello, según lo detallado en la Sección 2.5.1, el método de SVD aplica mejor para matrices con datos nulos, que con la gran cantidad de datos nulos (especialmente en las columnas de atributos) resulta idóneo para este caso.

Para una matriz TF-IDF M_{C_i} de una subcategoría C_i , la SVD se define como:

$$M_{C_i} = U \Sigma V^T$$

Donde:

- U es una matriz unitaria de tamaño $m \cdot m$.
- Σ es una matriz diagonal de tamaño $m \cdot n$ con valores singulares en diagonal.
- V^T es la transpuesta de una matriz unitaria de tamaño $n \cdot n$.

Para reducir la dimensionalidad, seleccionamos los k componentes más importantes, esto implica truncar U , Σ y V^T para mantener las k dimensiones más significativas. Las matrices truncadas se denotan como:

$$M_{C_i} \approx U_k \Sigma_k V_k^T$$

Donde:

- U_k es una matriz unitaria de tamaño $m \cdot k$.
- Σ_k es una matriz diagonal de tamaño $k \cdot k$.
- V_k^T es la transpuesta de una matriz unitaria de tamaño $k \cdot n$.

Para aplicar la técnica SVD a la matriz M_{C_i} y truncamos las matrices para obtener U_k , Σ_k y V_k^T :

$$M_{C_i} \approx U_k \Sigma_k V_k^T$$

Para calcular la similitud con el espacio de dimensiones reducidas, los vectores de características de los productos p_a y p_b se transforman a:

$$\mathbf{v}'_{p_a} = U_k \Sigma_k (\mathbf{v}_{p_a} V_k^T)^T$$

$$\mathbf{v}'_{p_b} = U_k \Sigma_k (\mathbf{v}_{p_b} V_k^T)^T$$

La similitud entre los productos p_a y p_b en el espacio reducido se calcula como:

$$\text{sim}(p_a, p_b) = \frac{\mathbf{v}'_{p_a} \cdot \mathbf{v}'_{p_b}}{\|\mathbf{v}'_{p_a}\| \|\mathbf{v}'_{p_b}\|}$$

3.2.4. Plataforma de Mercado Libre

La construcción del sistema esta basada ocupando netamente la información obtenida desde la API de Mercado libre, dada la flexibilidad, constancia, parámetros y funcionalidades ya testeadas en esta plataforma; esto nace según la necesidad del cliente, dado que, como se explica anteriormente, es la única plataforma que tiene registros oficiales de la empresa y la cual regularmente ocupa. Esto conlleva una falta de diversidad en la cantidad de productos a recomendar, pero permite una mayor precisión en el cálculo de similitud (al tener todos los productos los mismos campos), siendo un trabajo futuro adaptar mediante Web Scrapping y la identificación del código único (EAN) de cada producto para poder realizar buenas comparaciones de similitud. En este trabajo solamente se explora Mercado Libre.

Algoritmos de Generación de datos sintéticos

Si bien se cuenta con un registro inicial de datos obtenidas desde la empresa, para los estándares de un sistema de recomendación resultan demasiado acotados y escaso, generando que los resultados muestren un sobreajuste excesivo, por lo que se explora la opción de aplicar algoritmos de generación de datos sintéticos, como GANs (Redes Generativas Adversarias) o VAEs (Autoencoders variacionales) que permitirían crear una base de datos más amplia. Sin embargo, esto significa otro desarrollo de un sistema de mayor complejidad que el de recomendación para no caer en este mismo sobre ajuste (si el modelo de generación no esta bien afinado puede empeorar los resultados) y las interpretaciones no serían totalmente aplicadas a la realidad de la empresa; esto se acompaña que con el uso de la plataforma y el ingreso de más proveedores, plataformas y usuarios, el mismo sistema será enriquecido de información y variada, por lo que, no es necesidad vital a medio y largo alcance.

Calculo de similitud y pesos

Los procesos generan dos matrices de usuarios y dos de productos, una para características numéricas y otra para texto. Las características numéricas se estandarizan, mientras que las de texto se transforman utilizando vectores de palabras previamente entrenados, donde cada valor indica la relevancia de las palabras en el documento.

Se utiliza la similitud de coseno para calcular la similitud entre el registro del usuario y el catálogo de productos, generando una matriz de dimensiones $u * p$, siendo u la cantidad de registros del usuario y p la cantidad de productos que fueron parte del cálculo. Se aplica este proceso también a las matrices obtenidas del TF-IDF. Se pueden asignar pesos a las matrices resultantes para priorizar ciertas características, basándose en la cantidad de columnas utilizadas en el cálculo de cada tipo de característica, es decir, si 3 columnas numéricas fueron ocupadas y 7 columnas de texto, se atribuirá un peso respectivo a la cantidad de columnas ocupadas (30 % y 70 % respectivamente).

La matriz resultante de similitud proporciona la similitud entre cada producto del registro de usuarios y cada producto del catálogo. Para identificar los productos más similares para los usuarios, se suman las similitudes de todos los registros. Sin embargo, la suma directa puede exceder la escala entre cero y uno de la métrica de coseno. Por lo tanto, se pueden aplicar diferentes métodos de normalización, como la normalización L1 [2.5.1](#), L2 [2.5.1](#) y MaxMin [2.5.1](#).

3.2.5. Interpretación y recomendaciones

El proceso de interpretación y recomendación se puede ver reflejado en la siguiente Figura 3.28 la cual detalla las cuatro etapas que el sistema realiza para obtener y entregar la recomendación al usuario.

- **Recolección de datos:** Como su nombre lo indica, esta etapa corresponde a la recolección u obtención de los datos ya guardados en la base de datos respectiva, filtrando según la categoría que el usuario desea y obteniendo los dos archivos CSV almacenados en la base de datos, tanto los datos del usuario (registro de compras) como los datos del catálogo (productos que se encuentran en la plataforma mercado libre).
- **Categorización de columnas:** En esta etapa se separa la información de las matrices según 3 tipos de columnas, numéricas, de texto y de atributos, aplicando una estandarización a los valores obtenidos a la columna numérica, con la finalidad de que cada número aporte el mismo peso ponderado; además, en esta etapa está el proceso de aplicación de reducción de dimensionalidad a la columna de atributos.
- **Cálculo de similitud:** En esta etapa, las matrices se dividen para aplicar el vector tf-idf, considerando la información de ambas matrices y la frecuencia de palabras en toda la información disponible. Se aplica Stemming si es seleccionado, así como la ponderación de palabras clave para otorgar mayor peso a ciertas palabras. Cada columna de atributos tiene su propio proceso de vector tf-idf, y se calcula la similitud de coseno entre cada categoría de columnas, obteniendo así tres matrices de similitud entre productos de usuarios y productos del catálogo.
- **Suma y normalización:** En esta etapa, se realiza una suma ponderada de las tres matrices de similitud según la cantidad de columnas de cada categoría. Luego, se suma por el eje y para obtener una única matriz que representa la similitud total de cada registro de usuario con el catálogo. Finalmente, se aplica la normalización L1 para obtener un vector donde cada atributo indica la similitud de cada producto del catálogo al usuario, permitiendo así la recomendación de los productos más similares.

3.2.6. Iteración y mejora

El sistema de recomendación debe ir teniendo en cuenta cada nuevo registro que se incorpore al historial del usuario, con el fin de ir actualizando recomendaciones actualizadas al estado de la empresa, por consiguiente, el sistema debería permitir la compra de los productos recomendados o no. Dado que el sistema contiene las matrices definidas en los puntos anteriores, cuando el usuario realice una compra en la plataforma web, el sistema obtendrá el nuevo registro y agregará el registro ocupando la información almacenada en el respectivo catálogo; esto conlleva que el usuario no podrá comprar productos que no se encuentren en el catálogo, por lo que, el mantener actualizado este listado le dará un gran alcance y escalabilidad con el tiempo al sistema. Estos puntos se desarrollarán y explorarán en el desarrollo de la plataforma web.

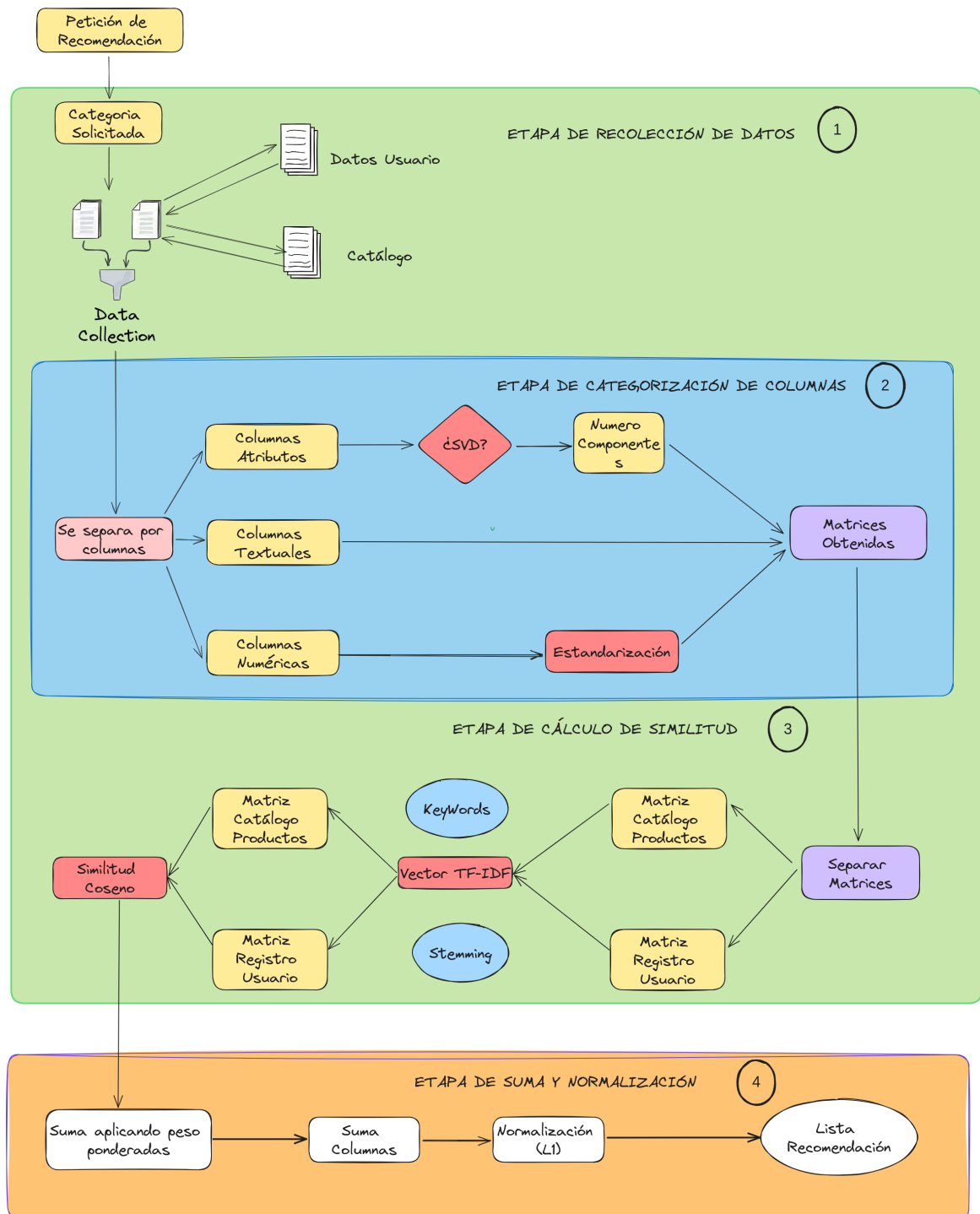


Figura 3.28 Diagrama de estructura sistema de recomendación

3.3. Desarrollo de aplicación web

El desarrollo de la aplicación web conlleva diferentes procesos paralelos, desde la definición de los objetivos y tecnologías más adecuadas para ello, como la arquitectura y el cómo se

relacionará cada componente para entregarle al usuario la experiencia de una buena recomendación. Esto conlleva un proceso de levantamiento de requisitos hacia el cliente, identificando y analizando cual son las necesidades y funcionalidades que deben implementarse en el sistema, definiendo conexiones, relaciones y por lo tanto, beneficios y limitaciones, teniendo en cuenta que tan escalable debe o podrá ser el sistema.

3.3.1. Modelo de Dominio

La definición del modelo de dominio tiene como elementos los usuarios, productos, proveedores y la empresa, en cuales exista el registro de las compras que realiza la usuario con los datos de la empresa que pertenece, de los productos provenientes del catálogo de los proveedores, estando referenciado en la orden de compra ambas entidades Modelo (3.29). Dentro de la decisión de diseño del esquema, es importante destacar que no existe un atributo de cantidad en los productos, es decir, cada producto tiene su propio ID, esto con la finalidad de otorgarle una independencia unitaria a cada uno, para así realizar un seguimiento adecuado entre las transacciones y permita una mayor escalabilidad entre proveedores y plataformas.

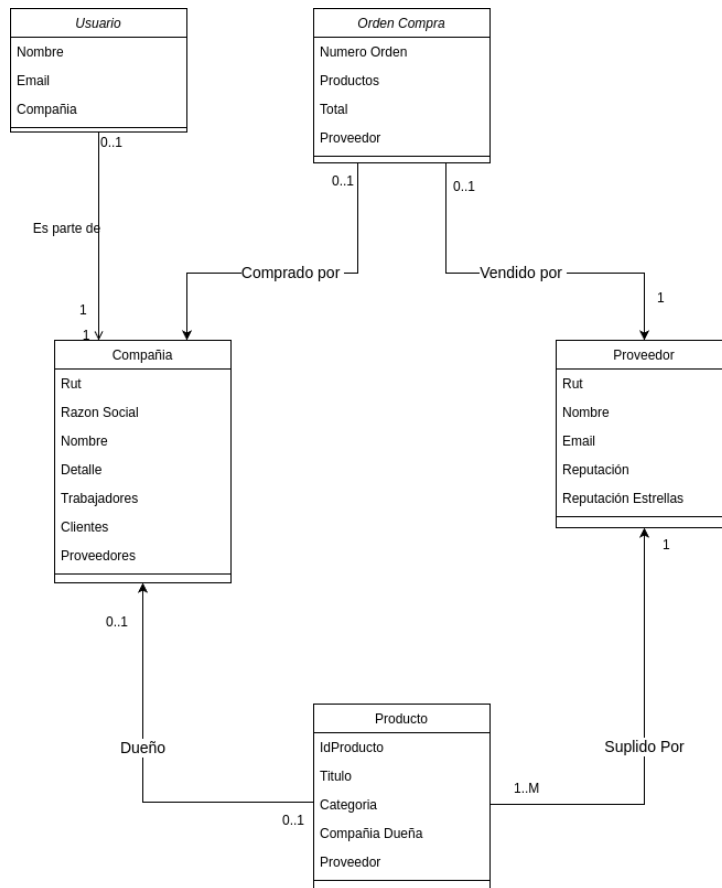


Figura 3.29 Modelo de Dominio de Sistema

3.3.2. Tecnologías

Las herramientas seleccionadas para el proyecto se centran en el stack MERN (MongoDB, Express, React, y NodeJS) Figura 3.30, con el objetivo de garantizar una experiencia de usuario fluida y un alto rendimiento. El desarrollo de la interfaz se realizó utilizando React junto con herramientas de diseño y persistencia como Material UI y Redux. Se abordaron diversas vistas esenciales para un comercio electrónico, como la visualización del catálogo de productos, configuraciones de la empresa, detalles históricos en el panel de control, visualización de registros individuales y la capacidad de compra. El servidor, desarrollado en Express, integra el sistema de recomendación creado en Python con FastAPI (Figura 3.31), se comunica con los servicios de Mercado Libre y gestiona los modelos de registros, inventario y catálogo para proporcionar información relevante al usuario.

La decisión de diseño de desarrollar el sistema usando Python, recae en la gran cantidad de librerías, herramientas y documentación existente para aplicar técnicas de análisis de datos, pre procesamiento, machine learning y manejo de datos (csv, excel, json, etc), siendo Python el lenguaje adecuado para estas funcionalidades.

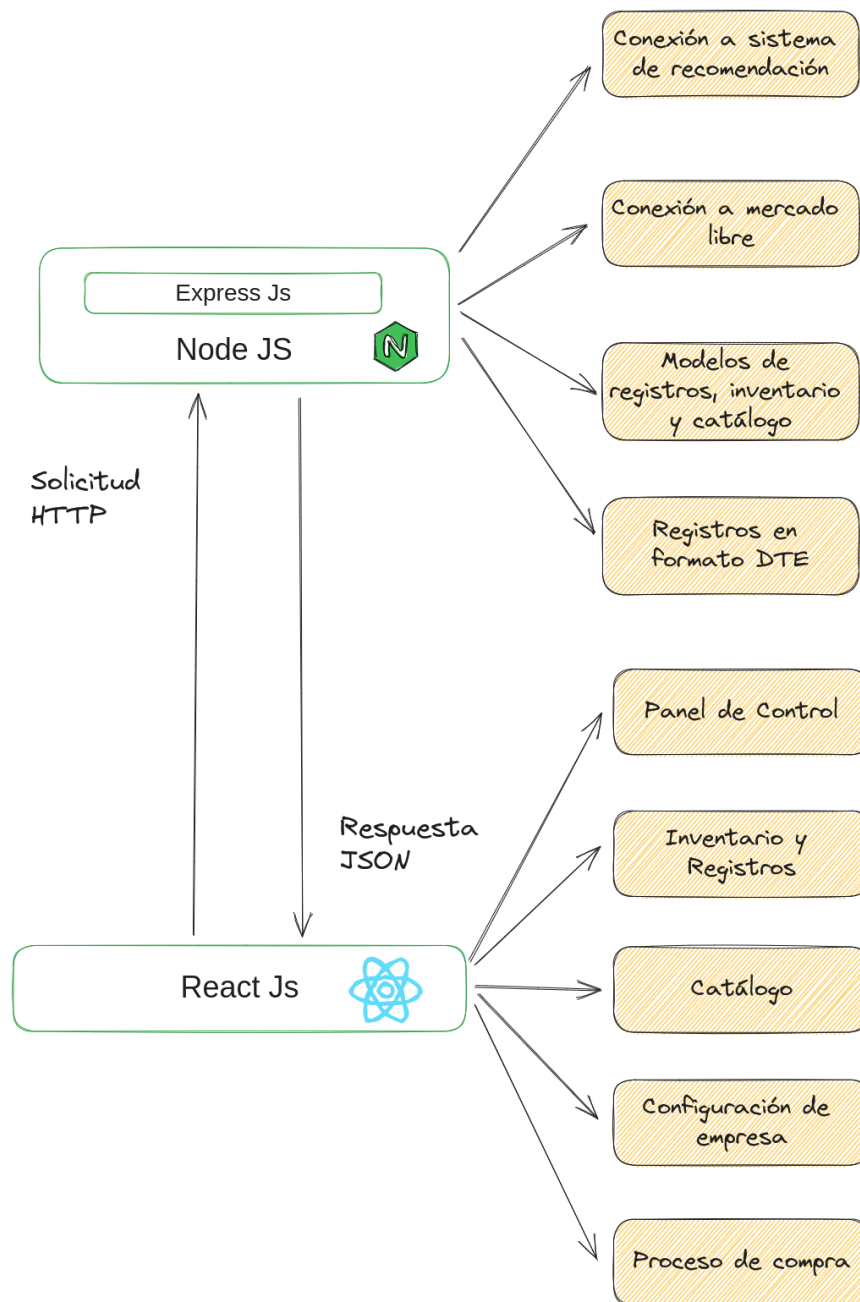


Figura 3.30 Diagrama que detalla la estructura de tecnologías MERN

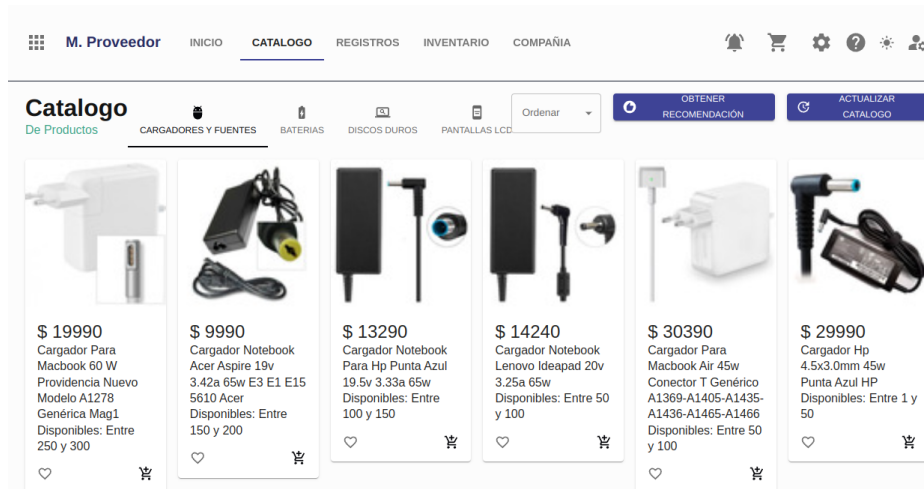


Figura 3.32 Imagen de interfaz de catálogo de los proveedores

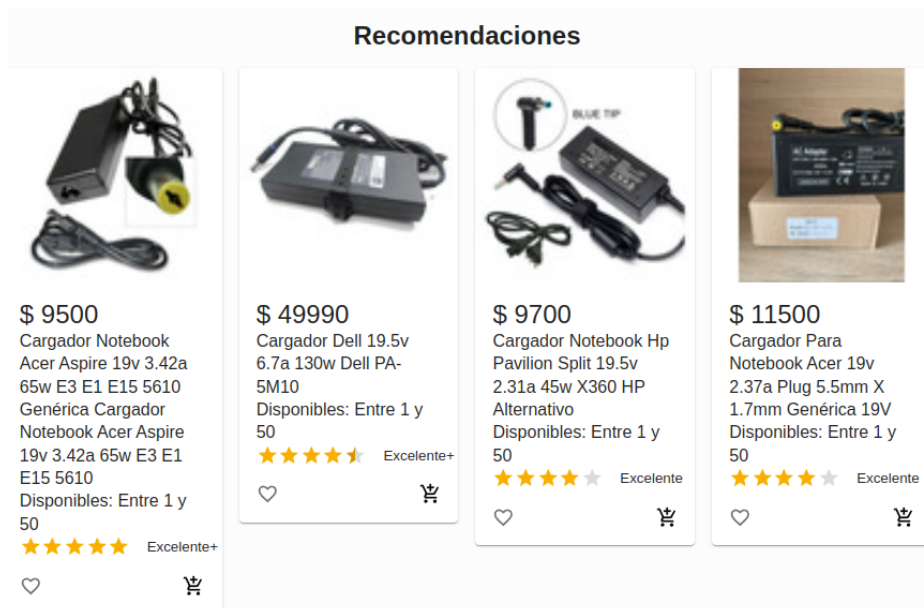


Figura 3.33 Imagen de interfaz de catálogo al obtener una recomendación

Obtener recomendaciones

Si desea obtener recomendación de productos, por favor ingrese la cantidad de productos que desea visualizar, además si desea agregarle prioridad según su importancia. **¡OJO!, que se tendrá en cuenta la prioridad que se definió en la encuesta inicial, si desea configurarla, vaya a la sección de configuración de compañía.**



¿Aplicar Stemming?

Número de componentes

0

Cantidad de productos

Cargadores y Fuentes ▼

AGREGAR PALABRA CLAVE

SALIR

OBTENER RECOMENDACION

Figura 3.34 Imagen de la interfaz que detalla la selección de recomendación

Cargador Lenovo 20v 3.25 Punta Cuadrada / Rectangular



Precio unidad **\$ 9990**
Precio total **1 x \$ 9990**

— +

ELIMINAR

Cargador Para Macbook 85 W Providencia Nuevo Modelo 1398



Precio unidad **\$ 21990**
Precio total **1 x \$ 21990**

— +

ELIMINAR

Total **\$ 92360**

LIMPIAR CARRO **INICIAR PAGO**

Figura 3.35 Imagen del carrito de compra de un usuario

El manejo de registros de compras y ventas, además de la visualización de los productos que se encuentran actualmente en la empresa, resulta necesario desarrollar una Tabla de productos y registros en donde muestre información clave como la fecha, precio, producto, stock actual y el proveedor a quien se le compro el ítem. Esto se puede visualizar en las siguiente Figuras 3.36 y 3.37, en donde resulta crucial la capacidad de seleccionar, filtrar y ordenar según diferentes opciones para el usuario.

