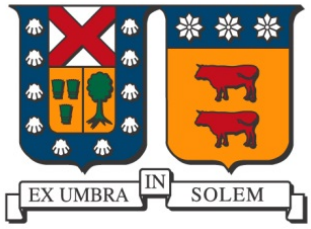
2018

NUEVO ALGORITMO PARA LA ATRIBUCIÓN DE AUTORES BASADO EN N-GRAMAS

GODOY ÁLVAREZ, GLORIA LORETO

UNIVERSIDAD TÉCNICA FEDERICO SANTA MARÍA
DEPARTAMENTO DE INFORMÁTICA
VALPARAÍSO - CHILE



"NUEVO ALGORITMO PARA LA ATRIBUCIÓN DE AUTORES BASADO EN N-GRAMAS"

GLORIA LORETO GODOY ALVAREZ

MEMORIA PARA OPTAR AL TÍTULO DE
INGENIERO CIVIL EN INFORMÁTICA

Profesor Guía: Claudio E. Torres, Ph.D.
Profesora Correferente: Raquel Pezoa, Ph.D.

Junio - 2018

DEDICATORIA

A mi familia, en especial a mis padres, cuyo apoyo incondicional y arduo sacrificio me ha permitido llegar a donde estoy el día de hoy.

RESUMEN

Resumen— Hay solamente una cosa que no podremos esconder por siempre y es a nosotros mismos. Cada persona crea una firma autentica y personal en la que si miramos detenidamente, podremos reconocer sin lugar a duda. El problema de *Atribución de Autores* basa sus necesidades en esta simple afirmación, si podemos identificar patrones de un autor podremos reconocer su autoría. Y el problema de autoría no es ajeno a nuestro día a día, se puede reconocer en áreas tan diversas como en el periodismo, ciencias políticas, criminalística e informática, quien sustituye las convencionales búsquedas manuales y agrupa diferentes propuestas lingüísticas y matemáticas para desarrollar aplicaciones y metodologías que contribuyan con el criterio de identificación.

Este trabajo presenta una investigación de las técnicas actuales de atribución de autores, para luego enfocarse en experimentos que identifiquen el aporte de utilizar n-gramas en el problema de verificación de autores.

Palabras Clave— Atribución de autores, N-grama

ABSTRACT

***Abstract*—** There is only one thing that we can't hide forever and that is ourselves. Each person creates an authentic and personal signature and if we look carefully, we can recognize this signature without a doubt. The problem of *Authorship Attribution* bases on this simple statement, if we can identify patterns of an author we can recognize their authorship. And the problem of authorship is not foreign to our daily life, it can be recognized in diferent areas like journalism, political science, criminalist and computer science, in which we can replaces conventional manual searches for a group of different proposals, linguistic and mathematical, to develop applications and methodologies that contribute with the identification criteria.

This paper presents an investigation of the current techniques on the Authorship Attribution problem, to later pass on experiments that identify the contribution of using n-grams in the problem of author verification.

***Keywords*—** Authorship Attribution, N-gram

ÍNDICE DE CONTENIDOS

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

INTRODUCCIÓN

El estilo de escritura de cada persona es un ejemplo de biometría conductual. Las palabras que ocupamos, el vocabulario y hasta la forma en que estructuramos nuestras oraciones son distintivas y pueden utilizarse para identificar al autor de un trabajo en particular, este problema se conoce como *Authorship Attribution* (Atribución o Identificación de Autores).

En la actualidad existen distintas áreas dedicadas a resolver el problema de identificación de autores, tal es el caso del derecho penal [Howard, 2008], el cual busca la identificación de posibles escritos, notas de rescate o cartas hostigadoras, de las leyes civiles [Ginsburg, 2002] que buscan esclarecer disputas de derecho intelectual, y de la seguridad informática [Zheng *et al.*, 2003] que realiza esfuerzos por identificar autores en correos, detección de *spam*, discriminación de personas o *bots* en comunidades, etc. El desarrollo de técnicas que permiten realizar esta identificación ha llevado a materializar grandes avances informáticos y lingüísticos que han aplicado en técnicas forenses y criminalistas, abriendo nuevos campos de acción y repercutiendo positivamente en numerosas investigaciones [Chaski, 2005, Li *et al.*, 2006].

Debido a sus múltiples aplicaciones es que estos estudios han cobrado tanta popularidad, algunos de ellos enfocándose en mejorar las técnicas utilizadas realizando pruebas controladas en pequeños *sets* de autores, mejorando el desempeño de sus resultados, y otros que específicamente se enfocan en estudiar los tamaños de dichas muestras para realizar pruebas con situaciones más reales y mejorar su precisión [Koppel *et al.*, 2011]. El estudio del segundo grupo se puede analizar en temas de cantidad de caracteres o palabras a utilizar, definiciones de variables o cantidad de autores, por ejemplo, al incrementar el número de autores en las muestras existiría disminución del rendimiento. En un experimento [Luyckx y Daelemans, 2008] de 100 pruebas con muestras aleatorias de 2 autores se obtendría una precisión promedio del 96,9 %, lo que se considera normal dentro de los resultados informados en estudios sobre pequeños grupos de autores, pero al crecer a un experimento con 5-10 autores, nuevamente con 100 pruebas, muestra una disminución del rendimiento, con un 88 % y 82 % de precisión promedio respectivamente, al aumentar a 20 autores esta precisión baja a un 76 % y ya con 50 y 100 autores la precisión es de bajo 52 %.