Inventario
De Empresa

AGREGAR PRODUCTO TOMA DE INVENTARIO

☐	SKU	Nombre	Categoría	Precio ↑	Stock	Proveedor
☐	MLC615322065	Cargador Notebook Acer 19v 2.1a 5.5*1.7mm Alternativo	MLC9240	7090	1	GLOBALKERRY.CL
☐	MLC981083215	Bateria Compatible Con Samsung S8	MLC5068	7900	1	FIXMOBILEVIÑA
☐	MLC518800936	Cargador Para Dell 19.5v 4.62a Inspiron Vostro Latitude	MLC9240	8325	1	GLOBALKERRY.CL
☐	MLC518395062	Cargador Notebook Hp Pavilion Split 19.5v 2.31a 45w X360	MLC9240	8990	1	PCTECNOCL
☐	MLC929084184	Cargador Notebook 19v 3.42a Toshiba Satellite Packard Bell	MLC9240	9890	2	GLOBALARCADE

Rows per page: 5 1-5 of 29

Figura 3.36 Imagen que detalla como se muestra el inventario de la empresa

Registros
De Empresa

CARGAR REGISTROS EDITAR REGISTROS

☐	Fecha	Proveedor	Cantidad Productos	Total	Estado Pago	Estado Envio
☐	2023-11-16T01:28:03.524Z	GLOBALKERRY.CL	1	8325	Pagado	Enviado
☐	2023-11-16T01:28:03.524Z	FLEXACOMP LTDA.	1	49990	Pagado	Enviado
☐	2023-11-16T01:28:03.524Z	PCTECNOCL	1	8990	Pagado	Enviado
☐	2023-11-16T01:28:03.524Z	AS1618727	1	12769	Pagado	Enviado
☐	2023-11-16T01:28:03.524Z	PCROCK_COMPUTACION	1	21000	Pagado	Enviado

Rows per page: 5 1-5 of 57

Figura 3.37 Imagen que detalla como se muestran los registros de compras de la empresa

Para finalizar, resulta necesario para la empresa y un usuario de este sistema poder visualizar un resumen histórico que permita visualizar el registro en un rango de tiempo (ya sea mes, año o semana), por lo que, en la vista de dashboard (en la Figura 3.38) detalla la cantidad de compras realizadas por la empresa en el actual año. Por otro lado, en la Figura 3.40 se muestra la información de la compañía, tanto el nombre, rut y otra información (los datos sensibles no se muestran), lo que da la opción de mostrar otra sección de preferencias (como en la Figura 3.41 en cual la empresa puede definir sus preferencias en la compra de productos.



Figura 3.38 Imagen que muestra la información de los registros de la empresa a través de un dashboard

The 'Configuración' page is accessed via the 'COMPAÑIA' tab. It includes a sidebar with links for 'COMPAÑIA', 'PREFERENCIAS', 'PRODUCTOS', and 'USUARIOS'. The main content is divided into two sections: 'Datos contacto' and 'Datos Formales'. The 'Datos contacto' section includes fields for 'Nombre' (CGTEC), 'Rut' (19291239), 'Ciudad' (Illapel), 'Direccion' (Constitution 234), and 'Email' (atompos@gmail.com). The 'Datos Formales' section includes fields for 'Detalle' (Detalle de la empresa) and 'Trabajadores' (3). An 'ACTUALIZAR DATOS' button is located at the top right of the configuration area.

Figura 3.39 Imagen que detalla las configuraciones y preferencias de la compañía

M. Proveedor INICIO CATALOGO REGISTROS INVENTARIO **COMPANIA**

Configuración [De Empresa](#) [ACTUALIZAR DATOS](#)

Datos contacto
Datos actuales de la empresa

Nombre
CGTEC

Rut
19291239

Ciudad
Illapel

Direccion
Constitucion 234

Email
atompos@gmail.com

Datos Formales
Datos actuales de la empresa

Detalle

Detalle de la empresa

Trabajadores
3

Figura 3.40 Imagen que detalla los datos de la compañía

Preferencias [De Empresa](#)

Precio
¿Qué tan importante es el precio para la empresa?

El precio es un factor importante para la empresa, ya que se busca que el producto sea de calidad y a un precio accesible.

★★★★★

Velocidad
¿Qué tan importante es la velocidad para la empresa?

La velocidad es un factor importante para la empresa, ya que se busca que el producto sea entregado en el menor tiempo posible.

★★★★☆

Confianza
¿Qué tan importante es la confianza para la empresa?

La confianza es un factor importante para la empresa, ya que se busca que el producto sea entregado en el menor tiempo posible.

★★★☆☆

Calidad
¿Qué tan importante es la calidad para la empresa?

La calidad es un factor importante para la empresa, ya que se busca que el producto sea entregado en el menor tiempo posible.

★★★★★

Figura 3.41 Imagen que detalla las preferencias de la compañía

Capítulo 4

Resultados y Análisis

En este capítulo se expondrán los resultados obtenidos en esta memoria, presentando las diferentes configuraciones experimentales que ayudarán a visualizar información valiosa relacionada a los productos y categorías establecidas en el sistema y además el aporte que entrega cada técnica computacional descrita previamente sobre estos. Además, se tienen en cuenta ciertas consideraciones y previo análisis que justifican los parámetros ocupados para los experimentos. Dentro de esto, resulta importante tener en cuenta que el análisis se realiza principalmente por cada categoría, es decir, se explorará como cambia los resultados de las técnicas con la misma información inicial (registro de la empresa), tanto en el orden de los productos, como en la cantidad de valores atípicos, etc. Para realizar análisis entre diferentes categorías resulta vital tener en cuenta este hecho, además de que ciertos registros al no tener información de atributos, impide una correcta comparación; para ello, se definieron parámetros iguales para todas las categorías (aún si no son los ideales) y se explora los resultados obtenidos los cuales fueron normalizados, lo cual nos permite llevar los valores obtenidos a una misma escala.

El primer acercamiento fue explorar la varianza, tiempo de ejecución, similitud de atributos y general obtenida por cada categoría usando la técnica de reducción de dimensionalidad, para lograr determinar que la cantidad de número de componentes a ocupar en todas las categorías es de dos componentes para SVD. Se debe prestar atención a este factor, ya que ciertas categorías no cuentan con la características de atributos, por lo que, no existe un número de componentes ideal a aplicar (con poco cambio en los resultados).

Las Tablas que también serán parte de este análisis, reportan un set de palabras obtenidas para cada categoría, en donde se diferencian las 20 palabras más relevantes y las 20 menos relevantes según cada vector calculado con cada técnica (un vector para los atributos y otro para el texto en general del producto). Con motivo de legibilidad y orden, se sitúa primero los resultados para los atributos, la cual, tendrá menor cantidad de palabras en general.

Los resultados son obtenidos son analizados a través de un test con la empresa, en donde se explora cuales son los productos que se consideran más relevantes para ellos, con la finalidad de obtener una métrica de evaluación por cada uno, denotando un bajo o alto rendimiento (medir la precisión y cantidad de falsas recomendaciones); esto se acompaña con la medición de recalls, o cantidad de productos recomendados que son parte del registro inicial, que nos entregará que tan cerca está la recomendación obtenida. Luego, se medirá también la distancia

que se obtiene de cada conjunto de recomendación, con el fin de cuantificar cuanta diferencia de similitud existe entre los productos obtenidos y así visualizar que técnica acopla o aleja los productos más cercanos.

Para finalizar, la siguiente Tabla entrega parámetros vitales para los siguientes experimentos, como las palabras claves que la empresa le resulta interesante analizar para el sistema y que se ocuparán 4 valores significativos para los resultados obtenidos.

Categoría	Cantidad Productos	Registros Usuario	Cantidad Atributos	Palabras Claves
Cargadores y Fuentes	1000	8	45	Dell, Original
Baterías	1000	4	36	Calidad, Adhesivo
Pantallas LCD	1000	7	40	Samsung, Garantía
Discos duros y SSD	1000	3	60	SATA, Interno

Tabla 4.1 Tabla de datos según categoría

4.1. Gráficos de Varianza y Tiempo

Los primeros gráficos a analizar se proporcionan al ir obteniendo la cantidad de varianza que se obtiene del algoritmo de SVD con cierta cantidad de componentes, que permite visualizar la cantidad de información que se captura al ir variando el parámetro del número de componentes (ver Figura A.1), apreciándose valores bastante parecidos para las diferentes categorías, siendo los resultados de tanto su versión acumulativa y derivada que denotan un aumento en la cantidad según el número de componentes. Los que entregaron información vital resultan ser la cantidad de outliers o valores atípicos (ver Figuras A.13 y A.14) en cual para ambos gráficos la categoría de baterías y discos duros y SSD denotan muy poco cambio en su valor de similitud obtenida de los productos; sin embargo, para las demás categorías apreciamos que a medida que se aumenta la cantidad de número de componentes, el rango obtenido del valor de similitud se acorta, generando así mayores valores atípicos (tanto para cargadores como para pantallas), teniendo esta última una gran suma de outliers; por otro lado, para similitudes menores que el primer cuartil, al aumentar la cantidad de componentes, el rango del cuartil aumenta arraigando estas similitudes.

Para estos gráficos se logra determinar un número que será parte del parámetro de los siguientes gráficos obtenidos, dado que la categoría de "Cargadores" obtiene valores bastante depurados con dos números de componentes, definiéndose como parámetro a ocupar.

4.2. Tabla de vector TF-IDF

Como se explica en el capítulo anterior, el principal mecanismo para calcular similitudes entre productos es el uso y cálculo de usar y calcular la frecuencia de los términos que se encuentran tanto en los productos del catálogo como en los productos del registro de compra. Esta técnica le asigna un valor de relevancia a cada término, siendo los términos con mayor valor, los que más representan a un producto, siendo necesario analizar y visualizar como afecta principalmente la técnica de stemming a las palabras, por lo que el listado de palabras no cambia en gran medida son las demás técnicas exploradas.

En la Tabla obtenida para cargadores (ver Tabla B.1) se muestran el set de palabras más relevantes encontradas para la categoría de cargadores son las relacionadas con su modelo de adaptador o de voltaje, siendo mayormente códigos que denotan la versión que es adaptable el producto, como lo son las palabras con el formato 'e1-422-3431', entre otras; por otro lado, las menos relevantes la lideran palabras como 'nuevo', 'hp', 'lenovo' y palabras comunes como la versión del computador y la marca de este mismo. Esto significa que los términos que son mayormente compartidos en esta categoría, como lo son las marcas de los computadores se verán menos beneficiados, siendo menos relevantes e importantes para el cálculo de similitud entre los productos; se suman palabras que define la misma plataforma de Mercado Libre, como son: 'immediate_payment', 'credits_profile', que son características que se entregan al marcar ciertas casillas en el registro del producto, además información compartida como la región, indica la centralización de éste tipo de productos de manera geográfica.

En la transformación obtenida con el uso de stemming, vemos que ciertas palabras escritas en inglés como 'credits_profile' no sufren una transformación, dada su nula presencia de sub-fijos como "ía" o "ias"; por otro lado, la palabra boleta se le elimina la última sílaba dado que puede haberla detectado como un verbo mas que un sustantivo, generando así un corte.

En Tabla obtenida de palabras para baterías (ver Tabla B.2) sucede algo similar a la categoría anterior, las palabras que son mayormente especificaciones resultan con una mayor relevancia, dada su misma naturaleza de ser únicas en las descripciones de los productos, aunque se aprecia que para la sección de stemming marcas y especificaciones que aún así fueron identificadas como menos relevantes. Para los resultados obtenidos en la segunda Tabla (ver Tabla B.2), se aprecia un cambio sustancial en las palabras más relevantes obtenidas con stemming, dado que el patrón de solo obtener especificaciones no ocurre en esta situación, siendo identificadas palabras como 'correspondiente', 'coquimb', 'copiap' que hacen referencia a la localidad y ciertas descripciones; esto hace entender, que las palabras obtenidas en el apartado de sin técnicas, fueron transformadas en (eliminandose la últimas letras en ocasiones) y/o las palabras más relevantes obtenidas se lograron diferenciar de su contraparte de otros productos, siendo lo suficientemente únicas para obtener una gran valorización en este apartado.

En la Tabla de palabras de los atributos de las pantallas (ver Tabla B.3) para el caso de los resultados de la Tabla de atributos, se visualiza que los modelos y marcas de los celulares que son parte estas pantallas, como: "redmi", "motorola", "huawei", etc. Esto se refleja también en los resultados obtenidos con stemming, en donde la naturaleza de las palabras no varía de manera sustancial. Por otro lado, en los resultados obtenidos para la Tabla de columnas principales (ver Tabla B.3) el cambio que se genera de las palabras obtenidas en sin técnicas con stemming son muy diferentes, siendo un patrón totalmente diferente a los casos anteriores, ya que sin stemming, las palabras más relevantes son estándares o más comunes de lo que se esperaría, detallando que estas características no se ven mucho en estos productos. Para el caso de transformación a sus raíces de ciertas palabras, se identifica algunas que se obtuvieron con caracteres como guión bajo "_hac" o algunas con guión simple "ldn-lx3", que resulta claramente en una mayor relevancia dentro del vector, dada su unicidad y particularidad dentro del producto identificado; por otro lado, para las palabras menos relevantes, si bien no se mantuvo en su totalidad las mismas palabras que sin técnicas, se mantiene el mismo conjunto de palabras menos relevantes identificadas.

Para el caso de las Tablas obtenidas para la última categoría (ver Tabla B.4), aplica patrones parecidos vistos con categorías anteriores, en donde la descripción de los modelos de los productos y descripción muy particulares son identificadas como mayormente relevantes, identificándose palabras que no se logra identificar el origen o la raíz que se intenta calcular con el uso de stemming. Para la Tabla con las características generales de textos B.4 el uso de stemming obtiene palabras con caracteres especiales vistos anteriormente, además de transformar palabras como `rm_metropolitana`, en cual fue limpiada por el algoritmo propio, cortando la palabra `metropolitana` a `metropolitan`, eliminando así la última sílaba. Además, que las palabras identificadas como menos relevantes son algunas propias del registro inicial del usuario, como la marca `"western"`, la abreviación de disco sólido `"ssd"` también es una no relevante, acompañado de ciertas marcas como `"kingston"`, `"toshiba"`, y su color como `"green"`, `"blue"`, `"purple"`.

4.3. Tabla de Resultados

Dentro del Apéndice C podemos encontrar los resultados de los 10 primeros productos obtenidos según técnica y por categoría, que se complementa con los gráficos obtenidos en el apéndice continuo, en donde mediante gráficos de box plot se obtiene la cantidad de outliers o valores atípicos presentes en cada tipo de características calculadas en el sistema. Además, las gráficas de barras presentes en el siguiente apéndice y las Tablas obtenidas desde el test que se realizó a la empresa, para este contexto, nos ayudará a visualizar cual de los productos son parte del registro inicial del usuario.

Como dato a tener en cuenta, que debido a que las categorías de Baterías y Discos Duros no poseen información de atributos en sus registros iniciales, no se realizará algún tipo de análisis en el uso de la técnica de reducción de dimensionalidad, dado que, los resultados no varían respecto al caso de base de no aplicar técnicas al sistema de recomendación.

4.3.1. Cargadores

Para los resultados de la primera categoría y sin técnicas (ver Tabla C.1), se observa que los productos primero, tercero y cuarto recomendados son parte del registro inicial del usuario (ver Tabla F.1), mostrando mayor similitud en características de texto. No obstante, desde el aspecto numérico, el segundo producto tiene incluso valores negativos en similitud y cercanía debido a la variabilidad en la calidad, vendedor y estado del producto. La media numérica se ve afectada por la reputación de los vendedores, provocando fluctuaciones en los valores. En la Tabla obtenida del test con el usuario, se muestra una precisión del 70 %, un 100 % de recall y una distancia promedio de 4.0772. Esto indica que la precisión es alta, aunque la variabilidad numérica puede influir en la percepción de la calidad de las recomendaciones.

Con el uso de la técnica de stemming, el orden de estos mismos productos se mantiene en la misma posición, sin embargo, se ve un aumento general de la similitud obtenida especialmente en la categoría de atributos, el cual sin técnicas tiene una cantidad de outliers de 47 (ver Figura D.1) que ocupando stemming disminuye hasta ser solamente 7 productos con valores atípicos a los obtenidos. La similitud de texto también se ve beneficiada en valores pero con una menor magnitud, siendo disminuida la cantidad de outliers de 59 a 46. Sin embargo, a pesar de la mejora de los valores obtenidos de similitud, la cantidad de valores atípicos no cambia aplicando stemming (siguen siendo 8 productos) por lo que en este aspecto, no se

aprecia una mejoría medible; tampoco en la Tabla de resultados obtenidas, dado que a pesar de que la distancia de las similitudes aumenta (de 4.0722 a 4.5443), la precisión obtenida y el recall no varía, se mantiene en un 60 %.

Al aplicar la reducción de dimensionalidad, el producto que antes ocupaba el primer lugar (Cargador Notebook Acer) desaparece del listado de recomendados, siendo reemplazado por un cargador con una palabra clave especificada por el usuario. El único producto que se mantiene en el top-10 es otro cargador Dell, cuya similitud de texto aumenta significativamente (de 0.9226 a 1.2981). Según los gráficos de box plot (ver Figura D.4), los valores atípicos aumentan de 8 a 89 productos, debido a la valorización y peso artificial de las palabras clave. Esto altera el uso de vectores TF-IDF, ya que el peso ya no se basa en la frecuencia de aparición. La precisión aumenta un 10 %, pero el recall cae drásticamente a solo uno, y la distancia promedio sube de 4.0722 a 5.5961.

La siguiente técnica a analizar es el uso de una reducción de dimensionalidad, en la cual la cantidad de productos que se mantienen del registro inicial se rebaja también a uno solo, siendo en este caso el principal producto visto en el caso base como en el uso de stemming, obteniendo también el primer puesto o la mayor similitud dentro de la Tabla de resultados, sin embargo, el aspecto de la precisión aumenta comparado con el stemming, en donde obtiene un valor de 70 % al igual que las palabras claves y la distancia sufre un aumento radical hasta un valor de 6.7489. Esta técnica reduce significativamente la gran cantidad de outliers visto en la similitud de atributos, pasando a no tener ningún valor atípico en este campo, lo que se traduce en ningún producto fuera de los cuartiles calculados en la similitud, siendo visualizado en el gráfico de box plot con un gran aumento de la similitud general obtenida, llegando hasta valores de 2.2520 (específicamente del único producto del registro inicial), esto generado por el aumento de los cuartiles y rango de la similitud de atributo calculada, sobrepasando los valores de 1.

Por último, la siguiente técnica se orienta a explorar la precisión que significa tener una eliminación de los valores atípicos de la similitud obtenida, aplicando un umbral basándose en este mismo cálculo de cuartiles del box plot. Para este caso, se obtiene una baja de la distancia promedio de los productos obtenidos (se obtiene 3.1072) que se acompaña de una baja drástica de la precisión de la recomendación siendo solamente de un 40 % con un 0 % de recall, es decir, sin ningún producto que es parte del registro inicial.

4.3.2. Baterías

Los resultados para la categoría de batería debe tener en cuenta la nula información que contiene el registro inicial del usuario en las características de atributos, además de que un solo producto registrado es parte del catálogo analizado, por lo que, los resultados son más acotados y ajustados.

Dentro de la Tabla obtenida para las baterías sin técnicas (ver Tabla C.2), podemos apreciar que este producto unitario parte del registro inicial del usuario, se sitúa en el primer puesto de la lista de recomendados, obteniendo un gran valor de similitud de texto con los registros mencionados. Este caso base, muestra en su gráfico de box plot (ver Figura D.5) una gran cantidad de valores atípicos en esta misma característica de texto, lo que lleva a un valor no ignorable de 46 outliers en la lista de recomendados, siendo estos primeros 10 productos parte de este conjunto. Respecto a los resultados en precisión y recall, solamente el 40 % de los productos

son considerados relevantes para el usuario, siendo el primero, el cuarto, quinto y el noveno de la Tabla; que son mayoritariamente productos del mismo modelo, es decir, una batería que es compatible para un iPhone 6s.

Para el caso de uso de stemming, apreciamos que los valores obtenidos de similitud en el texto aumenta hasta tener una media casi de 0.3 (ver Figura D.6), representando así un aumento del rango obtenido de la similitud general, en donde el mismo producto se mantiene como la principal referencia a entregarle al usuario. Sin embargo, a pesar del aumento de valores, la cantidad de outliers se mantiene igual, apareciendo un valor atípico menor, que posee un valor más bajo que lo obtenido de los cuartiles; esto se debe a este mismo aumento de valores, que significa, que para este caso la técnica de stemming entrega un mejor acercamiento en valores brutos de similitud, que se traduce en un aumento del 20 % en la precisión obtenida, obteniendo un valor de 60 %.

Al usar palabras clave para las baterías, se mantiene el producto inicial en primer lugar y aumenta la precisión a un 80 %. Sin embargo, la cantidad de outliers es alta, con 73 valores atípicos, lo que indica sobreajuste y un aumento en la distancia promedio de 2.82 a 3.03. El segundo producto, con el mismo nombre que el primero, también se mantiene en esa posición, pasando del quinto puesto en la base al segundo debido a su similitud en el título y diferencias en calificación y precio. Esto resalta la influencia de las palabras clave en la precisión, el distanciamiento entre productos y la diferencia de casi 2000 CLP\$.

En el uso de palabras claves para la batería, se mantiene en primer lugar el producto que es parte del registro inicial, pero además se ve un aumento considerable en la precisión de los productos relevantes determinado por el usuario, logrando un valor de 80 % para esta técnica. Sin embargo, como se ve en las técnicas anteriores, la cantidad de outliers es significativamente alta, siendo para esta técnica muy aumentada la cantidad, obteniendo un valor de 73 valores atípicos usando palabras claves; esto se traduce que, a pesar de tener una gran precisión, el sobre ajuste presente en los resultados remarcan que los productos que contienen palabras claves, genera un distanciamiento entre los productos, aumentando de 2.82 que era sin técnicas, hasta un valor de 3.03 con palabras claves. Además, se aprecia que el segundo producto que contiene el mismo nombre que el primero, mantiene su segundo puesto, siendo también su puesto para el stemming y pasando de ser el quinto en el caso base a este puesto específico; esto dado por su gran parecido en el título del primer puesto, pero que lo diferencia los valores de calificación del producto y el vendedor, y su propio valor de precio, teniendo el primero un valor monetario de 10490 pesos chilenos, al segundo que es de solamente 8990 pesos chilenos.

Para finalizar, el aplicar el umbral y eliminación de valores atípicos, se visualiza una baja de precisión de productos relevantes para el usuario, siendo productos que no sobrepasan los 0.8 de similitud con el registro inicial, siendo estos generalmente de marca iPhone, pero no del modelo de interés para la empresa, siendo el primer puesto una batería de Samsung, la cual no fue relevante para este mismo.

4.3.3. Pantallas

Para la categoría de pantallas, se debe tener en cuenta que es la categoría con más productos presentes en su catálogo, siendo de 4 productos de los 7 registrados por la empresa; además que presenta un porcentaje de 48 % de valores nulos en el catálogo y es la catego-

ría con mayor distancia de similitud entre los productos, llegando un promedio de casi 12 en distancia euclídeana; detalles que permitirá una mejor comprensión de los resultados e interpretaciones.

En el caso base (ver Tabla C.11), tres productos del registro inicial tienen alta similitud, ocupando el primer, cuarto y quinto lugar; todos relacionados con características de mica y marcas Samsung o Redmi. Esto resulta en un recall del 75 % y una precisión del 70 %, superior a otras categorías. En el gráfico de box plot (ver Figura D.9), se observan 68 valores atípicos en texto, con 18 productos superando el cuartil superior. El primer producto muestra una similitud de 3.7193, significativamente mayor que el segundo con 2.5192, debido a la prevalencia de productos Samsung en el registro inicial. Además, otros productos con títulos similares, como el quinto puesto, se ven afectados por diferencias en las características numéricas, especialmente en la reputación del vendedor (0.83 versus 0.66).

Al aplicar la técnica de stemming a esta categoría, se mantiene el sobre ajuste del primer producto, manteniendo el primer puesto de la Tabla, siendo el único con una similitud mayor a 3. Por otro lado, la precisión de la Tabla disminuye en un 10 % y el recall se mantiene en un 75 % con los mismos productos, sin embargo, los puestos de estos se mejoran, pasando el cuarto y quinto a ser el segundo y el tercero, representando una mejora en los resultados obtenidos con stemming (a pesar de la baja de precisión). En los resultados del gráfico de box plot (ver Figura D.10) se aprecia que la cantidad de outliers no cambia significativamente, siendo de solo dos productos la disminución de valores atípicos, principalmente vistos en el aspecto de atributos, el cual se disminuyó de 25 outliers a 6 usando stemming.

Para la técnica de palabras claves, se usaron las palabras: "Samsung" y "Garantía", que como se visualizó anteriormente, son los productos que mayormente han sido parte de la Tabla de resultados, pero han sido opacados por otros productos sin estas características; para este caso, podemos apreciar que el tercer producto de stemming ahora es el segundo puesto, pero ya no es la misma cantidad de productos en el recall, específicamente la pantalla con marca Redmi ya no aparece en el top 10 de productos recomendados usando palabras claves, ya que mayoritariamente de productos samsung. Esto se traduce, en que el peso asignado por el algoritmo usando palabras claves, genera que las palabras menos relevantes como garantía, significa que los productos que comparten esta característica (que es la mayoría) tengan un mejor valor de similitud, por ende, la similitud de texto sea más extendido los cuartiles y la diferencia sea menor entre los productos. Por otro lado, los resultados obtenidos con el test con la empresa, aumenta significativamente la precisión de 70 % a 80 %, pero con una baja significativa del recall de solo 2 productos (50 % de recall).

Con el uso de la reducción de dimensionalidad, se aprecia un aumento drástico de la distancia de similitudes entre los productos, siendo de 7 del caso base hasta un 15 de distancia euclídeana, que se representa en el cálculo de los cuartiles del gráfico de box plot ??, en donde principalmente la similitud de atributos se ve acotado de manera tajante, siendo ínfima el rango de valores que son parte de la media, pasando de 25 outliers a 173 usando SVD. Por otro lado, dentro de los resultados realizados al test con la empresa, se ve una baja de precisión de 10 % y de recall a comparación con el caso base, pasando de 75 % a solamente un 50 %, en donde los productos alcanzan una diferencia radical de puestos, siendo el primero este producto sobre ajustado de pantalla samsung, pero el producto de redmi su valor de similitud

aumenta, pero su diferencia de similitud no es tanta pero suficiente para posicionarse en el último puesto.

Por último, los resultados visto eliminando los valores atípicos del caso base, se aprecia una disminución radical de un 70 % de precisión a solamente un 10 % de relevancia en el top-10 listado en la Tabla, que se acompaña con una disminución de la distancia entre similitudes, en donde el único producto relevante fue una pantalla samsung con características medianamente parecidas a las vistas del producto ajustado en técnicas anteriores.

4.3.4. Discos Duros y SSD

La categoría de discos duros es el que tiene un registro menor a comparación de las otras categorías, siendo solamente de 3 productos de misma marca y especificaciones parecidas (marca Western Digital) los cuales no tienen información en sus características de atributos y especificaciones, sin embargo, solamente uno pertenece al catálogo de productos analizados, por lo que su recall solo varía entre 0 y 100 %. Además, dentro de las 4 categorías, los discos duros es quien presenta el mayor valor de porcentaje nulos, siendo más de la mitad de la matriz sin información adjunta.

Para la Tabla de resultados con el caso base (ver Tabla C.16), se observa que el único producto del registro inicial ocupa el primer puesto con una gran diferencia, con un valor de 1.5 en similitud de texto y 1.81 en similitud general. Este producto es uno de los dos outliers presentes en esta técnica, destacándose especialmente en la similitud textual. En el test con el usuario, se obtiene una distancia euclidiana de 4.2, que se utiliza para medir la separación en el espacio de características entre los productos recomendados y el registro inicial, proporcionando una forma intuitiva y cuantitativa de evaluar las diferencias. Además, se observa una precisión del 50 %, incluyendo los outliers seleccionados por la empresa, así como el cuarto, sexto y séptimo producto, todos discos "Western Digital". El único producto que no es de esta marca en la lista es el octavo, que presenta un alto valor de similitud numérica.

Con el uso de la técnica de stemming, se aprecia en el gráfico de box plot (ver Figura D.14) no cambia sustancialmente, manteniéndose la misma cantidad de valores atípicos en ambos aspectos (textual y general), y siendo los mismos productos determinados como relevantes para la empresa (con la misma precisión de un 50 %), con la diferencia que dos productos cambiaron de puesto, específicamente el cuarto y sexto puesto, ahora suben su puesto a ser el tercero y el cuarto puesto, en donde se ve una importancia a los productos que contenga mayor información específica en sus títulos, como se aprecia en los tres primeros puestos, en donde sus especificaciones la componen palabras largas y complicadas como: "Wds500g2b0a." "Wds480g2g0a".

En el caso de uso de palabras clave "sata" e "interno", la distancia entre similitudes aumenta de 4.23 a 5.75 y la precisión de productos relevantes disminuye un 10 %. Sin embargo, no se identifican outliers en el box plot, indicando que las palabras clave eliminan satisfactoriamente los valores atípicos. La palabra "interno" impacta significativamente, ya que todos los productos la contienen, ampliando el rango de similitud textual y abarcando el único outlier permitido.

Para finalizar, el uso de la técnica de umbral nos entrega una precisión mucho menor a la lista de recomendados de técnicas anteriores, solamente alcanzando un 30 %, en donde solamente se eliminan los dos primeros productos identificados como atípicos. Además, reduce la distancia entre las similitudes de un 4.2 hasta un 3.8 en promedio.

4.4. Test del Sistema de Recomendación

En esta sección se explora como fueron los resultados obtenidos del test realizado con la empresa, en donde se les presentó los 10 productos de cada categoría con cada técnica, con la finalidad de obtener cual determinan ellos que es relevante para la empresa. Cada Tabla se puede encontrar en Apéndice F.1, que fue siendo referenciado en cada categoría anteriormente, pero en esta sección, se gráfico el promedio de precisión, recall, distancia y outliers por técnica y categoría, con la finalidad de obtener una mejor visualización del panorama calculado.

4.4.1. Por categoría

En estos gráficos se puede apreciar las diferencias que obtuvieron cada categoría, partiendo por la distancia promedio obtenida por cada categoría, en donde pantallas presenta una clara diferencia de valores con las demás, siendo esta distancia menor en baterías. Por otro lado, en la cantidad de valores atípicos estas dos categorías presentan una gran diferencia a comparación con discos, la cual presenta un promedio de dos outliers en general.

Dentro del gráfico de promedio de precisión y recall por categoría, se visualiza, que baterías y discos presentan un promedio de recall muy cercano a 1, dado que solamente un producto de su registro inicial pertenece al catálogo analizado, por otro lado, cargadores y pantallas (este último quien tiene mas en catálogo) muestran un recall bastante cercano a 0.6. En el aspecto de la precisión, quien presenta una mejor precisión es la categoría de cargadores, siendo exactamente un 60 %, siguiéndole pantallas, baterías y discos al final.

Promedio Distancia por Categoría

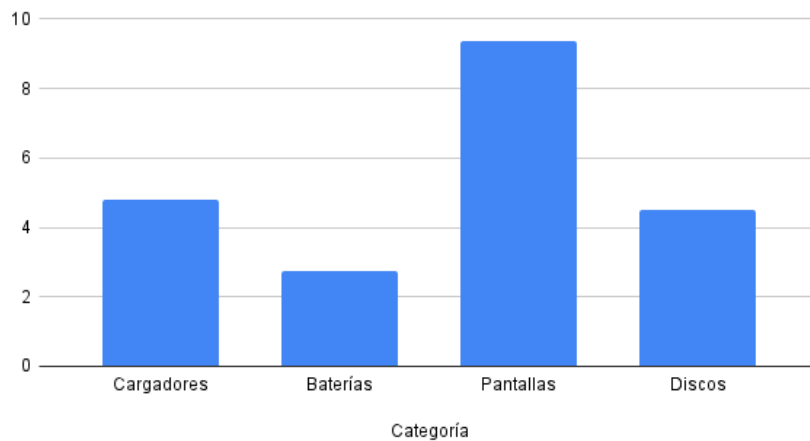


Figura 4.1 Gráfico de barras de promedio de distancia por categoría

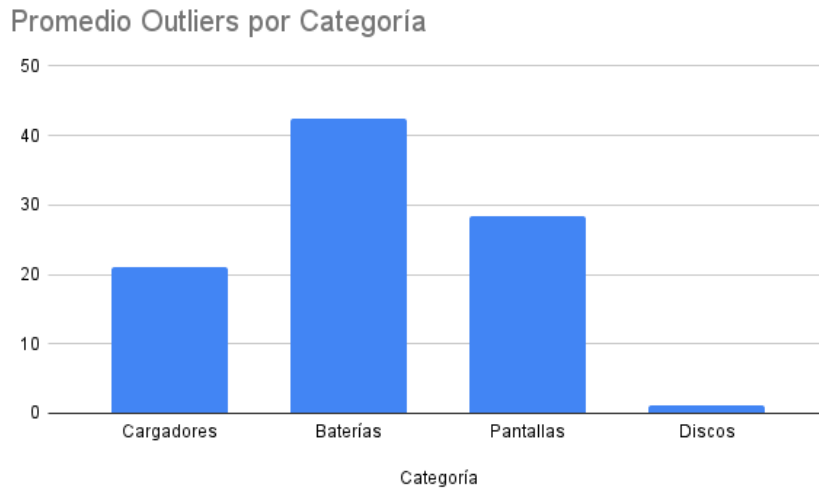


Figura 4.2 Gráfico de barras de promedio de outliers por Categoría

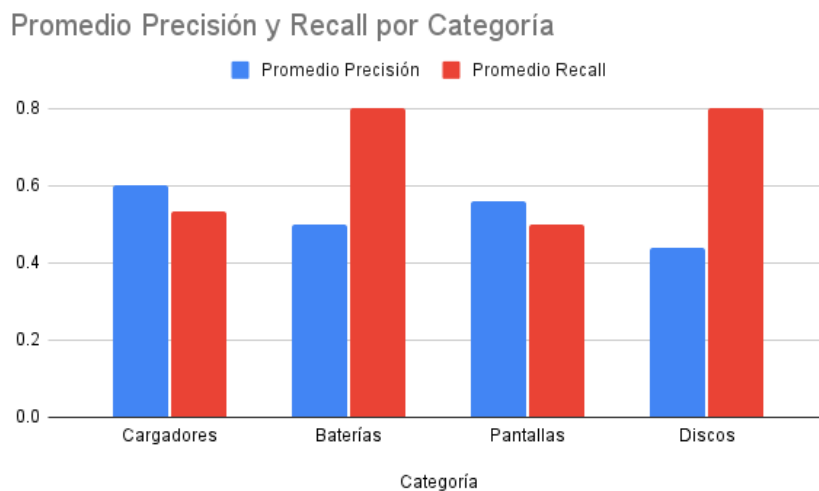


Figura 4.3 Gráfico de barras de promedio de precisión y recall por categoria

4.4.2. Por técnica

En estos gráficos se aprecia como cambia estas medidas según técnica, tomando el promedio obtenido entre todas las categorías, en el gráfico de distancia se observa que la técnica que presenta una mayor distancia en promedio es el SVD, siendo el menor el uso de un umbral eliminando los outliers. En el gráfico de outliers por técnica, se visualiza que la cantidad de outliers de forma general aumenta con el uso de palabras claves y SVD, siendo el menor valor del stemming y umbral.

El último gráfico es sobre el promedio de precisión y recall por técnica, se visualiza que la técnica que representa una mejor precisión es de las palabras claves, junto con SVD y stemming

que comparten el valor promedio de precisión; también se destaca, que de forma general, el uso de palabras claves y disminuye notoriamente el recall de la lista de recomendados.

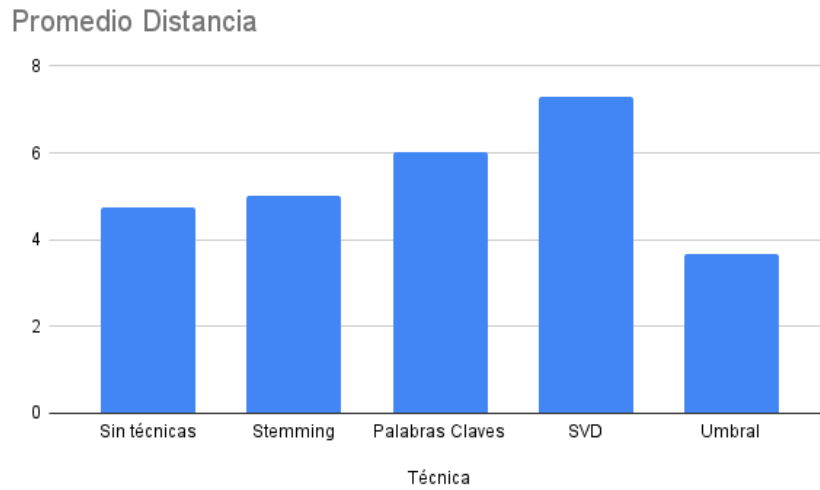


Figura 4.4 Gráfico de barras de promedio de distancia por técnica

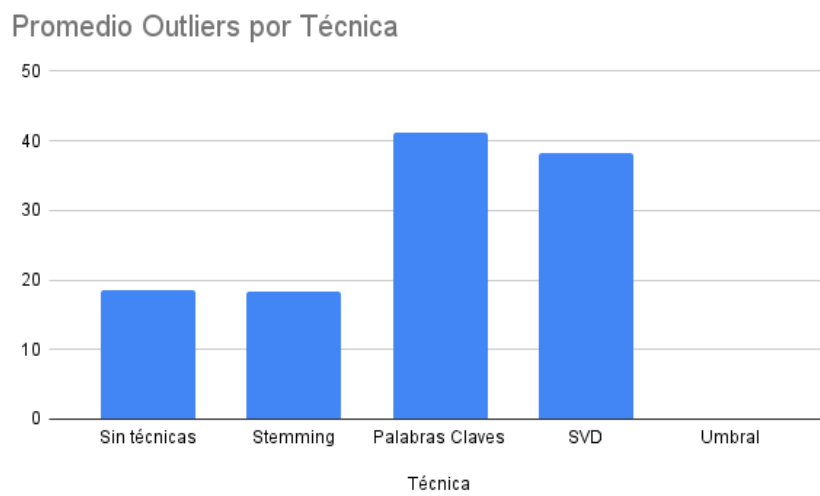


Figura 4.5 Gráfico de barras de promedio de Outliers por Técnica

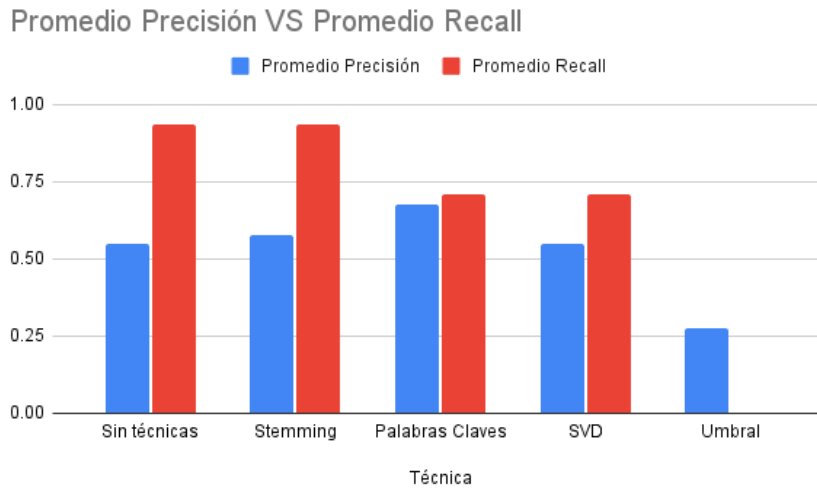


Figura 4.6 Gráfico de barras de promedio de precisión y recall por técnica

4.5. Análisis y Discusión

En esta sección se realizará una interpretación de los resultados vistos anteriormente, esto con el fin de estimar cual podrá ser la técnica adecuada y en que caso aplicaría mejor una sobre la otra.

Dentro de los aspectos de similitud, el aspecto numérico ha sido un valor que ha beneficiado y o perjudicado algunos productos, específicamente sirve para aquellos productos que comparten mucha información entre ellos (como el caso de las baterías) pero la valorización del vendedor, del producto, el precio, disponibilidad, ratings y los feedback recibidos de los compradores, altera los resultados para que se adecuen más al registro inicial obtenido de la empresa. Esto de primera forma, puede resultar beneficioso, ya que si el usuario/empresa tiene tendencias a comprar productos de cierta reputación o precio, se le recomendará productos similares a ellos, sin embargo, el problema radica en la existencia de descuentos o bajas de precios que se puede dar en este tipo de plataformas, lo que impediría al sistema, por mas que compartan información de especificaciones o de texto, recomendar productos que más le convengan al usuario. Para ello, el apartado numérico podría ser personalizado, es decir, el usuario podría seleccionar en sus preferencias cual es la información que quiere que se utilice en los cálculos, ya sea solamente la reputación, el precio del producto o el feedback; para ello, se podría aplicar un filtro en el listado total de recomendados, que le permitiría así ordenar según su propia preferencia.

Dentro del uso de la técnica de stemming en el sistema de recomendación entrega un cambio muy significativo en el tiempo de respuesta del sistema, incluso sobrepasando el tiempo permitido para las respuestas de la API en la conexión con la plataforma web, esto se muestra al realizar una petición en la plataforma online, que al elegir la técnica, se debe esperar un tiempo mayormente considerable en la respuesta.

Para el caso de las palabras mayor y menos relevantes obtenidas usando el vector TF-IDF, se repite el patrón de que identifica como relevantes las palabras mayormente específicas y com-

plicadas en su creación, palabras como “hdtb510xk3aa” demarcan modelos, especificaciones, adaptadores y otras características que contiene sólo algunos productos. Esto denota, que no existe un mérito o mecánica implementada en el vector que amortice o castigue a productos que puedan ocupar descripciones extensas, larguísimas que puedan irse extendiendo del contexto de la categoría y de este mismo. Por lo que los proveedores pueden agregarse más texto sin miedo a que su producto no se identificado como parte de una categoría u otra.

Con el uso de stemming que permite obtener en ciertos casos la raíz de la palabra, las que contienen caracteres especiales en algunas ocasiones no son del todo identificadas como raíz, dada que al tener caracteres que son parte de la palabra, no se puede directamente separarlas y analizarlas por separado, dado que perdería la cohesión y sentido de esta misma, como: “messages_as_seller”, de todas formas, si otro producto contiene la misma combinación de letras, se identificará con una misma valorización. Esto se amplifica cuando se agregan palabras de otro idioma en al descripción de los productos, lo cual genera que la biblioteca (usada en español) intente identificar estos subfijos tomados desde un lenguaje a otro aplicándole al inglés, generando así raíces inexistentes a palabras que ya lo eran, como sucede con la palabra “extended_warranty_eligible”, la transforma a la palabra “extended_warranty_eligibl”, en donde se identifica el subfijo lo identifica como la última sílaba.

De forma general, se puede determinar que usar stemming da un mejor acercamiento general entre los productos, dado que, a pesar de que las palabras específicas son las que son consideradas mas relevantes, el contexto de los productos que esta analizando resulta fundamental las especificaciones, modelos y otras características al momento de comprar, además que a pesar de ciertos cambios en las menos relevantes, siguen manteniendo ese patrón de ser palabras explícitas de la categoría (como notebook, conector u garantía); para otro tipo de usuarios los cuales busquen otros productos, podría ser una desventaja la prioridad que se le da a las especificaciones sobre otro tipo de atributos, como los numéricos u las descripciones en bruto.

Otro caso particular ocurre cuando dentro del registro inicial de los productos, existe alguno que se haya comprado más de una vez (lo que sucede en la categoría de pantallas), ya que, genera un sobre ajuste y valores atípicos que se alejan demasiado de los demás producto que no cumplan con estas características se verán muy opacados en sus resultados, esto se demasiado en la cantidad de distancia que contiene esta categoría, la cual no representa la mayor parte de outliers, pero si los más alejados entre ellos.

Esta métrica de distancia tiene diferentes interpretaciones, de manera cuantitativa mide que tan alejada esta la similitud entre los productos calculados, pero entre la categoría no hay algún caso explícito que detalle, sin embargo, entre las técnicas se aprecia que aplicar la técnica de palabras claves y especialmente de reducción de dimensionalidad (como se observa en la categoría de pantallas), esto debido a que la cantidad de valores nulos se encuentran principalmente en las características de atributos (esto se ve especialmente en la de baterías y discos dada su falta de información en este aspectos), y que al aplicar SVD, la ponderación en el cálculo se mantiene, pero la matriz de datos obtenida se ocupa con menos datos, por lo que, solamente ciertos atributos se ven parte del cálculo (los que representan una mayor varianza, que para este caso se ocuparon dos componentes). Esto se acompaña con que la precisión y el recall en promedio disminuye al aplicar la reducción de dimensionalidad, por lo que en general para estas categorías, aplicar SVD no proporciona buenos resultados en este contexto,

excepto para cargadores y fuentes, en donde la precisión aumenta un 10 % a costo de la baja de recall.

Los resultados con el uso de palabras claves depende del tipo de palabra que se ocupa para determinar el peso asignado en los vectores TF-IDF, generalmente se traduce en un aumento general de la cantidad de outliers, como se visualiza en la categoría de cargadores y fuentes y baterías; por otro lado, para pantallas LCD, al tener como palabra clave “garantía” genera una mejora en la cantidad de valores atípicos, dado que esta propiedad la brinda y exige mercado libre a sus vendedores proveedores con cierta cantidad de productos vendidos y cantidad de estrellas, por lo que, es una palabra que se repetirá generalmente en los productos, esto se puede ver en el apartado de palabras menos relevantes del vector TF-IDF de pantallas.

El uso de la técnica de palabras claves representa un aumento notorio en la precisión de los productos relevantes para la empresa, esto es debido, que dado la diferencia de peso asignada a las palabras, los productos que cumplían este criterio se veían mayormente beneficiados, por lo que para el usuario/empresa le resulta relevante que tengan mejor puesto estos productos. Sin embargo, para el caso de la categoría de discos, el uso de palabras generales y consideradas como no relevantes (como es la palabra sata e interno) genera que no exista una especificación clara del tipo de producto a obtener, dado que, son palabras que generalmente será parte de la mayoría de productos (al ser definido como no relevante por el vector, significa que es una palabra con mucha frecuencia y compartida por muchos productos), generando así una baja en la cantidad de outliers para este caso. Para las demás categorías, las palabras ocupadas detallan modelos o marcas de los productos, como Dell, Samsung, adhesivo o calidad, características que no necesariamente la comparten todos los productos.

A pesar de que algunas categorías como de discos en donde contiene el menor registro de productos, no existe un sobre ajuste desmedido a comparación de otras categorías, dado que, su promedio solamente fue de dos outliers por técnica, siendo eliminados satisfactoriamente sin necesidad de aplicar técnicas como de umbral.

Otro punto a analizar, es el resultado de aplicar normalización a las similitud general obtenida en cada categoría, la cual se ocupa con el intento de reducir la cantidad de valores atípicos en el resultados general, especialmente usando L1, pero como se detalla en los gráficos de box plot, en ninguna técnica ni categoría dio un resultado satisfactorio, solamente se cambia el rango de los valores pero manteniendo el orden. Esto conlleva al uso de la técnica de umbral, en donde se ocupa el mismo cálculo de cuartiles vistos en el gráfico de box plot; los resultados de aplicar este umbral fueron insatisfactorios, dado que, como se ve en la Tabla de resultados y en los gráficos de promedio, el índice de recall y precisión bajan drásticamente.

Para finalizar, la relación que se da entre la métrica de precisión y recall, no es directa, dado que ve en casos como de baterías y discos en donde el recall es drásticamente mayor, dada su baja cantidad de productos obtenidos en el catálogo, por lo que, para este análisis es mejor basarse en los resultados obtenidos en cargadores y pantallas, los cuales se da una relación más equilibrada, siendo generalmente el índice de precisión mayor al valor de recall. En el Gráfico 4.6 se aprecia que la precisión aumenta al reducir los valores de recall (como en palabras claves), sin embargo, esta técnica depende demasiado en las palabras ocupadas en el análisis (como ocupar la palabra garantía o interno), por lo que se puede concluir que tener un recall más alto es mejor, dado que, todos los productos que eran parte del recall eran seleccionados como productos relevantes por la empresa.

4.6. Limitaciones

Debido al bajo volumen de datos iniciales que se obtuvieron desde la empresa, los resultados y análisis de las técnicas no es tan rico en variación de resultados y opciones (son bastante sobre ajustados los valores obtenidos) siendo restringidos a un análisis por categoría de aquellas que se tuvieran más registros del conjunto de datos inicial y la cantidad de productos que aparecen en el catálogo inicial de cada categoría. Aunque es una limitación importante, no limita la validez de las conclusiones obtenidas para cada técnica, dado que son aplicados al conjunto de datos obtenido desde Mercado Libre. Sería deseable aumentar la cantidad de datos disponibles con más usuarios que permita una mayor visualización de patrones y tendencias en la compra de estos artículos, que permitiría un mejor análisis de las técnicas.

Dentro de esta misma limitación, los valores atípicos visualizados en el conjunto de datos inicial (tanto de precios, como de cantidad de transacciones) generan que los valores de similitud se vean afectados; en forma idónea estos registros deberían ser eliminados u obviados en el pre-procesamiento, sin embargo, dada la baja cantidad de registros, se prefiere mantener esta información con el objetivo de enriquecer (dentro de la medida) la variedad de resultados y limpieza de los datos, permitiendo aplicar la técnica de umbral y visualizar los resultados de precisión que conlleva eliminar estos valores.

Otra de las limitaciones que sufre este experimento, es obtener las opiniones y calificación de cada producto y de cada vendedor, dada la implementación de un límite restringido de 5 opiniones por llamada a la API de Mercado Libre, no permite la obtención de más información, por lo que, la similitud calculada en el aspecto numérico se ve restringida, dado que, existen una dependencia de las primeras opiniones y retroalimentación, dado que, si el producto contiene 100 calificaciones de los usuarios, pero solamente los 5 primeros usuarios son los que tienen una mejor calificación, genera un sesgo en los resultados (a pesar de que estos valores se ordenen según que tan útil le fue para otros usuarios la retroalimentación).

Capítulo 5

Conclusiones y Trabajos Futuros

En este capítulo resumiremos los principales resultados y contribuciones de esta investigación y desarrollo del sistema de recomendación. También mostraremos algunas opciones que podríamos tomar para trabajos futuros.

5.1. Conclusiones

En esta tesis se estudió el desarrollo de un módulo completo de comercio electrónico que permita a la empresa gestionar y comprar productos de otros proveedores, desde la creación de la interfaz con sus vistas necesarias para el funcionamiento, de la creación de la API del servidor encargado de conectarse a la base de datos y la aplicación de diferentes técnicas en el sistema de recomendación creado para la empresa CGTEC, siendo sus necesidades contar con una herramienta que pueda visualizar sus activos y decidir que proveedor le entrega una mayor ganancia.

Este módulo web aportó con un desarrollo multifacético y significativo, dada la descomposición en servicios independientes, siendo la interfaz que interactúa con los usuarios, la API del servidor que maneja la lógica del negocio y el sistema de recomendación que se especializa en el procesamiento de datos y generación de recomendaciones, estando integradas de tal forma que asegura la coherencia y eficiencia de los procesos, facilitando la experiencia del usuario. Para ello se definieron endpoints que evite la sobrecarga entre componentes, por ejemplo: Para el caso de carga del catálogo en la interfaz, se podría obtener la información de cada producto de manera individual y todos de manera asincrónica, generando así una sobrecarga y lentitud en la visualización del producto, siendo más eficiente tener un endpoint que entregue la información de 10 productos por llamada, permitiendo que en la interfaz, si el usuario quiere visualizar más ítems, basta con realizar un scroll vertical; traduciéndose en un mejor desempeño. Para esto, fue necesario gestionar el estado de cada componente usando los hooks de React (como la información de los productos) que pueda ser dinámica y cambiar al seleccionar otra categoría, gestionar la persistencia de la información usando Redux, como el inicio de sesión del usuario y sección seleccionada. Para el manejo de estilos y componentes se ocupara la biblioteca de MaterialUI que acelera el desarrollo de las vistas, sin embargo, la flexibilidad y personalización de estos componentes visuales resultan no del todo eficaces y

comunes, al ser estilos visuales basados en Google, que permite un mejor entendimiento, pero una baja identidad de producto.

Esto se acompaña con la definición de diferentes entidades en la API del servidor, dado que al ocupar una base de datos no relacional (Mongo) permite el guardado de la misma información en diferentes entidades (como la información de la compañía contenida en el mismo usuario), con esto en mente, el endpoint que se encarga de obtener la información inicial de la compañía además trae las referencias a las categorías que maneja, referencia a los registros obtenidos y proveedores relacionados con la empresa, permitiendo llamadas ligeras a la información de catálogo y proveedores. Para diseñar el modelo del servidor, se componentizó por entidades según la lógica del negocio, como: La compañía, usuario, producto, registros y proveedor. Una de las decisiones de diseño tomadas fue el como manejar la transacción entre empresas, dado que, si se ocupa un atributo de stock o cantidad para los productos, se pierde el registro histórico de cada uno, por lo que se decidió el manejar cada producto como un documento único, es decir, cada uno tiene la información detallada en la colección de producto pero manejándolo de forma unitaria; esto nos permite cambiar rápidamente el dueño, fecha de compra y proveedor de cada producto.

El resultado de ocupar una base de datos no relacional representa una ventaja en el desempeño inicial en el manejo de la información, dado que la información del producto y de la compañía pueden variar, sin embargo, frente al manejo de información sensible, como la sería el rut de la empresa, giro, transacciones resulta poco idóneo ocupar una no relacional, dada su falta de consistencia al ir escalando el proyecto, para ello se ocupa mejor base de datos relacionales que entregue una mejor consistencia y rigidez en esos datos. Por otro lado, la arquitectura de la infraestructura permite la actualización de la información del catálogo (en menor grado y no automatizada) pero no con eficacia como lo haría un lambda o una instancia EC2 que permita periódicamente ejecutarse y actualizar la base de datos del catálogo.

La exploración de diferentes técnicas permitió evaluar cómo afectaban la calidad y el nivel de similitud alcanzado para el conjunto de datos inicial, observándose varias diferencias entre ellas. El uso de la técnica de stemming permitió una mejor identificación de palabras más granuladas que compartían un mismo significado. Sin embargo, el uso de palabras técnicas en las descripciones de ciertos productos generaba la identificación de un conjunto de palabras no muy adecuadas al contexto, ya que las descripciones de las especificaciones de los productos estaban en inglés, generando errores debido al uso de *stopwords* en español. Esto resultó en resultados como 'air', 'conector', 'good_quality' identificados por el vector TF-IDF, describiendo algunos productos en español y otros en inglés, lo que provocó la identificación de palabras relevantes. Además, se identificaron palabras generales compartidas entre los productos (como "cargador", que al aplicar stemming, a veces se mantenían pero no seguían una regla estructurada, dada su baja precisión en la transformación de las palabras, en comparación con la técnica de lematización, que complementó el uso de stemming, determinando el listado de palabras que permitió obviar las diferencias y juntar estas variantes en un mismo término.

El uso de un vector TF-IDF permitió calcular la similitud entre los productos basándose en sus palabras y su frecuencia. Según los resultados, este fue el principal aspecto en el cual se basaron las recomendaciones (generalmente con mayores valores comparados a otras técnicas), aunque fue muy sensible a símbolos y palabras que mezclaban idiomas. Los valores numéricos resultaron más contraproducentes que favorables en el cálculo, como se apreció en el primer

resultado (ver Tabla C.1), donde productos más similares al conjunto inicial (como Acer Aspire) se vieron opacados por la diferencia en valores numéricos. Productos que se alejaban de los valores del conjunto inicial, aunque fueran mejores en términos representativos, se consideraron más alejados de los registros del usuario; productos con muchas valoraciones y opiniones por vendedor se encontraron más alejados del registro de la empresa. Para la empresa, era más interesante obtener recomendaciones de productos y proveedores con mayor calificación, pero si el usuario no reflejaba su propia tendencia en sus compras, el modelo se sesgaba por estos.

Para estas características, resultó más cómodo el filtrado implementado en las vistas de catálogo, permitiendo al usuario obtener productos según su preferencia. Esto se complementó con el sistema de Mercado Libre, que ofrece productos recomendados o destacados en base a la cantidad de ventas y reputación del vendedor, permitiendo ofrecer productos a menor precio, costeados por Mercado Libre. Además, sus recomendaciones se basan en el registro de visitas del usuario, información útil numéricamente en el cálculo final.

El uso de palabras claves representó un cambio en la precisión medida en la lista de recomendados para el usuario. Al cumplir con el criterio del conjunto de palabras, la mayor ponderación generó que los productos tomaran una mejor posición en la lista de recomendados. Sin embargo, esto aumentó la cantidad de valores atípicos, ya que productos menos similares al registro inicial de la empresa se beneficiaban de la mayor ponderación de estas palabras, ampliando el rango de los cuartiles en los gráficos box plot (como en el caso de pantallas). Esto generó una disminución en el valor de recall, ya que si el producto del conjunto inicial de la empresa no contenía la palabra clave, se veía opacado por otros productos.

La reducción de dimensionalidad usando SVD fue la técnica con menor resultado en este experimento (dada la nula información para dos categorías). El análisis se centró en las categorías de Cargadores y Fuentes y Pantallas LCD. Para la primera categoría, presentó una mejora en la cantidad de outliers, aumentando el rango de valores admitidos en los atributos. Sin embargo, aunque aumentó la precisión en un 10 %, hubo una baja considerable en el recall del listado de palabras. En la categoría de pantallas, al haber dos productos iguales en el registro inicial y la categoría con mayor porcentaje de nulos, se esperaba un mejor resultado. Sin embargo, la distancia aumentó drásticamente de 7 a 15 y los outliers de 18 a 105 valores atípicos, además de una baja de precisión de un 10 %. Por lo tanto, aunque SVD se desempeña mejor para matrices dispersas, no representó los mejores resultados para este sistema de recomendación.

Finalmente, la reducción de outliers a través de un umbral disminuyó la distancia entre las similitudes calculadas (hasta un promedio menor que 4), ya que al eliminar los productos fuera de los límites, la desviación y la media abarcaban a todos los productos. Sin embargo, aplicar esta técnica significó una disminución abrupta en la precisión determinada por la empresa, que no superó el 30 % de relevancia en el top 10 de los productos recomendados. Además, al eliminar los productos outliers, el recall disminuyó hasta no entregar ningún producto del registro inicial, ya que estos eran generalmente parte de los valores atípicos identificados por el sistema.

5.2. Trabajo Futuro

Como trabajo futuro es interesante la implementación de un sistema de recomendación basado en filtrados colaborativos, en un caso en donde ya varios usuarios sean parte del sistema y registren sus propias compras a proveedores, se tendrá a disposición los resultados y calificaciones de varios usuarios a productos compartidos, abriendo la opción de aplicar modelos de predicción y de similitud para entregar recomendaciones menos sesgadas y con mayor información de tendencias y patrones; esto no dejaría de lado el sistema basado en contenidos, dado que permitirá abordar casos de la entrada de nuevos productos al sistema, por lo que, la opción de implementar un sistema híbrido (como lo aplica el propio Mercado Libre o Netflix) resultará en recomendaciones más precisas y ricas en variedad.

Complementando al escalabilidad del sistema, se puede agregar diferentes fuentes de plataformas online a través de técnicas de scrapping, la cual permite identificar a través de un código de identificación los mismos productos ofrecidos en distintas páginas, lo que permitiría una mejor recomendación en riqueza de variedad, permitiendo así a los usuarios visualizar diferentes opciones de varios proveedores en diferentes plataformas sin necesidad de recorrerlas todas. Esto se logra con una migración de la información a algún servicio en la nube escalable, que pueda ser el uso de Amazon web services o de Azure de Microsoft.

Bibliografía

- [Aggarwal, 2015] Aggarwal, C. C. (2015). *Data Mining: The Textbook*. Springer, Cham.
- [Aggarwal, 2016] Aggarwal, C. C. (2016). *Recommender Systems - The Textbook*. Springer.
- [Alibaba, 2022] Alibaba, G. (2022). Authentication & verification. information center. <<https://seller.alibaba.com/memberships/ggs/avprocess.html?spm=5386.1223793.1998081280.22.AbKdOv>> (último acceso: 9.12.22).
- [Autoventa, 2022] Autoventa (2022). Software para preventa y autoventa. sección funciones. <<https://www.autoventa.io/funciones>> (último acceso: 11.12.22).
- [Cardemil, 2022] Cardemil, M. (2022). *E-BUSINESS: Definición de la estrategia de internet business, para una empresa editorial*. PhD thesis, Universidad Técnica Federico Santa María. <<https://repositorio-usm-cl.usm.idm.oclc.org/handle/11673/43129>>.
- [CERTUS, 2020] CERTUS (2020). 8 tipos de modelos de negocios que debes conocer. <<https://www.certus.edu.pe/blog/tipos-modelos-negocios/>> (último acceso: 10.12.22).
- [Conecta, 2022a] Conecta, M. (2022a). La toma de decisiones en ecommerce. <<https://www.conectasoftware.com/magazine/analytics/la-toma-de-decisiones-en-ecommerce/>> (último acceso: 7.12.22).
- [Conecta, 2022b] Conecta, M. (2022b). Los KPIS de un ecommerce. <<https://www.conectasoftware.com/magazine/analytics/kpi-ecommerce/>> (último acceso: 7.12.22).
- [Cárdenas, 2020] Cárdenas, J. (2020). Conoce la historia del comercio electrónico y cómo ha evolucionado hasta hoy. <<https://rockcontent.com/es/blog/historia-del-comercio-electronico/>> (último acceso: 8.12.22).
- [Drewel and Gausemeier, 2018] Drewel, M. and Gausemeier, J. (2018). Digital B2B-platforms and how to find the right one. In *European Conference on Innovation and Entrepreneurship*.
- [Electrónica, 2021] Electrónica, M. (2021). SRM (supplier relationship management) para el departamento de compras. <<https://blog.mercadoe.com/es/espanol-srm-supplier-relationship-management-para-el-departamento-de-compras/>> (último acceso: 11.12.22).
- [Fensel et al., 2001] Fensel, D., Ding, Y., Omelayenko, B., Schulten, E., Botquin, G., Brown, M., and Flett, A. (2001). Product data integration in B2B e-commerce. *IEEE Intelligent Systems*, 16(4):54-59.
- [Gestioo, 2022] Gestioo (2022). Software para talleres. reparación y servicio técnico. <<https://gestioo.com/producto/software-para-talleres>> (último acceso: 11.12.22).
- [Hat, 2022] Hat, R. (2022). ¿Qué es la Api? <<https://www.redhat.com/es/topics/api/what-are-application-programming-interfaces>> (último acceso: 11.12.22).
- [Honrado et al., 2000] Honrado, A., Leon, R., O'Donnel, R., and Sinclair, D. (2000). A word stemming algorithm for the Spanish language. In *Proceedings Seventh International Symposium on String Processing and Information Retrieval. SPIRE 2000*, pages 139-145.

- [L., 2021] L., C. (2021). ¿qué es un ecommerce? tipos, cómo crearlo y ejemplos. <<https://www.cyberclick.es/numerical-blog/que-es-un-ecommerce-tipos-como-crearlo-y-ejemplos>> (último acceso: 8.12.22).
- [Libre,] Libre, M. Mercado libre: somos la plataforma líder de comercio electrónico de América Latina. <<https://www.mercadolibre.cl/institucional/hacemos/mercado-libre-nuestro-marketplace#:~:text=Actualmente%2C%20tenemos%20operaciones%20en%2018,>>> (último acceso: 11.12.22).
- [López, 2016] López, C. (2016). Curso organización y sistemas de información (OSI). <<https://docs.google.com/presentation/d/1pJ8iszlCG6dhSnUvUfJdIDVNWeBXZ1vjsEwf5hWP0s/edit#slide=id.p3>> (último acceso: 6.11.22).
- [MercadoLibre, 2023] MercadoLibre (2023). *Documentacion para desarrollador de mercado libre*.
- [OSI, 2018] OSI, W. (2018). Tipos de sistemas de información. <https://lc-ip81.inf.utfsm.cl/wiki/index.php/Tipos_de_Sistemas_de_Información> (último acceso: 11.12.22).
- [OVHcloud, 2024] OVHcloud (2024). ¿Qué es una api rest?
- [Pateli and Giaglis, 2008] Pateli, A. and Giaglis, G. (2008). A framework for understanding and analysing ebusiness models. *BLED 2003 proceedings*.
- [Richardson and Innovation, 2010] Richardson, A. and Innovation, X. (2010). Why a company toughest problems are its greatest advantage. *Adam Jossey-Bass*.
- [Sarwar et al., 2000] Sarwar, B., Karypis, G., Konstan, J., and Riedl, J. (2000). Analysis of recommendation algorithms for e-commerce. *Proceedings of ACM E-Commerce*, 1.
- [Silveira et al., 2019] Silveira, T., Zhang, M., Lin, X., Liu, Y., and Ma, S. (2019). How good your recommender system is? a survey on evaluations in recommendation. *International Journal of Machine Learning and Cybernetics*, 10:813-831.
- [Villalonga, 2022] Villalonga, R. (2022). Sistemas recomendadores basados en contenido. *Medium*.
- [Wei et al., 2007] Wei, K., Huang, J., and Fu, S. (2007). A survey of e-commerce recommender systems. In *2007 International Conference on Service Systems and Service Management*, pages 1-5.
- [Zeibak, 2017] Zeibak, L. (2017). How Alibaba works: Advice for buyers & sellers. <<https://tinuiti.com/blog/ecommerce/what-is-alibaba/>> (último acceso: 9.12.22).
- [Zhang et al., 2008] Zhang, Y., Yang, Z., and Lin, H. (2008). Rating of e-commerce integrity based on decision tree. In *2008 4th International Conference on Wireless Communications, Networking and Mobile Computing*.

Apéndice A

Gráficos de varianza de SVD y similitudes

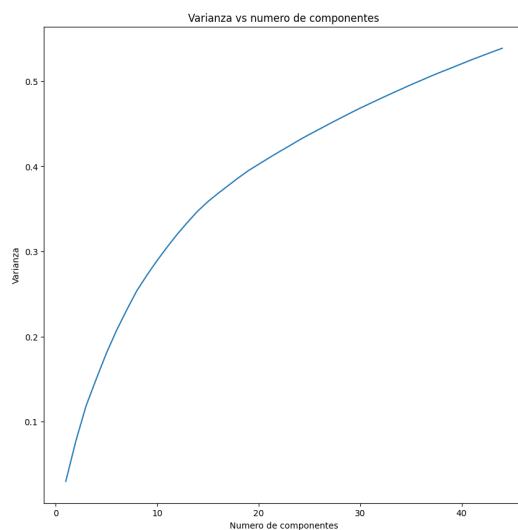


Figura A.1 Varianza de Cargadores

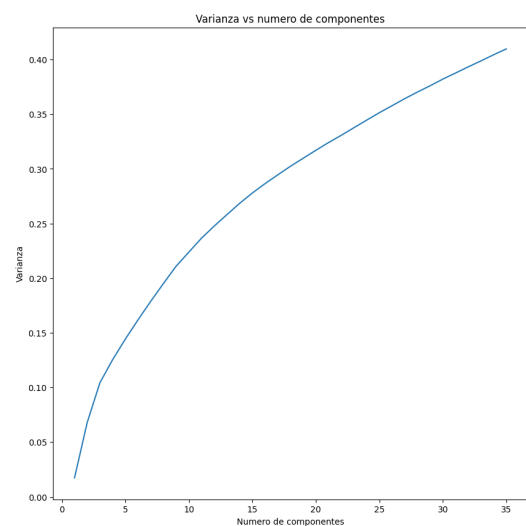


Figura A.2 Varianza de Baterías

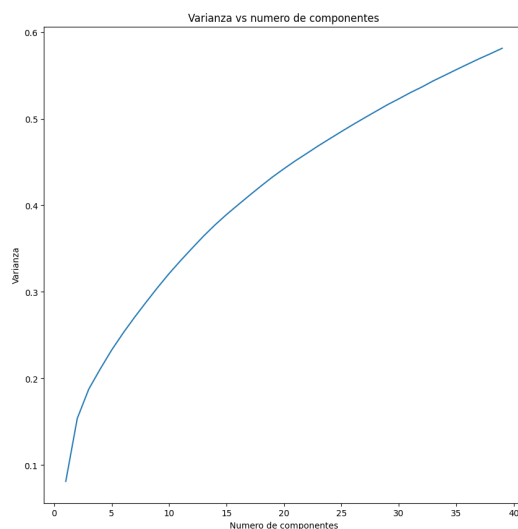


Figura A.3 Varianza de Pantallas LCD

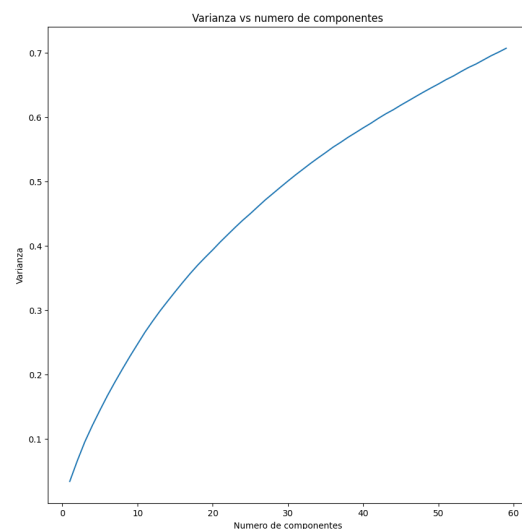


Figura A.4 Varianza de Discos duros y ssd

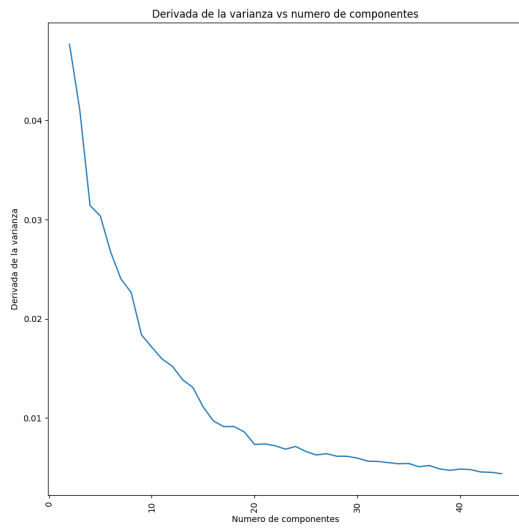


Figura A.5 Varianza Derivada de Cargadores y Fuentes

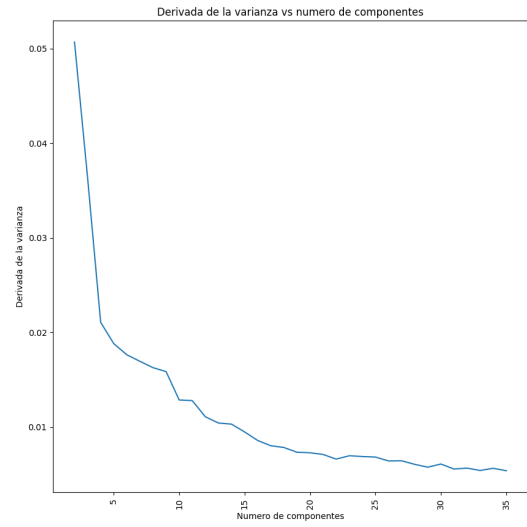


Figura A.6 Varianza Derivada de Baterías

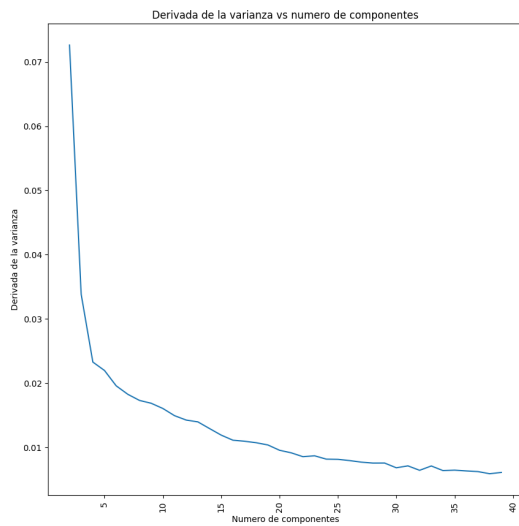


Figura A.7 Varianza Derivada de Pantallas LCD

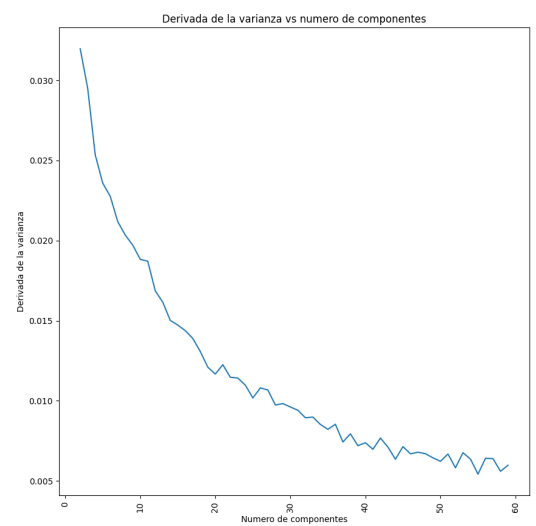


Figura A.8 Varianza Derivada de Discos duros y ssd

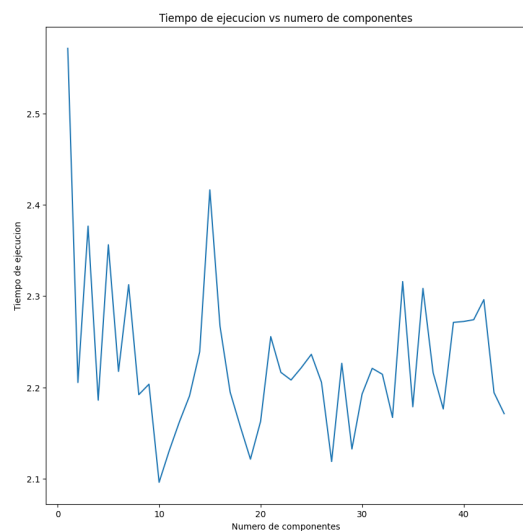


Figura A.9 Tiempo de Cargadores

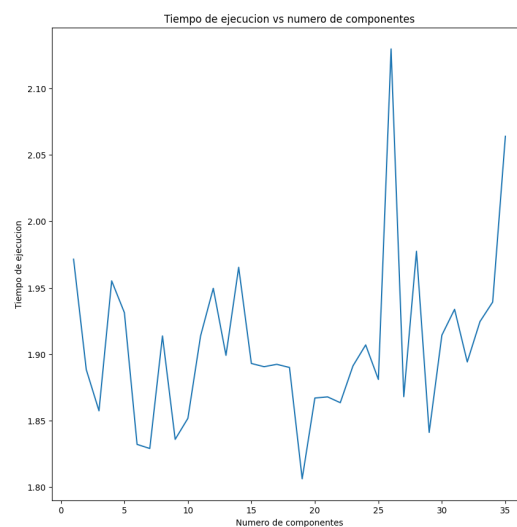


Figura A.10 Tiempo de Baterías

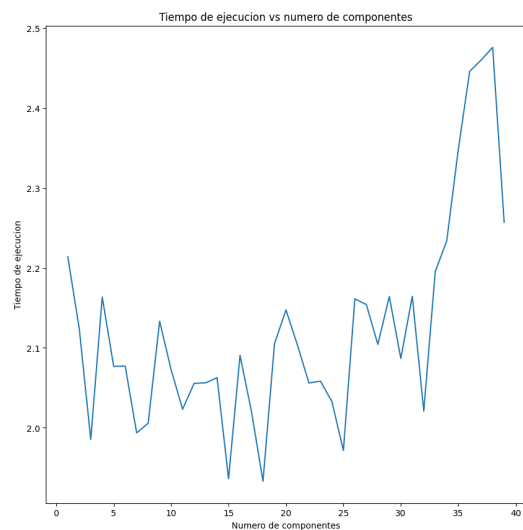


Figura A.11 Tiempo de Pantallas LCD

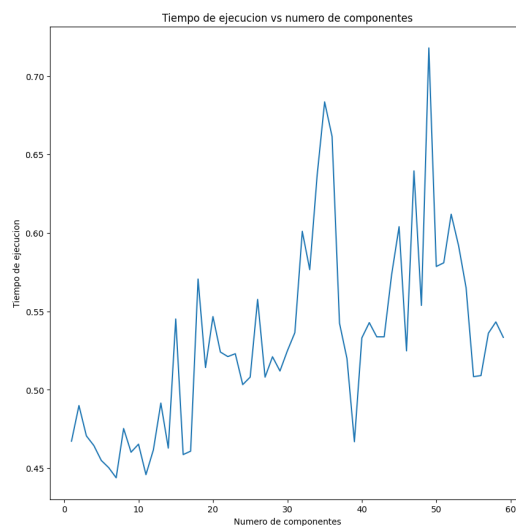


Figura A.12 Tiempo de Discos duros y ssd

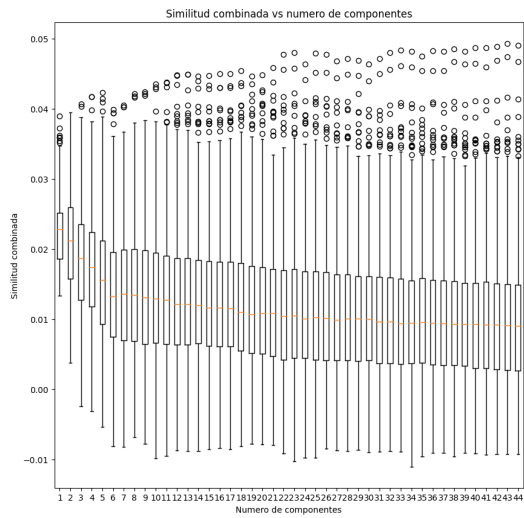


Figura A.13 Similitud de Cargadores

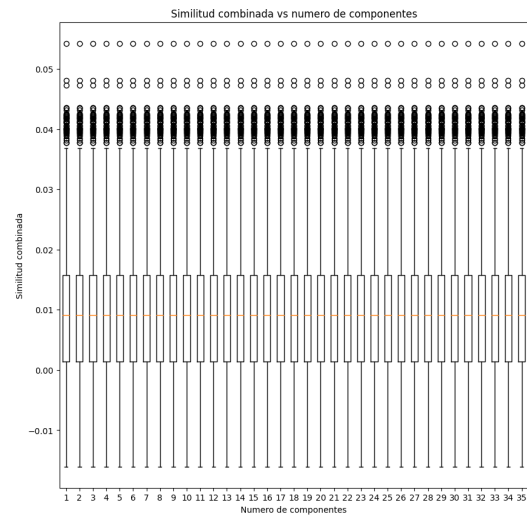


Figura A.14 Similitud de Baterías

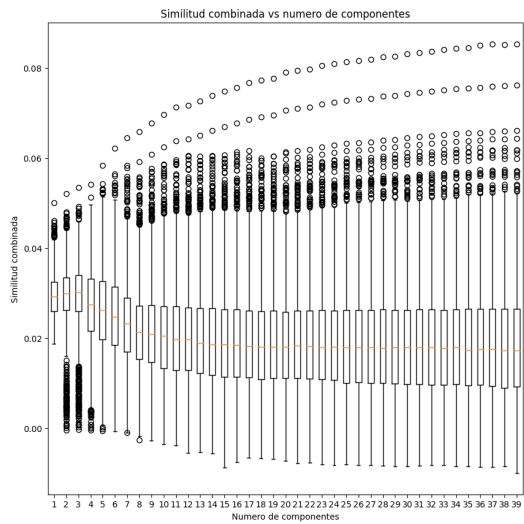


Figura A.15 Similitud de Pantallas LCD

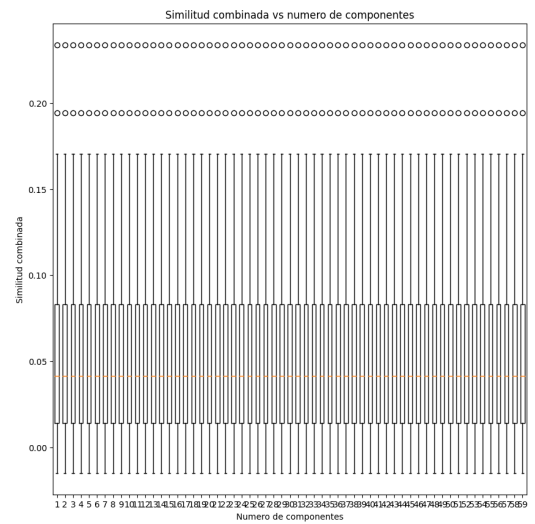


Figura A.16 Similitud de Discos duros y ssd

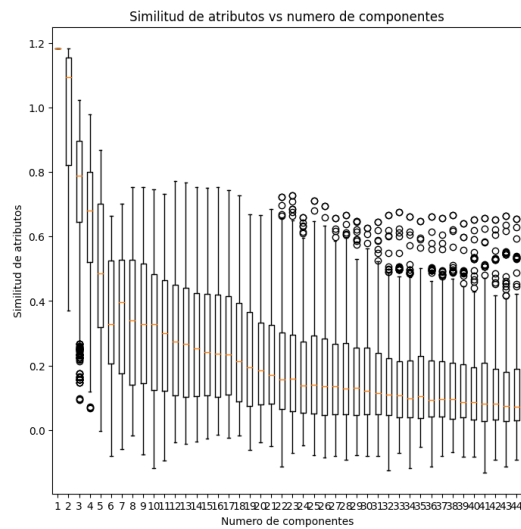


Figura A.17 Similitud de atributos de Cargadores

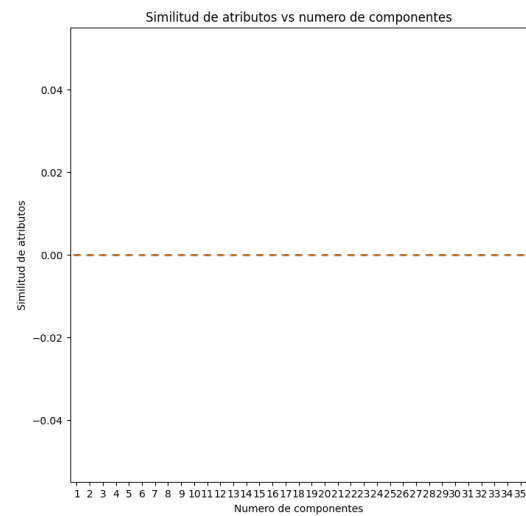


Figura A.18 Similitud de atributos de Baterías

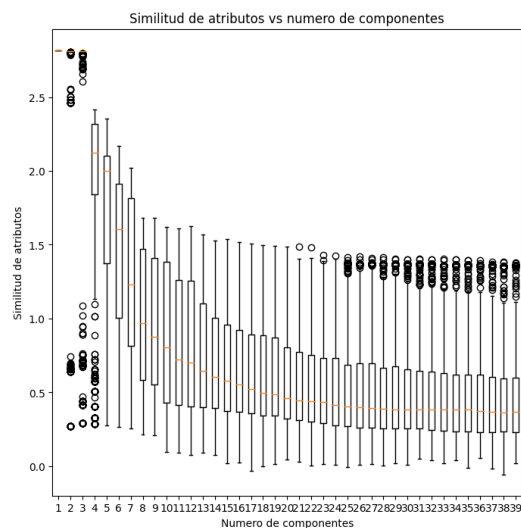


Figura A.19 Similitud de Pantallas LCD

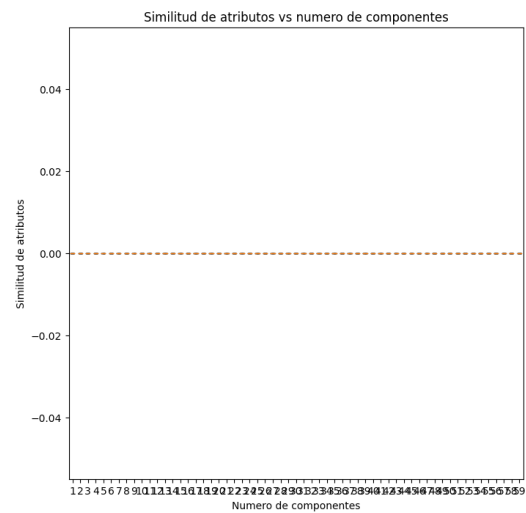


Figura A.20 Similitud de atributos de Discos duros y SSD

Apéndice B

Tabla de resultados de vector TF-IDF

B.1. Cargadores y Fuentes

Técnica	Más relevante	Menos relevante
Sin Técnicas	'zenbook' 'alimentación' 'amp' 'amarillo' 'amarilla' 'am070la' '1040' 'alt' 'al007la' '102w' 'ak005la' 'ahom' 'ahgase' 'ah0004la' 'ah0003la' 'ag0002la' '1030' 'aust' 'c403la' 'blanco'	'nuevo' '65' '45' 'cm' 'hp' 'lenovo' 'acer' 'genérica' 'dell' 'asus' 'genérico' '11' '90' '60' 'pavilion' 'aspire' '14' 'adp' '120' 'pa'
Stemming	'zenbook' 'g1' 'fx553v' '65w-typec' 'g-tec' 'g-xhh06' 'g-xhh07' 'g-xhh12' 'g-xhh15' 'g-xhh16' 'g-xhh23' 'g-xhh60' 'g-xhh75' 'g2' 'gabinet' 'g3-571' 'g3-572' '65w-4.5mm' 'g40-30' 'g40-80'	'nuev' 'w' 'cm' 'g' 'hp' 'gener' 'lenov' 'acer' 'dell' 'asus' 'pavilion' 'aspir' 'inspiron' 'cargador' 'ideap' '19v' 'macbook' 'samsung' 'magsaf' 'universal'
Palabras Claves	'zenbook' 'alimentación' 'amp' 'amarillo' 'amarilla' 'am070la' '1040' 'alt' 'al007la' '102w' 'ak005la' 'ahom' 'ahgase' 'ah0004la' 'ah0003la' 'ag0002la' '1030' 'aust' 'c403la' 'blanco'	'nuevo' '65' '45' 'cm' 'hp' 'lenovo' 'acer' 'genérica' 'dell' 'asus' 'genérico' '11' '90' '60' 'pavilion' 'aspire' '14' 'adp' '120' 'pa'
SVD	'zenbook' 'alimentación' 'amp' 'amarillo' 'amarilla' 'am070la' '1040' 'alt' 'al007la' '102w' 'ak005la' 'ahom' 'ahgase' 'ah0004la' 'ah0003la' 'ag0002la' '1030' 'aust' 'c403la' 'blanco'	'nuevo' '65' '45' 'cm' 'hp' 'lenovo' 'acer' 'genérica' 'dell' 'asus' 'genérico' '11' '90' '60' 'pavilion' 'aspire' '14' 'adp' '120' 'pa'

Tabla B.1 Palabras más y menos relevantes por vector TF-IDF en atributos

Técnica	Más relevante	Menos relevante
Sin Técnicas	'332t' 'acumulado' 'ad6019' 'is' 'iskb' 'it' 'itil' 'ad2129020' 'iturra"iv' 'ad2108020' 'ad045g4' 'j002dx' 'j003cl' 'j003xx' 'j006cl' 'j007cl' 'j009wm' 'j01' '2301wlmi'	'messages_as_seller' 'normal' 'immediate_payment' 'credits_profile' 'cart_eligible' 'cargador' 'extended_warranty_eligible' 'kvs_primary' 'mshops' 'metropolitana' 'rm' 'good_quality_thumbnail' 'factura' 'notebook' 'original' 'conector' 'good_quality_picture' 'boleata' 'entrega' 'nuevo'
Stemming	'salida:12v-15v-16v-18v-19v-20v-24v' 'dv5040us' 'e/' 'e0w59u' 'e0w60u' 'e1-422-3431' 'e1-431-2818' 'e1-531-4436' 'e1-531g' 'e1-570g' 'e1-571' 'e1-571g' 'e1-571p' 'e1-571pg' 'e1-731' 'e1501' 'e201' 'e201n' 'e201na' 'e202'	'messages_as_sell' 'normal' 'immediate_payment' 'credits_profil' 'cart_elig' 'cargador' 'extended_warranty_elig' 'kvs_primary' 'mshops' 'metropolitan' 'rm' 'good_quality_thumbnail' 'factur' 'bolet' 'notebook' 'original' 'entreg' 'conector' 'good_quality_pictur' 'model'
Palabras Claves	'332t' 'acumulado' 'ad6019' 'is' 'iskb' 'it' 'itil' 'ad2129020' 'iturra"iv' 'ad2108020' 'ad045g4' 'j002dx' 'j003cl' 'j003xx' 'j006cl' 'j007cl' 'j009wm' 'j01' '2301wlmi'	'messages_as_seller' 'normal' 'immediate_payment' 'credits_profile' 'cart_eligible' 'cargador' 'extended_warranty_eligible' 'kvs_primary' 'mshops' 'metropolitana' 'rm' 'good_quality_thumbnail' 'factura' 'notebook' 'original' 'conector' 'good_quality_picture' 'boleata' 'entrega' 'nuevo'
SVD	'332t' 'acumulado' 'ad6019' 'is' 'iskb' 'it' 'itil' 'ad2129020' 'iturra"iv' 'ad2108020' 'ad045g4' 'j002dx' 'j003cl' 'j003xx' 'j006cl' 'j007cl' 'j009wm' 'j01' '2301wlmi'	'messages_as_seller' 'normal' 'immediate_payment' 'credits_profile' 'cart_eligible' 'cargador' 'extended_warranty_eligible' 'kvs_primary' 'mshops' 'metropolitana' 'rm' 'good_quality_thumbnail' 'factura' 'notebook' 'original' 'conector' 'good_quality_picture' 'boleata' 'entrega' 'nuevo'

Tabla B.2 Palabras más y menos relevantes obtenidas por vector TF-IDF

B.2. Baterías

Técnica	Más relevante	Menos relevante
Sin Técnicas	'zte' 'hb474284rbc' 'hb406689ecw' 'hb417094ebc' 'hb4242b4ebw' 'hb426489eew' 'b600bc' 'hb436486cw' 'b2pzf100' 'hb486486cw' 'b800bc' 'b2pw4100' 'b2ps6100' 'b100ae' 'hb526488eew' 'alternativo' 'hd40' 'he316' 'b650ac' 'b800be'	'nuevo' 'genérico' 'eb' 'genérica' 'samsung' 'reacondicionado' 'iphone' 'xiaomi' 'compatible' '616' 'alternativa' 'bl' 'huawei' 'apple' 'galaxy' 'motorola' 'marca' 'plus' 'jk50' 'zoex'
Stemming	'bn54' 'flashlight' 'eb-bj710cbe' 'eb-bj720ab' 'eb-bm526aby' 'eb-bn910bbe' 'eb-bn915bbc' 'eb-bn980aby' 'eb-bt111lab' 'eb-bt530fbc' 'eb-bt530fbe' 'eb-bt530fbu' 'eb20' 'eb425161lu' 'eb615268vu' 'electronics' 'ep-500' 'ep350' 'esm48100b1' 'ey30'	'nuev' 'gener' 'samsung' 'reacondicion' 'iphon' 'xiaomi' 'altern' 'compat' 'huawei' 'apple' 'galaxy' 'motorol' 'marc' 'plus' 'jk50' 'zoex' 'hb366481ecw' '6s' 'eb-ba315aby' 'eb-ba505'
Palabras Claves	'zte' 'hb474284rbc' 'hb406689ecw' 'hb417094ebc' 'hb4242b4ebw' 'hb426489eew' 'b600bc' 'hb436486cw' 'b2pzf100' 'hb486486cw' 'b800bc' 'b2pw4100' 'b2ps6100' 'b100ae' 'hb526488eew' 'alternativo' 'hd40' 'he316' 'b650ac' 'b800be'	'nuevo' 'genérico' 'eb' 'genérica' 'samsung' 'reacondicionado' 'iphone' 'xiaomi' 'compatible' '616' 'alternativa' 'bl' 'huawei' 'apple' 'galaxy' 'motorola' 'marca' 'plus' 'jk50' 'zoex'
SVD	'zte' 'hb474284rbc' 'hb406689ecw' 'hb417094ebc' 'hb4242b4ebw' 'hb426489eew' 'b600bc' 'hb436486cw' 'b2pzf100' 'hb486486cw' 'b800bc' 'b2pw4100' 'b2ps6100' 'b100ae' 'hb526488eew' 'alternativo' 'hd40' 'he316' 'b650ac' 'b800be'	'nuevo' 'genérico' 'eb' 'genérica' 'samsung' 'reacondicionado' 'iphone' 'xiaomi' 'compatible' '616' 'alternativa' 'bl' 'huawei' 'apple' 'galaxy' 'motorola' 'marca' 'plus' 'jk50' 'zoex'

Tabla B.3 Palabras más y menos relevantes por vector TF-IDF en Atributos

Técnica	Más relevante	Menos relevante
Sin Técnicas	'g9300' 'fuego' 'g928c' 'g9280' 'g928' 'g925v' 'g925t' 'g925l' 'g925k' 'g925i' 'g925a' 'g920t' 'g920s' 'g920l' 'g920k' 'g920i' 'g920f' 'g920a' 'g9200' 'g928g'	'messages_as_seller' 'normal' 'immediate_payment' 'credits_profile' 'kvs_primary' 'rm' 'metropolitana' 'batería' 'producto' 'mshops' 'good_quality_thumbnail' 'bateria' 'extended_warranty_eligible' 'good_quality_picture' 'compatible' 'santiago' 'servicio' 'meses' 'instalación' 'capacidad'
Stemming	'hb4242b4ebw' 'matriz' 'cp185' 'coup' 'cotiz' 'cotic' 'mañan' 'mb520' 'mb525' 'mb855' 'mb865' 'mci' 'mci_electronics' 'me525' 'me526' 'me865' 'correspondient' 'coquimb' 'copiap' 'mediocr'	'messages_as_sell' 'normal' 'immediate_payment' 'credits_profil' 'kvs_primary' 'rm' 'metropolitan' 'product' 'bat' 'mshops' 'good_quality_thumbnail' 'bateri' 'extended_warranty_elig' 'good_quality_pictur' 'compat' 'envi' 'santiag' 'tecnic' 'instal' 'servici'
Palabras Claves	'g9300' 'fuego' 'g928c' 'g9280' 'g928' 'g925v' 'g925t' 'g925l' 'g925k' 'g925i' 'g925a' 'g920t' 'g920s' 'g920l' 'g920k' 'g920i' 'g920f' 'g920a' 'g9200' 'g928g'	'messages_as_seller' 'normal' 'immediate_payment' 'credits_profile' 'kvs_primary' 'rm' 'metropolitana' 'batería' 'producto' 'mshops' 'good_quality_thumbnail' 'bateria' 'extended_warranty_eligible' 'good_quality_picture' 'compatible' 'santiago' 'servicio' 'meses' 'instalación' 'capacidad'
SVD	'g9300' 'fuego' 'g928c' 'g9280' 'g928' 'g925v' 'g925t' 'g925l' 'g925k' 'g925i' 'g925a' 'g920t' 'g920s' 'g920l' 'g920k' 'g920i' 'g920f' 'g920a' 'g9200' 'g928g'	'messages_as_seller' 'normal' 'immediate_payment' 'credits_profile' 'kvs_primary' 'rm' 'metropolitana' 'batería' 'producto' 'mshops' 'good_quality_thumbnail' 'bateria' 'extended_warranty_eligible' 'good_quality_picture' 'compatible' 'santiago' 'servicio' 'meses' 'instalación' 'capacidad'

Tabla B.4 Palabras más y menos relevantes por vector TF-IDF

B.3. Pantallas LCD

Técnica	Más relevante	Menos relevante
Sin Técnicas	'g9play' '25' 'hauwei' 'h0' 'gnerico' 'glass' '20e' 'gen' 'galazy' '40' 'g31' 'zhigao' 'fusión' 'fe' '193' 'f2' 'eda004164705' 'e32s' 'e20' '2nd'	'nuevo' 'samsung' 'generica' 'genérica' 'compatible' 'generico' 'servitec' 'huawei' 'redmi' 'iphone' 'xiaomi' 'completa' 'motorola' '5g' 'note' 'dc' 'full' 'mobile' 'tel' 'pro'
Stemming	'j2' 'k42/k52/k62' 'k40' 'k10' 'j5' 'j400m' 'j400' 'zhiga' 'ipad' 'incel' 'hyper' 'hd' 'hauwei' 'gneric' 'glass' 'gen.' 'galazy' 'g9play' 'g31' 'fe'	'nuev' 'gener' 'samsung' 'compat' 'servitec' 'complet' 'huawei' 'redmi' 'iphon' 'xiaomi' 'motorol' '5g' 'not' 'full' 'dc' 'mobil' 'z-tel' 'pro' '4g' 'lit'
Palabras Claves	'g9play' '25' 'hauwei' 'h0' 'gnerico' 'glass' '20e' 'gen' 'galazy' '40' 'g31' 'zhigao' 'fusión' 'fe' '193' 'f2' 'eda004164705' 'e32s' 'e20' '2nd'	'nuevo' 'samsung' 'generica' 'genérica' 'compatible' 'generico' 'servitec' 'huawei' 'redmi' 'iphone' 'xiaomi' 'completa' 'motorola' '5g' 'note' 'dc' 'full' 'mobile' 'tel' 'pro'
SVD	'g9play' '25' 'hauwei' 'h0' 'gnerico' 'glass' '20e' 'gen' 'galazy' '40' 'g31' 'zhigao' 'fusión' 'fe' '193' 'f2' 'eda004164705' 'e32s' 'e20' '2nd'	'nuevo' 'samsung' 'generica' 'genérica' 'compatible' 'generico' 'servitec' 'huawei' 'redmi' 'iphone' 'xiaomi' 'completa' 'motorola' '5g' 'note' 'dc' 'full' 'mobile' 'tel' 'pro'

Tabla B.5 Palabras más y menos relevantes por vector TF-IDF en Atributos

Técnica	Más relevante	Menos relevante
Sin Técnicas	'inteligente' 'perdido' 'perfcto' 'empezaba' 'empaquetan' 'empaquetadas' 'empaquetada' 'permanecio' 'empacado' 'emiratos' 'perono' 'emergencia' 'embase' 'personalizado' 'personalmente' 'pesar' 'pesca' 'embalados' 'embaladas' 'pestañas'	'messages_as_seller' 'normal' 'immediate_payment' 'cart_eligible' 'pantalla' 'good_quality_thumbnail' 'credits_profile' 'kvs_primary' 'mshops' 'producto' 'rm' 'metropolitana' 'garantía' 'calidad' 'si' 'compatible' 'envíos' 'cualquier' 'productos' 'instalación'
Stemming	'...hac' 'ldn-lx3' 'ktech' 'l03' 'l09' 'l1' 'laboral' 'lamina+' 'las' 'lastim' 'lcd/touch' 'ldn-l01' 'led' 'j710' 'lenov' 'leon' 'leserq' 'lgm-k120l' 'lgm-k120s' 'lid'	'messages_as_sell' 'normal' 'immediate_payment' 'cart_elig' 'pantall' 'good_quality_thumbnail' 'credits_profil' 'kvs_primary' 'product' 'mshops' 'rm' 'metropolitan' 'envi' 'instal' 'garant' 'compat' 'calid' 'si' 'compr' 'tecnic'
Palabras Claves	'inteligente' 'perdido' 'perfcto' 'empezaba' 'empaquetan' 'empaquetadas' 'empaquetada' 'permanecio' 'empacado' 'emiratos' 'perono' 'emergencia' 'embase' 'personalizado' 'personalmente' 'pesar' 'pesca' 'embalados' 'embaladas' 'pestañas'	'messages_as_seller' 'normal' 'immediate_payment' 'cart_eligible' 'pantalla' 'good_quality_thumbnail' 'credits_profile' 'kvs_primary' 'mshops' 'producto' 'rm' 'metropolitana' 'garantía' 'calidad' 'si' 'compatible' 'envíos' 'cualquier' 'productos' 'instalación'
SVD	'inteligente' 'perdido' 'perfcto' 'empezaba' 'empaquetan' 'empaquetadas' 'empaquetada' 'permanecio' 'empacado' 'emiratos' 'perono' 'emergencia' 'embase' 'personalizado' 'personalmente' 'pesar' 'pesca' 'embalados' 'embaladas' 'pestañas'	'messages_as_seller' 'normal' 'immediate_payment' 'cart_eligible' 'pantalla' 'good_quality_thumbnail' 'credits_profile' 'kvs_primary' 'mshops' 'producto' 'rm' 'metropolitana' 'garantía' 'calidad' 'si' 'compatible' 'envíos' 'cualquier' 'productos' 'instalación'

Tabla B.6 Palabras más y menos relevantes por vector TF-IDF

B.4. Discos duros y sólidos

Técnica	Más relevante	Menos relevante
Sin Técnicas	'hdtca10xr3aa' 'for' 'hdtb440xk3ca' 'hdtb420xk3aa' 'hdd' 'gstfs31480gntd' 'gstfs31240gntd' 'gsm2ne8256gntd' 'gsm2ne3256gntd' 'gen' 'game' 'g325e500g' 'g325e1tb' 'firecud' 'hdtb520xk3aa' 'er500' 'eesn' 'e100' 'duro' 'drive'	nuevo' 'western' 'digital' 'wd' 'cm' 'kingston' 'kg' 'seagate' 'cambio' 'toshiba' 'green' 'blue' 'crucial' '10' '11' '149' '40' 'ssd' 'purple' '32'
Stemming	'firecud' 'ct2000p3ssd8' 'sa510' 'sa400s37/960g' 'sa400s37/480g' 'hdd' 'sa400s37/120g' 'comput' 'rig' 'rgb' 'ct1000bx500ssd1' 'ct1000mx500ssd1' 'ps4' 'ct1000p2ssd8' 'ct1000p3ssd8' 'pc' 'passport' 'p3-256' 'ct2000mx500ssd1' 'oz'	'nuev' 'g' 'western' 'digital' 'wd' 'cm' 'kingston' 'kg' 'seagat' 'canvi' 'toshib' 'gre' 'blu' 'crucial' 'purpl' 'ssd' 'gigabyt' 'extern' 'nv2' 'a400'
Palabras Claves	'hdtca10xr3aa' 'for' 'hdtb440xk3ca' 'hdtb420xk3aa' 'hdd' 'gstfs31480gntd' 'gstfs31240gntd' 'gsm2ne8256gntd' 'gsm2ne3256gntd' 'gen' 'game' 'g325e500g' 'g325e1tb' 'firecud' 'hdtb520xk3aa' 'er500' 'eesn' 'e100' 'duro' 'drive'	nuevo' 'western' 'digital' 'wd' 'cm' 'kingston' 'kg' 'seagate' 'cambio' 'toshiba' 'green' 'blue' 'crucial' '10' '11' '149' '40' 'ssd' 'purple' '32'
SVD	'hdtca10xr3aa' 'for' 'hdtb440xk3ca' 'hdtb420xk3aa' 'hdd' 'gstfs31480gntd' 'gstfs31240gntd' 'gsm2ne8256gntd' 'gsm2ne3256gntd' 'gen' 'game' 'g325e500g' 'g325e1tb' 'firecud' 'hdtb520xk3aa' 'er500' 'eesn' 'e100' 'duro' 'drive'	nuevo' 'western' 'digital' 'wd' 'cm' 'kingston' 'kg' 'seagate' 'cambio' 'toshiba' 'green' 'blue' 'crucial' '10' '11' '149' '40' 'ssd' 'purple' '32'

Tabla B.7 Palabras más y menos relevantes por vector TF-IDF de atributos

Técnica	Más relevante	Menos relevante
Sin Técnicas	'instalo' 'gsm2ne8256gntd' 'hdtx110xk3aa' 'hdtca40x' 'hdtca20x' 'hdtb540xk3ca' 'hdtb520xk3aa' 'hdtb510xk3aa' 'hdtb440xk3ca' 'hdtb420xk3aa' 'hdtb410xk3aa' 'haga' 'hacer' 'gstfs31480gntd' 'gstfs31240gntd' 'gsm2ne3256gntd' 'hdtx140xk3ca' 'graaaaaan' 'gigas' 'gigabit'	'messages_as_seller' 'immediate_payment' 'cart_eligible' 'disco' 'credits_profile' 'datos' 'mshops' 'rm' 'metropolitana' 'good_quality_thumbnail' 'velocidad' 'alcance' 'producto' 'normal' 'rendimiento' 'almacenamiento' 'través' 'información' 'interfaz' 'ssd'
Stemming	'-instal' 'ct1000bx500ssd1' 'cuidad' 'passport' 'curic' 'particion' 'part' 'pareci' 'curios' 'cyber' 'pan' 'pagin' 'p3-256' 'dab' 'otebook' 'oscur' 'dan' 'deal_of_the_day' 'oprer' 'peg'	'messages_as_sell' 'immediate_payment' 'cart_elig' 'disc' 'credits_profil' 'dat' 'mshops' 'metropolitan' 'rm' 'veloc' 'good_quality_thumbnail' 'product' 'alcanc' 'normal' 'rendimient' 'almacen' 'traves' 'inform' 'disposit' 'interfaz'
Palabras Claves	'instalo' 'gsm2ne8256gntd' 'hdtx110xk3aa' 'hdtca40x' 'hdtca20x' 'hdtb540xk3ca' 'hdtb520xk3aa' 'hdtb510xk3aa' 'hdtb440xk3ca' 'hdtb420xk3aa' 'hdtb410xk3aa' 'haga' 'hacer' 'gstfs31480gntd' 'gstfs31240gntd' 'gsm2ne3256gntd' 'hdtx140xk3ca' 'graaaaaan' 'gigas' 'gigabit'	'messages_as_seller' 'immediate_payment' 'cart_eligible' 'disco' 'credits_profile' 'datos' 'mshops' 'rm' 'metropolitana' 'good_quality_thumbnail' 'velocidad' 'alcance' 'producto' 'normal' 'rendimiento' 'almacenamiento' 'través' 'información' 'interfaz' 'ssd'
SVD	'instalo' 'gsm2ne8256gntd' 'hdtx110xk3aa' 'hdtca40x' 'hdtca20x' 'hdtb540xk3ca' 'hdtb520xk3aa' 'hdtb510xk3aa' 'hdtb440xk3ca' 'hdtb420xk3aa' 'hdtb410xk3aa' 'haga' 'hacer' 'gstfs31480gntd' 'gstfs31240gntd' 'gsm2ne3256gntd' 'hdtx140xk3ca' 'graaaaaan' 'gigas' 'gigabit'	'messages_as_seller' 'immediate_payment' 'cart_eligible' 'disco' 'credits_profile' 'datos' 'mshops' 'rm' 'metropolitana' 'good_quality_thumbnail' 'velocidad' 'alcance' 'producto' 'normal' 'rendimiento' 'almacenamiento' 'través' 'información' 'interfaz' 'ssd'

Tabla B.8 Palabras más y menos relevantes por vector TF-IDF

Apéndice C

Tabla de resultados ordenados por Categoría

C.1. Cargadores y Fuentes

Nombre	Similitud Texto	Similitud Numérica	Similitud Atributos	Similitud Estándar	Similitud L1	Similitud L2
Cargador Notebook Acer Aspire 19v 3.42a 65w E3 E1 E15 5610	0.8441	0.2238	0.6059	1.6738	0.0051	0.1336
Cargador Notebook Acer Aspire 19v 3.42a 65w E3 E1 E15 5610	0.8384	0.2087	0.5174	1.5647	0.0048	0.1249
Cargador Dell 19.5v 6.7a 130w	0.9226	-0.1247	0.6059	1.4038	0.0043	0.1121
Cargador Notebook Hp Pavilion Split 19.5v 2.31a 45w X360	1.0456	0.1940	0.0365	1.2762	0.0039	0.1019
Cargador Para Notebook Acer 19v 2.37a Plug 5.5mm X 1.7mm	0.5934	0.4223	0.1980	1.2138	0.0037	0.0969
Cargador Notebook Acer 19v/3.42a Acer Aspire	0.8348	0.1424	0.1920	1.1694	0.0036	0.0934
Cargador Notebook Acer Aspire 19v 3.42a 65w E3 E1 E15 5610	0.8591	0.1916	0.1129	1.1638	0.0036	0.0929
Cargador Notebook Hp 19.5v 2.31a Punta Azul / Boleta	0.6230	0.5058	0.0273	1.1561	0.0035	0.0923
Cargador Notebook Toshiba Satellite 19v 3.42a /packard Bell	0.7188	0.1972	0.1501	1.0662	0.0032	0.0851
Cargador Hp Envy Pavilion G1 19.5v 3.33a 65w Punta Azul	0.8668	0.1330	0.0514	1.0514	0.0032	0.0839

Tabla C.1 Resultados para la categoría Cargadores y Fuentes sin Técnicas

Nombre	Similitud Texto	Similitud Numérica	Similitud Atributos	Similitud Estándar	Similitud L1	Similitud L2
Cargador Notebook Acer Aspire 19v 3.42a 65w E3 E1 E15 5610	0.8538	0.2238	0.6422	1.7199	0.0040	0.1072
Cargador Notebook Acer Aspire 19v 3.42a 65w E3 E1 E15 5610	0.8486	0.2087	0.6288	1.6862	0.0039	0.1051
Cargador Dell 19.5v 6.7a 130w	1.0027	-0.1247	0.6422	1.5202	0.0036	0.0948
Cargador Notebook Hp Pavilion Split 19.5v 2.31a 45w X360	1.1779	0.1940	0.1211	1.4932	0.0035	0.0931
Cargador Para Notebook Acer 19v 2.37a Plug 5.5mm X 1.7mm	0.6538	0.4223	0.2997	1.3759	0.0032	0.0858
Cargador Notebook Acer Aspire 19v 3.42a 65w E3 E1 E15 5610	0.8814	0.1916	0.2665	1.3396	0.0031	0.0835
Cargador Notebook Acer 19v 3.42a Compatible Con 19v 2.37a	0.6822	0.4529	0.1875	1.3227	0.0031	0.0825
Cargador Notebook Acer Chromebook 19v 3.42a 65w 3.0 X 1.0mm	0.8239	0.2576	0.2146	1.2962	0.0030	0.0808
Cargador Notebook Toshiba Satellite 19v 3.42a /packard Bell	0.8194	0.1972	0.2534	1.2702	0.0030	0.0792
Cargador Hp Envy Pavilion G1 19.5v 3.33a 65w Punta Azul	0.9778	0.1330	0.1449	1.2558	0.0029	0.0783

Tabla C.2 Resultados para la categoría Cargadores y Fuentes con Stemming

Nombre	Similitud Texto	Similitud Numérica	Similitud Atributos	Similitud Estándar	Similitud L1	Similitud L2
Cargador Para Dell 19.5v 3.34a (4.5x3.0mm) Inspiron 11-14-15	1.0645	0.5313	0.5212	2.1170	0.0044	0.1058
Cargador Dell Usb Tipo C 45w 20v 2.25a Original	1.1000	0.1567	0.5406	1.7974	0.0037	0.0898
Cargador Dell Original 19.5v 3.34a 65w 4.5x3.0mm	1.1573	0.0773	0.5450	1.7797	0.0037	0.0890
Cargador Dell 19.5v 6.7a 130w	1.2981	-0.1247	0.5973	1.7707	0.0037	0.0885
Cargador Notebook Dell 19.5v 3.34a 4.5mm Punta Fina	1.0871	0.1191	0.5384	1.7447	0.0036	0.0872
Cargador Notebook Dell 19.5v 2.31a 45w Pin Central	1.2149	-0.0612	0.5507	1.7044	0.0036	0.0852
Cargador Original Notebook Dell Inspiron 15 5570 Nuevo	1.1870	-0.0295	0.5334	1.6908	0.0035	0.0845
Cargador Original Dell 19,5v - 2,31a 45w (tipo Azul)	1.1486	0.1612	0.3766	1.6865	0.0035	0.0843
Cargador Original Notebook Dell Vostro 5490 (19.5v - 3.34a)	1.1828	-0.0567	0.5546	1.6806	0.0035	0.0840
Cargador Original Notebook Dell Inspiron 14 5100 Nuevo	1.1826	-0.0567	0.5546	1.6805	0.0035	0.0840

Tabla C.3 Resultados para la categoría Cargadores y Fuentes con Palabras Claves

Nombre	Similitud Texto	Similitud Numérica	Similitud Atributos	Similitud Estándar	Similitud L1	Similitud L2
Cargador Notebook Acer Aspire 19v 3.42a 65w E3 E1 E15 5610	0.8441	0.2238	1.1840	2.2520	0.0018	0.0567
Cargador Notebook Acer Aspire 19v 3.42a 65w E3 E1 E15 5610	0.8384	0.2087	1.1833	2.2306	0.0018	0.0562
Cargador Notebook Acer Aspire 19v 3.42a 65w E3 E1 E15 5610	0.8591	0.1916	1.1796	2.2304	0.0018	0.0562
Cargador Para Notebook Acer 19v 2.37a Plug 5.5mm X 1.7mmc	0.5934	0.4223	1.1713	2.1870	0.0018	0.0551
Cargador Notebook Acer 19v /3.42a Acer Aspire	0.8348	0.1424	1.1815	2.1589	0.0018	0.0544
Cargador Notebook Toshiba Satellite 19v 3.42a /packard Bell	0.7188	0.1972	1.1721	2.0882	0.0017	0.0526
Cargador Notebook Acer Chromebook 19v 3.42a 65w 3.0 X 1.0mm	0.6395	0.2576	1.1837	2.0809	0.0017	0.0524
Cargador Notebook Compatible Acer Aspire 19v 3.42a 5.5x1.7mm	0.8370	0.0969	1.1313	2.0652	0.0017	0.0520
Cargador Para Acer 19v 3.42a 65w / 3.0*1.0mm Punta Fina	0.6053	0.2281	1.1835	2.0170	0.0016	0.0508
Cargador Hp Envy Pavilion G1 19.5v 3.33a 65w Punta Azul	0.8668	0.1330	1.0138	2.0137	0.0016	0.0507

Tabla C.4 Resultados para la categoría Cargadores y Fuentes con SVD

Nombre	Similitud Texto	Similitud Numérica	Similitud Atributos	Similitud Estándar	Similitud L1	Similitud L2
Cargador Notebook Toshiba Satellite 19v 3.42a /packard Bell	0.7188	0.1972	0.1501	1.0662	0.0032	0.0851
Cargador Hp Envy Pavilion G1 19.5v 3.33a 65w Punta Azul	0.8668	0.1330	0.0514	1.0514	0.0032	0.0839
Cargador Notebook Compatible Acer Aspire 19v 3.42a 5.5x1.7mm	0.8370	0.0969	0.0895	1.0234	0.0031	0.0817
Cargador Notebook Acer Chromebook 19v 3.42a 65w 3.0 X 1.0mm	0.6395	0.2576	0.1069	1.0042	0.0031	0.0802
Cargador Para Acer 19v 3.42a 65w /3.0*1.0mm Punta Fina	0.6053	0.2281	0.1650	0.9985	0.0030	0.0797
Cargador Notebook Acer Aspire 19v 3.42a 65w	0.8568	-0.0461	0.1838	0.9945	0.0030	0.0794
Cargador Notebook Acer 19v 3.42a Compatible Con 19v 2.37a	0.3423	0.4529	0.1457	0.9410	0.0029	0.0751
Cargador Notebook Hp Pavilion Split 19.5v 2.31a 45w X360	0.7479	0.1652	0.0086	0.9218	0.0028	0.0736
Cargador Original Notebook Acer Aspire E1-531 Nuevo	0.7175	-0.0274	0.2166	0.9067	0.0028	0.0724
Acer - Cargador Acer Original 19v - 3,42a 65w Notebook	0.6821	0.0778	0.1415	0.9015	0.0027	0.0720

Tabla C.5 Resultados para la categoría Cargadores y Fuentes con Umbral

C.2. Baterías

Nombre	Similitud Texto	Similitud Numérica	Similitud Atributos	Similitud Estándar	Similitud L1	Similitud L2
Batería Para iPhone 6s + Adhesivo - Dcompras	0.7828	0.2915	0.0	1.0744	0.0045	0.1146
Batería Compatible Con Batería Huawei Mate 20 Lite	0.7245	0.2280	0.0	0.9525	0.0040	0.1016
Batería Compatible Con iPhone 8g	0.7061	0.2308	0.0	0.9370	0.0040	0.0999
Batería Compatible Con iPhone 6s	0.7289	0.1350	0.0	0.8640	0.0036	0.0922
Batería Para iPhone 6 + Adhesivo - Dcompras	0.5898	0.2702	0.0	0.8601	0.0036	0.0917
Batería Compatible Con iPhone X	0.7189	0.1365	0.0	0.8554	0.0036	0.0912
Batería Compatible Con iPhone 8 Plus	0.7081	0.1355	0.0	0.8436	0.0036	0.0900
Batería Compatible Con Samsung A30s	0.7046	0.1358	0.0	0.8404	0.0035	0.0896
Batería Compatible Con Samsung A20	0.7038	0.1350	0.0	0.8389	0.0035	0.0895
Batería Compatible Con Samsung S10	0.7023	0.1340	0.0	0.8364	0.0035	0.0892

Tabla C.6 Resultados para la categoría Baterías

Nombre	Similitud Texto	Similitud Numérica	Similitud Atributos	Similitud Estándar	Similitud L1	Similitud L2
Bateria Para iPhone 6s + Adhesivo - Dcompras	0.8407	0.2915	0.0	1.1323	0.0040	0.1058
Batería Compatible Con iPhone 8g	0.7229	0.2308	0.0	0.9538	0.0034	0.0891
Bateria Para iPhone 6 + Adhesivo - Dcompra	0.6763	0.2702	0.0	0.9465	0.0034	0.0884
Batería Compatible Con Batería Huawei Mate 20 Lite	0.7013	0.2280	0.0	0.9294	0.0033	0.0868
Batería Compatible Con iPhone 6s	0.7560	0.1350	0.0	0.8911	0.0032	0.0832
Bateria Para iPhone 7 Plus + Adhesivo - Dcompras	0.6820	0.2044	0.0	0.8864	0.0032	0.0828
Bateria Para iPhone 8 + Adhesivo - Dcompras	0.6820	0.1962	0.0	0.8783	0.0031	0.0820
Batería Compatible Con iPhone 8 Plus	0.7342	0.1355	0.0	0.8697	0.0031	0.0812
Batería Compatible Con Samsung A30s	0.7331	0.1358	0.0	0.8689	0.0031	0.0812
Batería Compatible Con Samsung A20	0.7322	0.1350	0.0	0.8673	0.0031	0.0810

Tabla C.7 Resultados para la categoría Baterías con Stemming

Nombre	Similitud Texto	Similitud Numérica	Similitud Atributos	Similitud Estándar	Similitud L1	Similitud L2
Bateria Para iPhone 6s + Adhesivo - Dcompra	0.8002	0.2915	0.0	1.0917	0.0038	0.0995
Bateria Para iPhone 6 + Adhesivo - Dcompras	0.7494	0.2702	0.0	1.0197	0.0035	0.0929
Bateria Para iPhone 7 Plus + Adhesivo - Dcompras	0.7929	0.2044	0.0	0.9973	0.0035	0.0909
Bateria Para iPhone 8 + Adhesivo - Dcompras	0.7885	0.1962	0.0	0.9848	0.0034	0.0897
Bateria Para iPhone X + Adhesivo - Dcompras	0.7959	0.1729	0.0	0.9688	0.0034	0.0883
Batería Para iPhone SE 2020 + Adhesivo Regalo - Dcompras	0.7955	0.1390	0.0	0.9346	0.0032	0.0851
Batería Compatible Con Batería Huawei Mate 20 Lite	0.6918	0.2280	0.0	0.9198	0.0032	0.0838
Batería Compatible Con iPhone 8g	0.6736	0.2308	0.0	0.9044	0.0031	0.0824
Bateria Para iPhone 8 Plus + Adhesivo - Dcompras	0.7508	0.1382	0.0	0.8891	0.0031	0.0810
Bateria Para iPhone 7 + Adhesivo - Dcompras	0.7828	0.0779	0.0	0.8607	0.0030	0.0784

Tabla C.8 Resultados para la categoría Baterías con Palabras Claves

Nombre	Similitud Texto	Similitud Numérica	Similitud Atributos	Similitud Estándar	Similitud L1	Similitud L2
Bateria Para iPhone 6s + Adhesivo - Dcompras	0.7828	0.2915	0.0	1.0744	0.0045	0.1146
Batería Compatible Con Batería Huawei Mate 20 Lite	0.7245	0.2280	0.0	0.9525	0.0040	0.1016
Batería Compatible Con iPhone 8g	0.7061	0.2308	0.0	0.9370	0.0040	0.0999
Batería Compatible Con iPhone 6s	0.7289	0.1350	0.0	0.8640	0.0036	0.0922
Bateria Para iPhone 6 + Adhesivo - Dcompras	0.5898	0.2702	0.0	0.8601	0.0036	0.0917
Batería Compatible Con iPhone X	0.7189	0.1365	0.0	0.8554	0.0036	0.0912
Batería Compatible Con iPhone 8 Plus	0.7081	0.1355	0.0	0.8436	0.0036	0.0900
Batería Compatible Con Samsung A30s	0.7046	0.1358	0.0	0.8404	0.0035	0.0896
Batería Compatible Con Samsung A20	0.7038	0.1350	0.0	0.8389	0.0035	0.0895
Batería Compatible Con Samsung S10	0.7023	0.1340	0.0	0.8364	0.0035	0.0892

Tabla C.9 Resultados para la categoría Baterías con SVD

Nombre	Similitud Texto	Similitud Numérica	Similitud Atributos	Similitud Estándar	Similitud L1	Similitud L2
Bateria Compatible Samsung Galaxy Note 20 N980 4300 Mah	0.6414	0.0873	0.0	0.7288	0.0031	0.0777
Batería Motorola G4/g4 Plus	0.5920	0.1334	0.0	0.7255	0.0031	0.0774
Bateria Para iPhone 5s + Adhesivo - Dcompras	0.5950	0.0858	0.0	0.6808	0.0029	0.0726
Batería Para iPhone SE 2020 + Adhesivo Regalo - Dcompras	0.5180	0.1390	0.0	0.6571	0.0028	0.0701
Batería Para iPhone XS Max + Adhesivo - Dcompras	0.5799	0.0752	0.0	0.6552	0.0027	0.0699
Bateria Para iPhone SE + Adhesivo - Dcompras	0.5833	0.0670	0.0	0.6503	0.0027	0.0694
Bateria Para iPhone 7 + Adhesivo - Dcompras	0.5356	0.0779	0.0	0.6136	0.0026	0.0654
Bateria Para iPhone 6s Plus + Adhesivo - Dcompras	0.6232	-0.0499	0.0	0.5733	0.0024	0.0611
Bateria Compatible Con Huawei Mate 10 Lite / P30 Lite	0.1468	0.4257	0.0	0.5725	0.0024	0.0610
Bateria Para iPhone XR + Adhesivo - Dcompras	0.5417	0.0269	0.0	0.5686	0.0024	0.0606

Tabla C.10 Resultados para la categoría Baterías con Umbral

C.3. Pantallas LCD

Nombre	Similitud Texto	Similitud Numérica	Similitud Atributos	Similitud Estándar	Similitud L1	Similitud L2
Pantalla Para Samsung A32 4g Con Marco + Mica Regalo - Ltc	1.7208	0.8395	1.1589	3.7193	0.0065	0.1598
Pantalla iPhone X Calidad Incell + Lamina Completa De Regalo	1.2018	0.9345	0.3828	2.5192	0.0044	0.1082
Pantalla Para iPhone 11 Calidad Incell + Lamina De Regalo	1.2020	0.8954	0.3877	2.4852	0.0043	0.1068
Pantalla Para Redmi Note 11/ 11s 4g	1.0446	0.8267	0.6026	2.4740	0.0043	0.1063
Pantalla Compatible Samsung A32 4g Oled + Lamina Full Mobile	1.1809	0.6662	0.6031	2.4502	0.0042	0.1053
Pantalla Para Samsung A32 4g Con Marco + Mica Regalo - Ltc	1.6931	0.0592	0.5428	2.2952	0.0040	0.0986
Pantalla Samsung A32 4g C/m Oled Ztel	0.7904	1.0072	0.4467	2.2444	0.0039	0.0964
Pantalla Para Huawei P40 Lite + Mica De Regalo - Ltc	1.1952	0.7396	0.2235	2.1585	0.0037	0.0927
Pantalla Para Samsung Galaxy A03 + Mica De Regalo	1.1811	0.8475	0.0704	2.0991	0.0036	0.0902
Pantalla Para Samsung A12 Con Marco + Mica Regalo - Ltc	1.2808	0.7298	0.0732	2.0839	0.0036	0.0895

Tabla C.11 Resultados para la categoría Pantallas sin Técnicas

Nombre	Similitud Texto	Similitud Numérica	Similitud Atributos	Similitud Estándar	Similitud L1	Similitud L2
Pantalla Para Samsung A32 4g Con Marco + Mica Regalo - Ltc	1.8466	0.8395	1.1628	3.8490	0.0055	0.1428
Pantalla Para Redmi Note 11/ 11s 4g	1.1533	0.8267	0.6771	2.6572	0.0038	0.0986
Pantalla Compatible Samsung A32 4g Oled + Lamina Full Mobile	1.2935	0.6662	0.6712	2.6309	0.0037	0.0976
Pantalla Para iPhone 11 Calidad Incell + Lamina	1.3225	0.8954	0.4090	2.6271	0.0037	0.0974
Pantalla Samsung A32 4g C/m Oled Ztel	0.9880	1.0072	0.4794	2.4747	0.0035	0.0918
Pantalla Para Samsung A32 4g Con Marco + Mica Regalo - Ltc	1.8157	0.0592	0.5843	2.4593	0.0035	0.0912
Pantalla iPhone X Calidad Incell + Lamina Completa	1.2797	0.9345	0.0916	2.3059	0.0033	0.0855
Pantalla Para Huawei P40 Lite + Mica De Regalo - Ltc	1.3223	0.7396	0.2270	2.2891	0.0032	0.0849
Pantalla Para Samsung Galaxy A03 + Mica De Regalo	1.2777	0.8475	0.0730	2.1983	0.0031	0.0815
Pantalla Para Samsung A32 Versión 4g Incell (Con Marco)	1.1547	0.5047	0.5271	2.1867	0.0031	0.0811

Tabla C.12 Resultados para la categoría Pantallas con Stemming

Nombre	Similitud Texto	Similitud Numérica	Similitud Atributos	Similitud Estándar	Similitud L1	Similitud L2
Pantalla Para Samsung A32 4g Con Marco + Mica Regalo - Ltc	2.2320	0.8395	1.1589	4.2305	0.0044	0.1135
Pantalla Compatible Samsung A32 4g Oled + Lamina Full Mobile	1.9821	0.6662	0.6031	3.2514	0.0033	0.0872
Pantalla Samsung A32 4g C/m Oled Ztel	1.7491	1.0072	0.4467	3.2031	0.0033	0.0859
Pantalla Para Samsung A53 5g Oled Con Marco / Ztel	1.6934	1.1921	0.0111	2.8967	0.0030	0.0777
Pantalla Para Samsung A12 Con Marco + Mica Regalo - Ltc	2.0394	0.7298	0.0732	2.8424	0.0029	0.0763
Pantalla Para Samsung Galaxy A03 + Mica De Regalo	1.9024	0.8475	0.0704	2.8204	0.0029	0.0757
Pantalla Para Samsung A32 4g Con Marco + Mica Regalo - Ltc	2.2168	0.0592	0.5428	2.8189	0.0029	0.0756
Pantalla Para Samsung A21s + Mica De Regalo - Ltc	1.9311	0.7536	0.0723	2.7571	0.0028	0.0740
Pantalla Para Samsung Galaxy A03 + Mica De Regalo - Ltc	1.9661	0.7186	0.0704	2.7551	0.0028	0.0739
Pantalla Para Samsung A51 Con Marco + Mica Regalo - Ltc	1.9858	0.6968	0.0684	2.7511	0.0028	0.0738

Tabla C.13 Resultados para la categoría Pantallas con Palabras Claves

Nombre	Similitud Texto	Similitud Numérica	Similitud Atributos	Similitud Estándar	Similitud L1	Similitud L2
Pantalla Para Samsung A32 4g Con Marco + Mica Regalo - Ltc	1.7208	0.8395	2.8103	5.3708	0.0018	0.0552
Pantalla iPhone X Calidad Incell + Lamina Completa De Regalo	1.2018	0.9345	2.8095	4.9460	0.0016	0.0508
Pantalla Para iPhone 11 Calidad Incell + Lamina De Regalo	1.2020	0.8954	2.8135	4.9110	0.0016	0.0505
Pantalla Para Samsung Galaxy A03 + Mica De Regalo	1.1811	0.8475	2.8116	4.8403	0.0016	0.0498
Pantalla Para Samsung A51 Con Marco + Mica Regalo - Ltc	1.3068	0.6968	2.8118	4.8155	0.0016	0.0495
Pantalla Para Samsung A12 Con Marco + Mica Regalo - Ltc	1.2808	0.7298	2.8030	4.8137	0.0016	0.0495
Pantalla Para Samsung Galaxy A03 + Mica De Regalo - Ltc	1.2667	0.7186	2.8116	4.7970	0.0016	0.0493
Pantalla Para Samsung A21s + Mica De Regalo - Ltc	1.2029	0.7536	2.8124	4.7690	0.0016	0.0490
Pantalla Para Huawei P40 Lite + Mica De Regalo - Ltc	1.1952	0.7396	2.8122	4.7471	0.0016	0.0488
Pantalla Para Redmi Note 11/ 11s 4g	1.0446	0.8267	2.8051	4.6765	0.0015	0.0481

Tabla C.14 Resultados para la categoría Pantallas con SVD

Nombre	Similitud Texto	Similitud Numérica	Similitud Atributos	Similitud Estándar	Similitud L1	Similitud L2
Pantalla Para Samsung A53 5g Oled Con Marco /Ztel	0.6238	1.1921	0.0386	1.8546	0.0032	0.0797
Pantalla Para Samsung Galaxy A20s + Mica De Regalo	1.2700	0.5114	0.0694	1.8509	0.0032	0.0795
Pantalla Para iPhone XR Calidad Incell + Lamina De Regalo	0.9697	0.7798	0.0694	1.8190	0.0031	0.0781
Pantalla Para Motorola Moto One Fusion + Mica Regalo - Ltc	1.1462	0.6250	0.0461	1.8174	0.0031	0.0781
Pantalla Compatible Samsung A32 4g Ojo Oled Y Huella	1.0802	0.1308	0.6031	1.8142	0.0031	0.0779
Pantalla Para P30 Lite Completa+ Pegamento +kit De Desarme	0.7014	0.5757	0.5183	1.7954	0.0031	0.0771
Pantalla Para Samsung Galaxy A13 4g Mas Mica De Regalo	1.2106	0.4683	0.0556	1.7347	0.0030	0.0745
Pantalla Para iPhone XR Original + Mica De Regalo - Ltc	1.1782	0.4576	0.0694	1.7053	0.0029	0.0732
Pantalla Para Samsung A70 Con Marco + Mica Regalo - Ltc	1.2600	0.3597	0.0660	1.6858	0.0029	0.0724
Pantalla Para iPhone 11 Calidad Incell + Pegamento+envio	0.6421	0.5743	0.4518	1.6683	0.0029	0.0717

Tabla C.15 Resultados para la categoría Pantallas con Umbral

C.4. Discos Duros y SSD

Nombre	Similitud Texto	Similitud Numérica	Similitud Atributos	Similitud Estándar	Similitud L1	Similitud L2
Disco Sólido Ssd Interno Western Digital Wds500g2b0a 500gb Azul	1.5181	0.2928	0.0	1.8109	0.0274	0.2526
Disco Sólido Ssd Interno Western Digital Wd Green Wds480g2g0a 480gb Verde	1.3132	0.1924	0.0	1.5057	0.0228	0.2100
Disco Duro Interno Western Digital Wd Purple Wd20purx 2tb Púrpura	1.0035	0.3178	0.0	1.3214	0.0200	0.1843
Disco Sólido Ssd Interno Western Digital Wd Green Wds240g2g0a 240gb Verde	1.2212	0.0732	0.0	1.2945	0.0196	0.1805
Disco Duro Interno Western Digital Wd Purple Wd60purz 6tb Púrpura	0.9739	0.3126	0.0	1.2865	0.0195	0.1794
Disco Sólido Ssd Interno Western Digital Wd Green Wds480g3g0a 480gb Negro	1.1306	0.1511	0.0	1.2817	0.0194	0.1788
Disco Sólido Ssd Interno Western Digital Wd Green Wds120g2g0a 120gb Verde	1.2289	0.0182	0.0	1.2472	0.0189	0.1739
Ssd M.2 2tb Wd Black Pcie Gen4 Nvme Consola Ps5	0.9117	0.2894	0.0	1.2012	0.0182	0.1675
Disco Duro Interno Western Digital Wd Purple 4tb Wd42purz	0.8349	0.3222	0.0	1.1572	0.0175	0.1614
Disco Duro Interno Western Digital Wd Purple Surveillance Wd40purz 4tb	0.8134	0.3360	0.0	1.1495	0.0174	0.1603

Tabla C.16 Resultados para la categoría Discos

Nombre	Similitud Texto	Similitud Numérica	Similitud Atributos	Similitud Estándar	Similitud L1	Similitud L2
Disco Sólido Ssd Interno Western Digital Wds500g2b0a 500gb Azul	1.5659	0.2928	0.0	1.8587	0.0253	0.2432
Disco Sólido Ssd Interno Western Digital Wd Green Wds480g2g0a 480gb Verde	1.3389	0.1924	0.0	1.5314	0.0208	0.2004
Disco Sólido Ssd Interno Western Digital Wd Green Wds480g3g0a 480gb Negro	1.2195	0.1511	0.0	1.3707	0.0187	0.1793
Disco Duro Interno Western Digital Wd Purple Wd20purx 2tb Púrpura	1.0503	0.3178	0.0	1.3682	0.0186	0.1790
Disco Sólido Ssd Interno Western Digital Wd Green Wds240g2g0a 240gb Verde	1.2600	0.0732	0.0	1.3333	0.0181	0.1744
Disco Duro Interno Western Digital Wd Purple Wd60purz 6tb Púrpura	1.0191	0.3126	0.0	1.3318	0.0181	0.1742
Disco Sólido Ssd Interno Western Digital Wd Green Wds120g2g0a 120gb Verde	1.2530	0.0182	0.0	1.2713	0.0173	0.1663
Disco Duro Interno Western Digital Wd Purple4tb Wd42purz	0.9071	0.3222	0.0	1.2294	0.0167	0.1609
Ssd M.2 2tb Wd Black Pcie Gen4 Nvme Consola Ps5	0.9349	0.2894	0.0	1.2244	0.0167	0.1602
Disco Duro Interno Western Digital Wd Purple Surveillance Wd40purz 4tb	0.8841	0.3360	0.0	1.2201	0.0166	0.1596

Tabla C.17 Resultados para la categoría Discos con Stemming

Nombre	Similitud Texto	Similitud Numérica	Similitud Atributos	Similitud Estándar	Similitud L1	Similitud L2
Disco Sólido Ssd Interno Western Digital Wds500g2b0a 500gb Azul	1.8114	0.2928	0.0	2.1043	0.0184	0.1810
Disco Sólido Ssd Interno Western Digital Wd Green Wds480g2g0a 480gb Verde	1.7325	0.1924	0.0	1.9250	0.0168	0.1656
Disco Duro Interno Western Digital Wd Purple Wd20purx 2tb Púrpura	1.5794	0.3178	0.0	1.8973	0.0166	0.1632
Disco Duro Interno Western Digital Wd Purple Wd60purz 6tb Púrpura	1.5584	0.3126	0.0	1.8710	0.0163	0.1609
Disco Duro Interno Western Digital Wd Purple 4tb Wd42purz	1.5231	0.3222	0.0	1.8453	0.0161	0.1587
Disco Duro Interno Western Digital Purple Pro 10tb 3.5in	1.4909	0.2403	0.0	1.7312	0.0151	0.1489
Disco Sólido Ssd Interno Western Digital Wd Green Wds240g2g0a 240gb Verde	1.6239	0.0732	0.0	1.6972	0.0148	0.1460
Disco Sólido Ssd Interno Western Digital Wd Green Wds480g3g0a 480gb Negro	1.5447	0.1511	0.0	1.6959	0.0148	0.1459
Disco Duro Interno Western Digital Wd Purple Surveillance Wd40purz 4tb	1.3596	0.3360	0.0	1.6956	0.0148	0.1458
Disco Duro Interno Wd 2.5in 500gb Hdd Sata Iii Open Box	1.6343	0.0578	0.0	1.6922	0.0148	0.1455

Tabla C.18 Resultados para la categoría Discos con Palabras Claves

Nombre	Similitud Texto	Similitud Numérica	Similitud Atributos	Similitud Estándar	Similitud L1	Similitud L2
Disco Sólido Ssd Interno Western Digital Wds500g2b0a 500gb Azul	1.518	0.2928	0.0	1.8109	0.0274	0.2526
Disco Sólido Ssd Interno Western Digital Wd Green Wds480g2g0a 480gb Verde	1.3132	0.1924	0.0	1.5057	0.0228	0.2100
Disco Duro Interno Western Digital Wd Purple Wd20purx 2tb Púrpura	1.0035	0.3178	0.0	1.3214	0.0200	0.1843
Disco Sólido Ssd Interno Western Digital Wd Green Wds240g2g0a 240gb Verde	1.2212	0.0732	0.0	1.2945	0.0196	0.1805
Disco Duro Interno Western Digital Wd Purple Wd60purz 6tb Púrpura	0.9739	0.3126	0.0	1.2865	0.0195	0.1794
Disco Sólido Ssd Interno Western Digital Wd Green Wds480g3g0a 480gb Negro	1.1306	0.1511	0.0	1.2817	0.0194	0.1788
Disco Sólido Ssd Interno Western Digital Wd Green Wds120g2g0a 120gb Verde	1.2289	0.0182	0.0	1.2472	0.0189	0.1739
Ssd M.2 2tb Wd Black Pcie Gen4 Nvme Consola Ps5	0.9117	0.2894	0.0	1.2012	0.0182	0.1675
Disco Duro Interno Western Digital Wd Purple 4tb Wd42purz	0.8349	0.3222	0.0	1.1572	0.0175	0.1614
Disco Duro Interno Western Digital Wd Purple Surveillance Wd40purz 4tb	0.8134	0.3360	0.0	1.1495	0.0174	0.1603

Tabla C.19 Resultados para la categoría Discos con SVD

Nombre	Similitud Texto	Similitud Numérica	Similitud Atributos	Similitud Estándar	Similitud L1	Similitud L2
Disco Duro Interno Western Digital Wd Purple Wd20purx 2tb Púrpura	1.0035	0.3178	0.0	1.3214	0.0200	0.1843
Disco Sólido Ssd Interno Western Digital Wd Green Wds240g2g0a 240gb Verde	1.2212	0.0732	0.0	1.2945	0.0196	0.1805
Disco Duro Interno Western Digital Wd Purple Wd60purz 6tb Púrpura	0.9739	0.3126	0.0	1.2865	0.0195	0.1794
Disco Sólido Ssd Interno Western Digital Wd Green Wds480g3g0a 480gb Negro	1.1306	0.1511	0.0	1.2817	0.0194	0.1788
Disco Sólido Ssd Interno Western Digital Wd Green Wds120g2g0a 120gb Verde	1.2289	0.0182	0.0	1.2472	0.0189	0.1739
Ssd M.2 2tb Wd Black Pcie Gen4 Nvme Consola Ps5	0.9117	0.2894	0.0	1.2012	0.0182	0.1675
Disco Duro Interno Western Digital Wd Purple 4tb Wd42purz	0.8349	0.3222	0.0	1.1572	0.0175	0.1614
Disco Duro Interno Western Digital Wd Purple Surveillance Wd40purz 4tb	0.8134	0.3360	0.0	1.1495	0.0174	0.1603
Disco Sólido Ssd Interno Western Digital Wds100t2b0a 1tb Azul	1.2662	-0.1175	0.0	1.1487	0.0174	0.1602
Disco Duro Interno Wd 2.5in 500gb Hdd Sata Iii Open Box	1.0572	0.0578	0.0	1.1151	0.0169	0.1555

Tabla C.20 Resultados para la categoría Discos con Umbral

Apéndice D

Gráficos Box Plot de similitud ordenados por categoría

D.1. Cargadores y Fuentes

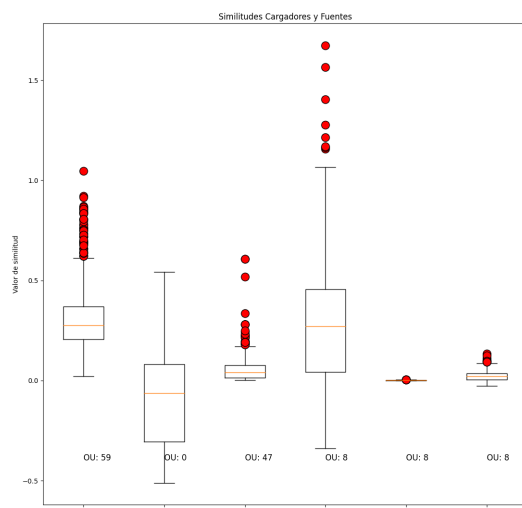


Figura D.1 Similitud sin técnicas

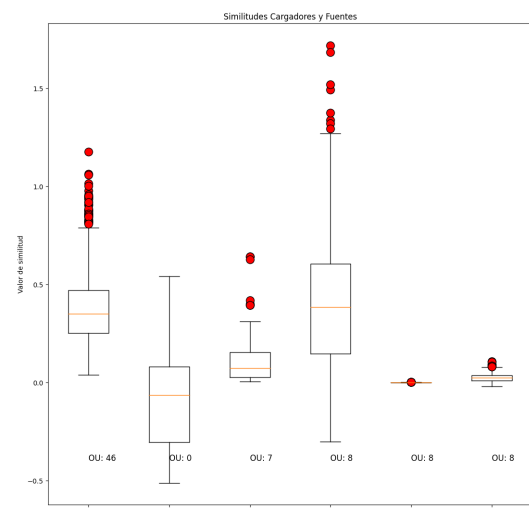


Figura D.2 Similitud con Stemming

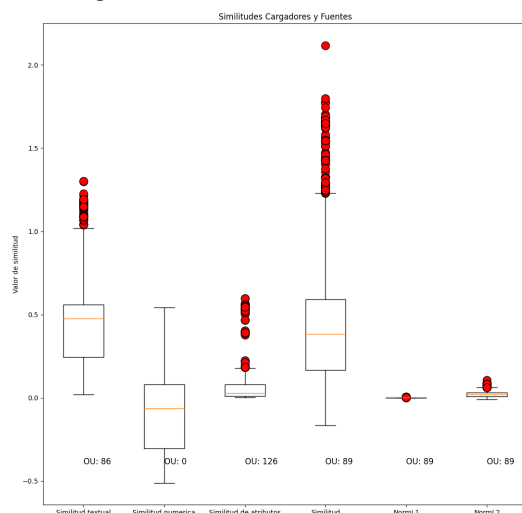


Figura D.3 Similitud con Palabras Claves

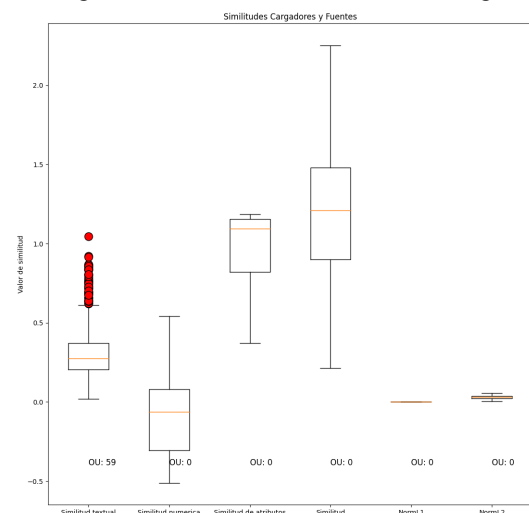


Figura D.4 Similitud con SVD

D.2. Baterías

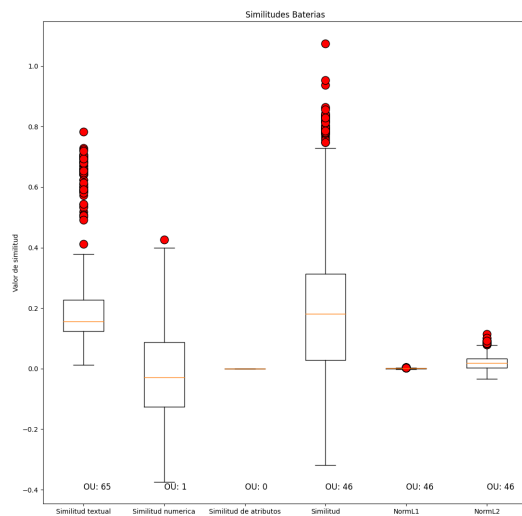


Figura D.5 Similitud sin técnicas

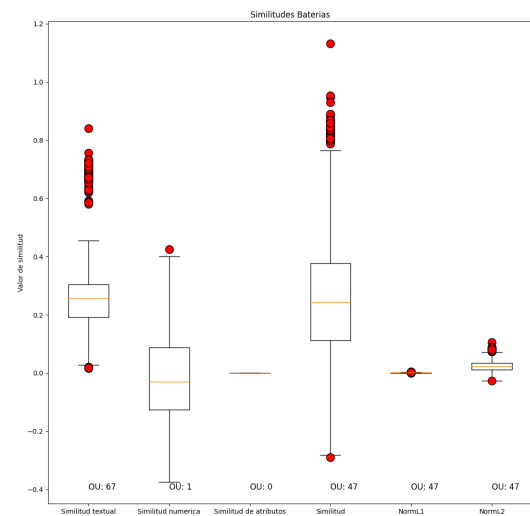


Figura D.6 Similitud con Stemming

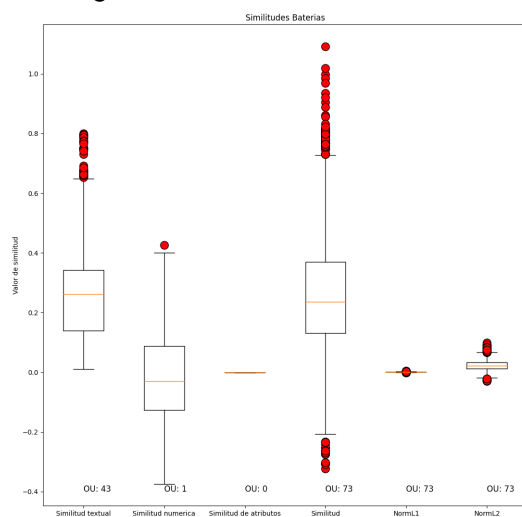


Figura D.7 Similitud con Palabras Claves

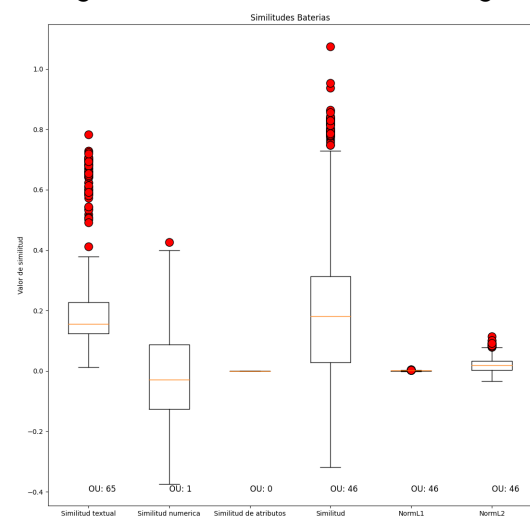


Figura D.8 Similitud con SVD

D.3. Pantallas

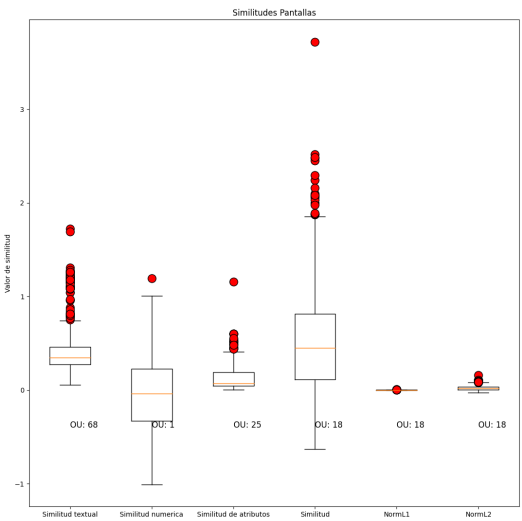


Figura D.9 Similitud sin técnicas

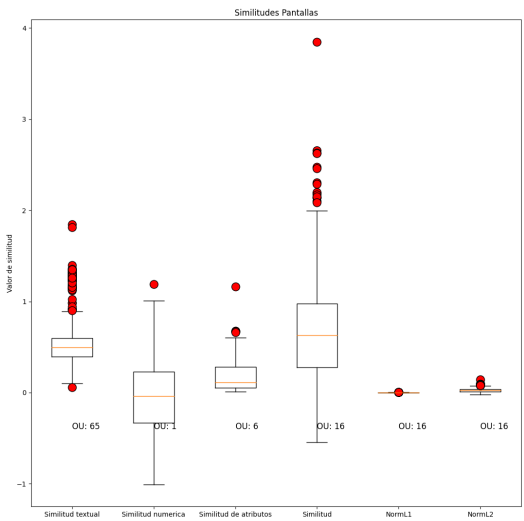


Figura D.10 Similitud con Stemming

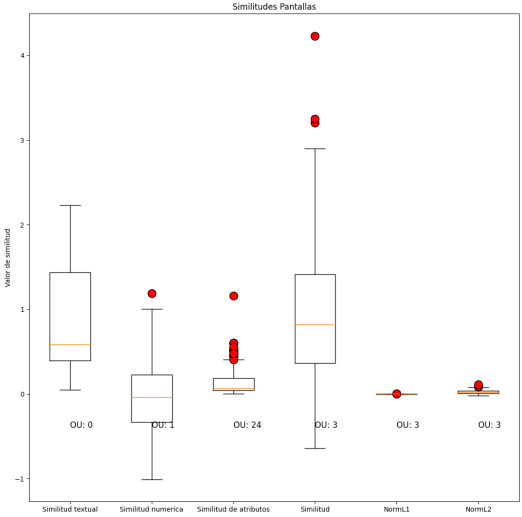


Figura D.11 Similitud con Palabras Claves

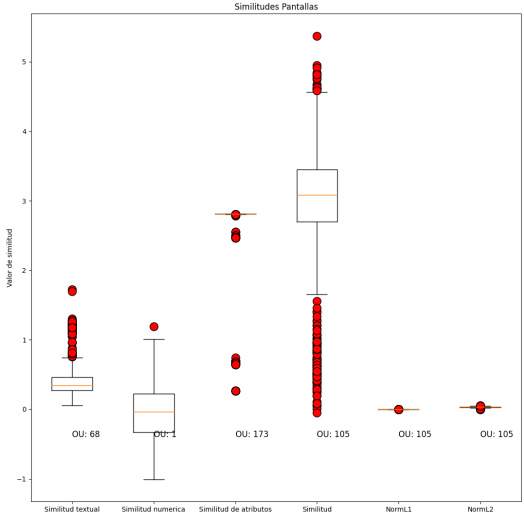


Figura D.12 Similitud con SVD

D.4. Discos Duros y SSD

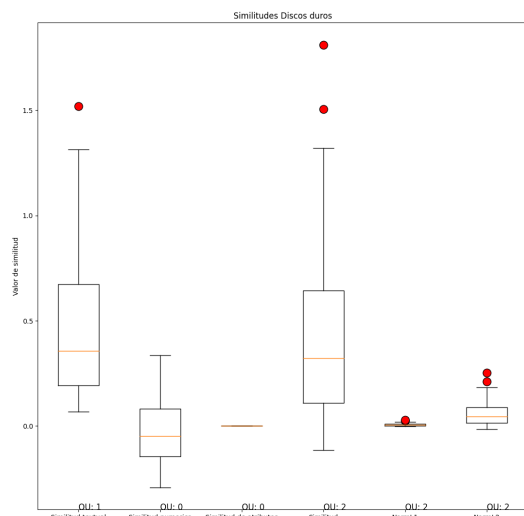


Figura D.13 Similitud sin técnicas

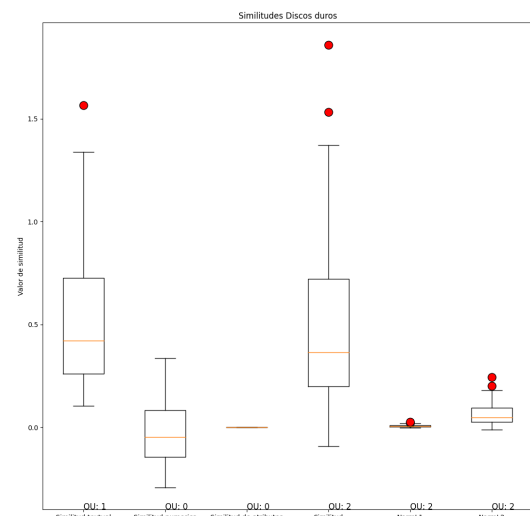


Figura D.14 Similitud con Stemming

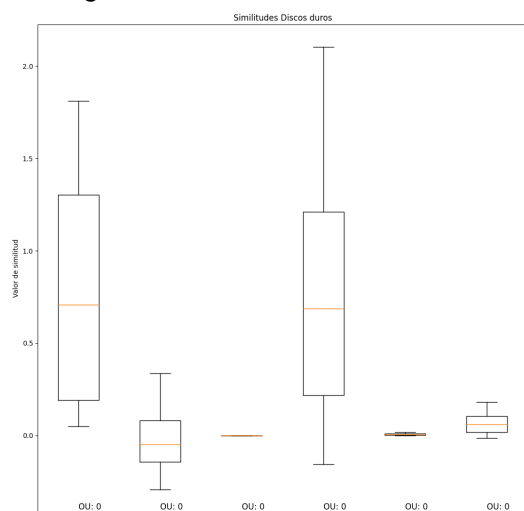


Figura D.15 Similitud con Palabras Claves

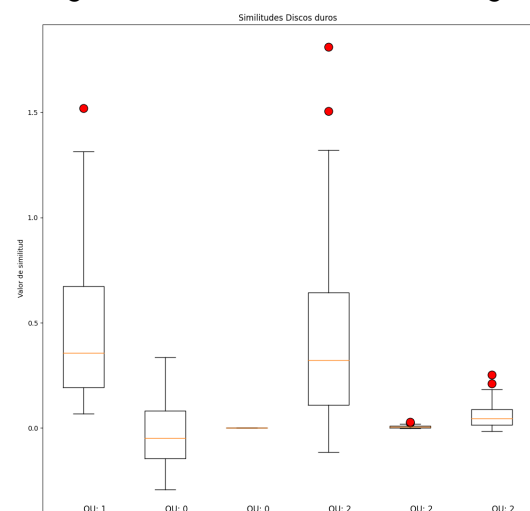


Figura D.16 Similitud con SVD

Apéndice E

Gráfico de barras similitud v/s producto

E.1. Cargadores y Fuentes

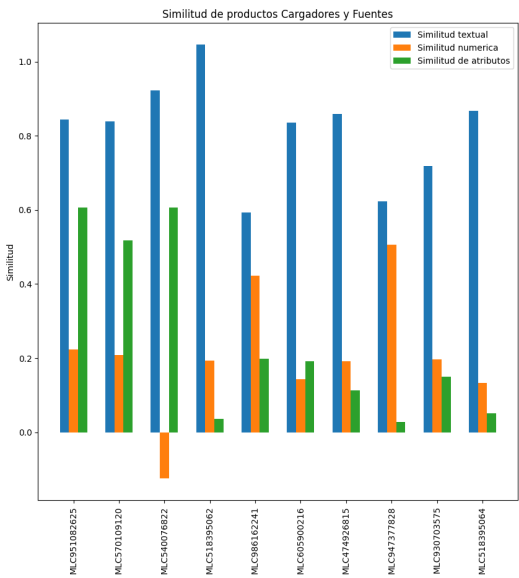


Figura E.1 Similitud sin técnicas

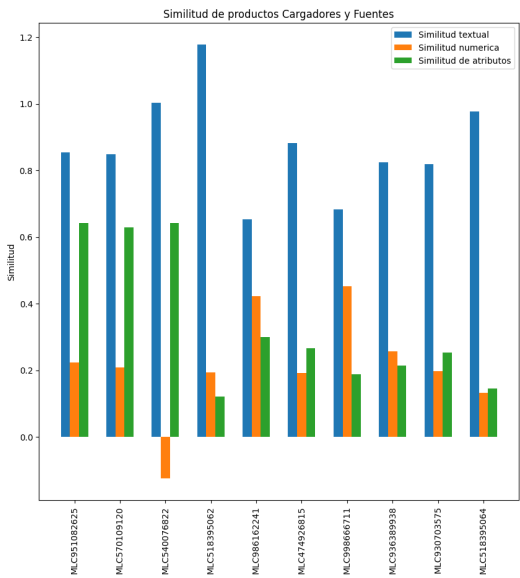


Figura E.2 Similitud con Stemming

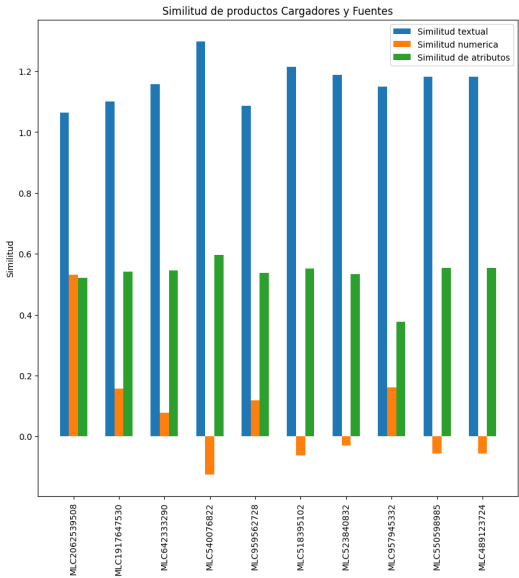


Figura E.3 Similitud con Palabras Claves

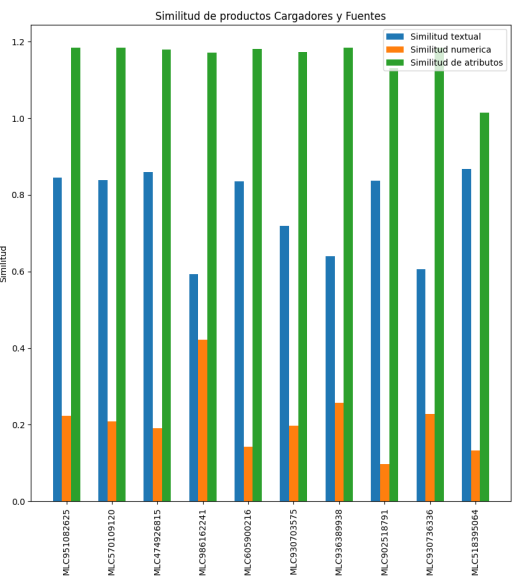


Figura E.4 Similitud con SVD

E.2. Baterías

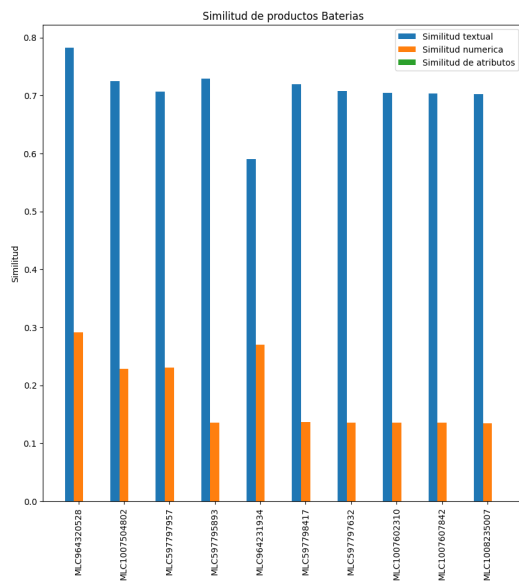


Figura E.5 Similitud sin técnicas

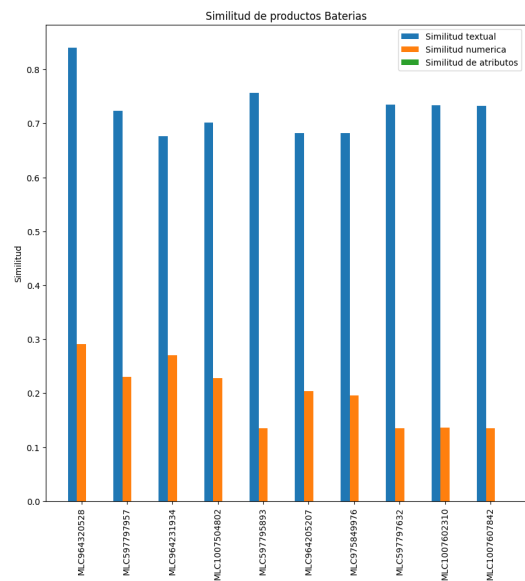


Figura E.6 Similitud con Stemming

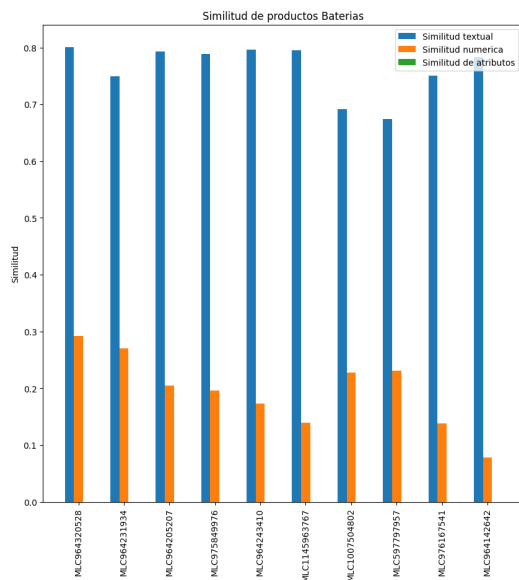


Figura E.7 Similitud con Palabras Claves

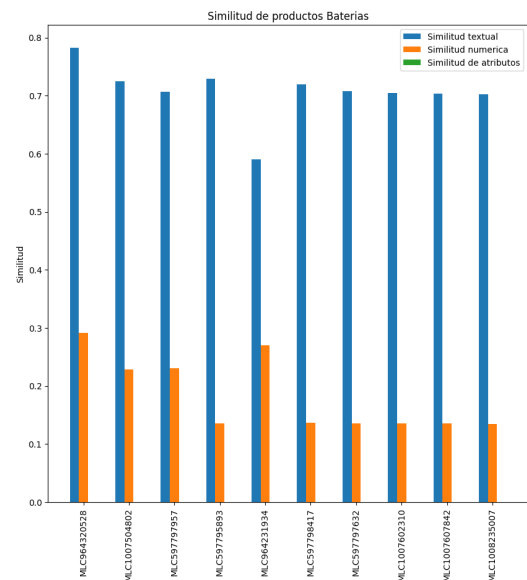


Figura E.8 Similitud con SVD

E.3. Pantallas

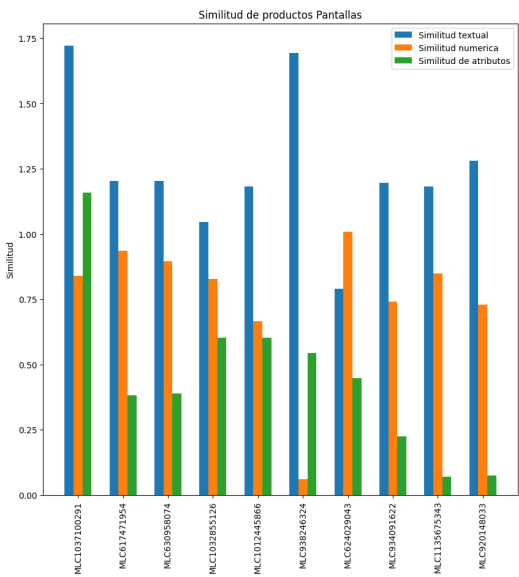


Figura E.9 Similitud sin técnicas

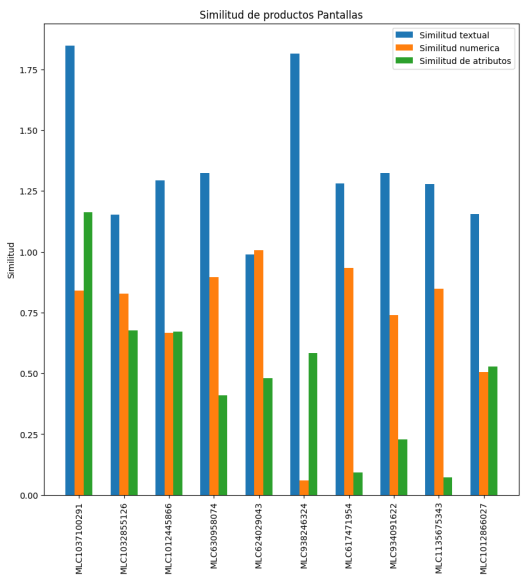


Figura E.10 Similitud con Stemming

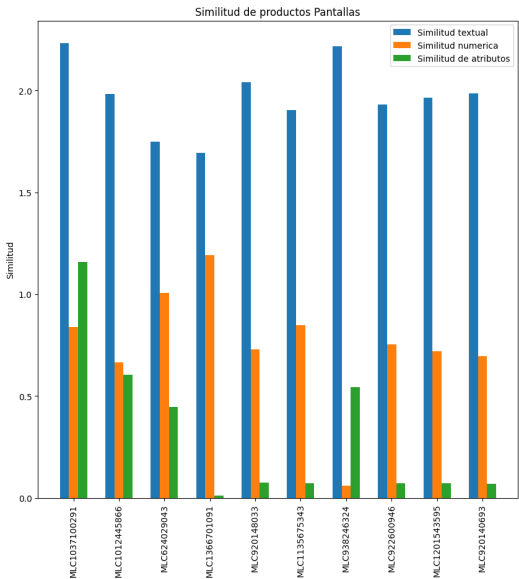


Figura E.11 Similitud con Palabras Claves

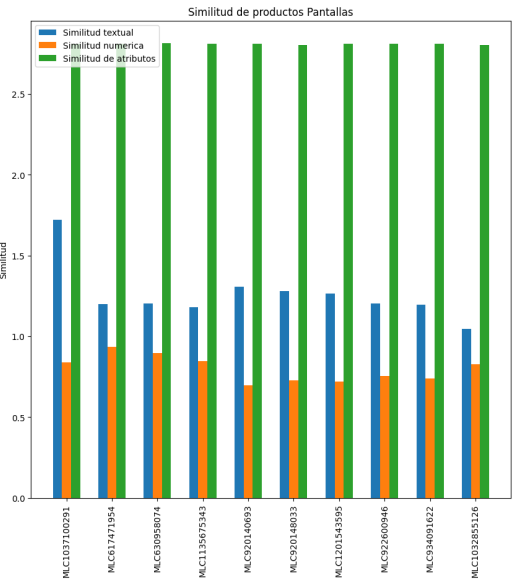


Figura E.12 Similitud con SVD

E.4. Discos Duros y SSD

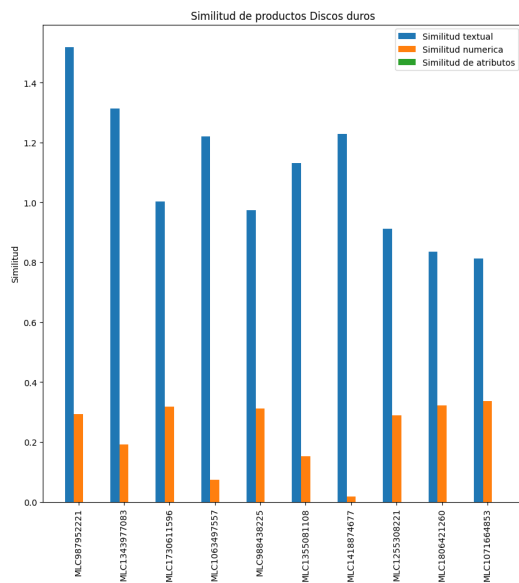


Figura E.13 Similitud sin técnicas

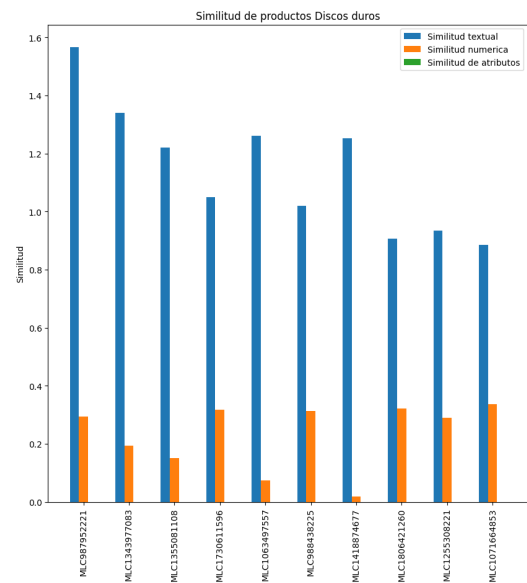


Figura E.14 Similitud con Stemming

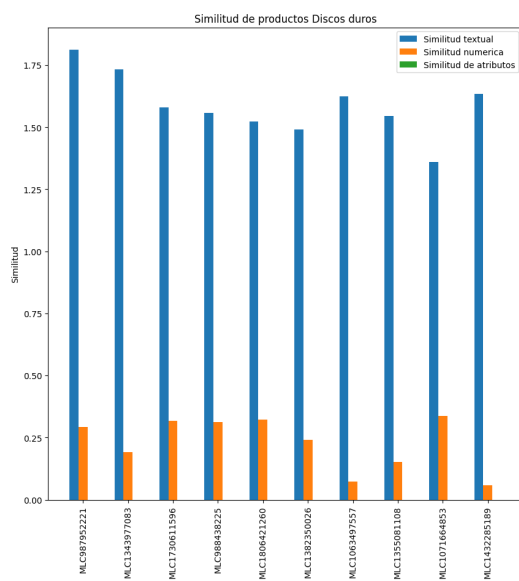


Figura E.15 Similitud con Palabras Claves

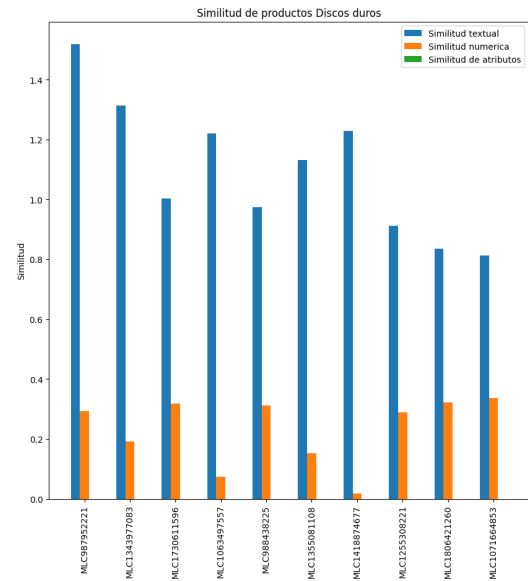


Figura E.16 Similitud con SVD

Apéndice F

Resultados de evaluación de Sistema de recomendación

F.1. Cargadores y Fuentes

Técnica	Distancia Euclideana	Precision	Productos	Recall	Productos
Sin Técnicas	4.0772	60 %	'MLC951082625' (1) 'MLC570109120' (2) 'MLC540076822' (3) 'MLC986162241' (5) 'MLC605900216' (6) 'MLC474926815' (7)	100 %	'MLC951082625' (1) 'MLC540076822' (3) 'MLC518395062' (4)
Stemming	4.5443	60 %	'MLC951082625' (1) 'MLC570109120' (2) 'MLC540076822' (3) 'MLC986162241' (5) 'MLC474926815' (6) 'MLC998666711' (7)	100 %	'MLC951082625' (1) 'MLC540076822' (3) 'MLC518395062' (4)
Palabras Claves	5.5961	70 %	'MLC2062539508' (1) 'MLC642333290' (3) 'MLC540076822' (4) 'MLC959562728' (5) 'MLC518395102' (6) 'MLC957945332' (8) 'MLC550598985' (9)	33 %	'MLC540076822'
SVD	6.7489	70 %	'MLC951082625' (1) 'MLC570109120' (2) 'MLC474926815' (3) 'MLC986162241' (4) 'MLC605900216' (5) 'MLC902518791' (8) 'MLC930736336' (9)	33 %	'MLC951082625'
Umbral	3.1072	40 %	'MLC902518791' (3) 'MLC930736336' (5) 'MLC959555891' (6) 'MLC998666711' (7)	0 %	

Tabla F.1 Resultados de sistema para Cargadores y Fuentes

F.2. Baterías

Técnica	Distancia Euclidea	Precision	Productos	Recall	Productos
Sin Técnicas	2.8248	40 %	'MLC964320528' (1) 'MLC597795893' (4) 'MLC964231934' (5) 'MLC1007607842' (9)	100 %	'MLC964320528' (1)
Stemming	2.9269	60 %	'MLC964320528' (1) 'MLC964231934' (3) 'MLC597795893' (5) 'MLC964205207' (6) 'MLC975849976' (7) 'MLC1007607842' (10)	100 %	'MLC964320528' (1)
Palabras Claves	3.0338	80 %	'MLC964320528' (1) 'MLC964231934' (2) 'MLC964205207' (3) 'MLC975849976' (4) 'MLC964243410' (5) 'MLC1145963767' (6) 'MLC976167541' (9) 'MLC964142642' (10)	100 %	'MLC964320528' (1)
SVD	2.8248	40 %	'MLC964320528' (1) 'MLC597795893' (4) 'MLC964231934' (5) 'MLC1007607842' (9)	100 %	'MLC964320528' (1)
Umbral	2.0400	30 %	'MLC964200970' (3) 'MLC964149980' (8) 'MLC964320970' (10)	0 %	

Tabla F.2 Resultados de sistema para Baterías

F.3. Pantallas LCD

Técnica	Distancia Euclideana	Precision	Productos	Recall	Productos
Sin Técnicas	7.8864	70 %	'MLC1037100291' (1) 'MLC617471954' (2) 'MLC1012445866' (5) 'MLC938246324' (6) 'MLC624029043' (7) 'MLC1135675343' (9) 'MLC920148033' (10)	75 %	'MLC1037100291' (1) 'MLC1032855126' (4) 'MLC1012445866' (5)
Stemming	8.2489	60 %	'MLC1037100291' (1) 'MLC1012445866' (3) 'MLC624029043' (5) 'MLC938246324' (6) 'MLC1135675343' (9) 'MLC1012866027' (10)	75 %	'MLC1037100291' (1) 'MLC1032855126' (2) 'MLC1012445866' (3)
Palabras Claves	9.6883	80 %	'MLC1037100291' (1) 'MLC1012445866' (2) 'MLC624029043' (3) 'MLC1366701091' (4) 'MLC920148033' (5) 'MLC1135675343' (6) 'MLC922600946' (8) 'MLC1201543595' (9)	50 %	'MLC1037100291' (1) 'MLC1012445866' (2)
SVD	15.4070	60 %	'MLC1037100291' (1) 'MLC1135675343' (4) 'MLC920140693' (5) 'MLC920148033' (6) 'MLC1201543595' (7) 'MLC922600946' (8)	50 %	'MLC1037100291' (1) 'MLC1032855126' (10)
Umbral	5.6157	10 %	'MLC1366701091' (1)	0 %	

Tabla F.3 Resultados de sistema para Pantallas LCD

F.4. Discos Duros y SSD

Técnica	Distancia Euclídeana	Precision	Productos	Recall	Productos
Sin Técnicas	4.2339	50 %	'MLC987952221' (1) 'MLC1343977083' (2) 'MLC1063497557' (4) 'MLC1355081108' (6) 'MLC1418874677' (7)	100 %	'MLC987952221' (1)
Stemming	4.3839	50 %	'MLC987952221' (1) 'MLC1343977083' (2) 'MLC1355081108' (3) 'MLC1063497557' (5) 'MLC1418874677' (7)	100 %	'MLC987952221' (1)
Palabras Claves	5.7561	40 %	'MLC987952221' (1) 'MLC1343977083' (2) 'MLC1063497557' (3) 'MLC1355081108' (8)	100 %	'MLC987952221' (1)
SVD	4.2339	50 %	'MLC987952221' (1) 'MLC1343977083' (2) 'MLC1063497557' (4) 'MLC1355081108' (6) 'MLC1418874677' (7)	100 %	'MLC987952221' (1)
Umbral	3.8655	30 %	'MLC1063497557' (2) 'MLC1355081108' (4) 'MLC1418874677' (5)	0 %	

Tabla F.4 Resultados de sistema para Discos Duros y SSD