Pero ¿podría realizarse manualmente este trabajo y prescindir de la tecnología?, la respuesta es no. No, si se quiere lograr precisión y buenos resultados. Aún cuando se entiende que la variedad de escritos puede ser desde enormes libros donde tal vez el problema sea la cantidad de palabras y no el método a elegir, ya que sabremos que el método requerirá una cantidad de texto bastante alta para proporcionar resultados precisos y que no será el mismo a utilizar en textos relativamente cortos, como podría ser una carta o un *tweet*. Aún cuando podamos definir este método a utilizar, será imposible revisar las variables que utilizan cada una de las técnicas a implementar, revisar el léxico, la sintaxis y el análisis de datos relacionados que revisaremos con más detalle en este documento.

Entonces, no es difícil entender porque siguen aumentando las investigaciones sobre el tema, incluso motivando eventos mundiales como es el caso de la competencia PAN [PAN, 2017], o páginas *online* como: [PLA, 2017] la cual detecta el plagio de escritos revisando oraciones y comparándolas con textos, páginas e información que pueda ser encontrada en la web a través de *google*, *yahoo*, *scholar books*, o [PS-, 2017] la cual realiza este servicio de forma más profesional entregando informes con la lista de coincidencias y cálculo del porcentaje de plagio obtenido, o [GIT, 2017] herramienta en java que permite utilizar técnicas de *machine learning* en problemas de atribución de texto y [AIC, 2017] herramienta básica en python que permite realizar experimentos de autoría.

En el presente trabajo se expondrá un marco para el entendimiento de Atribución de Autores y nos enfocaremos en analizar las variaciones del problema y sus técnicas más representativas.

CAPÍTULO 1
DEFINICIÓN DEL PROBLEMA

Si comenzamos estudiando la raíz del problema, podemos generalizar los apartados de análisis de autoría como el problema de identificar a un autor oculto de un documento cualquiera, documento que puede ser desde textos científicos, libros, cartas, textos cortos y hasta programas informáticos. Cada autor y cada documento con sus propias características.

Debido a la cantidad de variables que pueden ser analizadas, el estudio de autoría ha sido un problema desafiante al intentar marcar los límites con los cuales se realizará la búsqueda o comparación. Hay que definir si se quiere comparar sólo por cantidad y similitud de palabras o analizar también los párrafos, el léxico, la sintaxis, el sentido que están tomando o hasta las emociones que el autor esta utilizando. Cada análisis de variables puede ser estudiada por si sola o como dentro de un conjunto, si es que se determina que las variables poseen cierta relación entre ellas, o en si esto deberá ser también determinado. Si un libro es traducido, no sólo el lenguaje influirá sobre éste, sino que también múltiples combinaciones como la puntuación, reglas gramaticales, abreviaciones y diferencias culturales, podrán influenciar para determinar que un escrito posea las mismas características del original y que su autoría posea dichas características propias de su clase. Si en vez de eso, la autoría es desconocida, el problema se puede ver determinado por dichos rasgos como posibles patrones que puedan esclarecer la autoría del escrito, o determinar si dichas características pertenecen a un sólo autor o si existe alguna probabilidad de que múltiples autores estén involucrados.

Es por esto, que el problema general puede ser segregado en tres categorías fundamentales: verificación de autoría, caracterización de autor y detección de estilos de escritura. A continuación se detallan sus definiciones para profundizar en las técnicas que se utilizan en cada problema para el análisis de texto, y entre la que destaca la utilización de n-gramas. El estudio de esta técnica será la que nos permitirá abordar un problema de autoría y comparar resultados utilizando unigrama y bigrama.

1.1. Problema General

Identificación de Autores

Problema que se utiliza para determinar la probabilidad de que un escrito sea de la autoría de un individuo en particular, teniendo en consideración escritos anteriores del individuo.

La Identificación del Autor o *Author Clustering* tiene como objetivo agrupar los escritos de una misma autoría. En la tarea menos complicada de la atribución de conjuntos cerrados se proporciona documentos conocidos de un conjunto pequeño y finito de autores candidatos y se presenta un documento de autoría desconocida, debiendo agrupar el documento des-

conocido en una colección de segmentos que correspondan a dichos autores. En su tarea más complicada se revisa esta agrupación de autores donde no se dan documentos identificados por su autoría si no que se revisa una colección de documentos desconocidos y se debe agrupar dichos documentos por el mismo autor para que cada *cluster* corresponda a un autor diferente, esta tarea puede entenderse como el establecimiento de enlaces (distancias) de autoría entre los documentos.

1.2. Categorías Fundamentales

1. **Verificación de autoría**

   Problema en el cual un texto desconocido puede o no ser de determinado autor. Se tiene cierta cantidad de texto perteneciente a un autor determinado y se desea responder si cierto texto desconocido, que puede tener mayor o menor largo que el original, pertenece o no al autor determinado. Una variable de este problema es poder dar una asignación de pertenencia, en vez de tener una respuesta binaria de si corresponde o no, poder también entregar qué tanto (puede ser representado en un número o porcentaje) el algoritmo califica el texto desconocido dentro de la autoría.

2. **Caracterización del autor**

   Problema en el cual se utilizan técnicas, para determinar ciertos atributos de un individuo, por ejemplo género, edad, entre otros.

   La caracterización o predicción de características como el género o la edad, o posibles combinaciones entre las diferentes características como por ejemplo la variedad de lenguaje con el género [Rangel Pardo *et al.*, 2017], tienen como objetivo enfrentar la tarea de analizar las diferentes cualidades, para crear una descripción a la cual podamos identificar claramente como correspondiente de un autor en particular.

3. **Detección de estilos de escritura**

   Este problema se puede encontrar al intentar determinar a los individuos de un texto de múltiple autoría.

   Mientras que el objetivo de Identificación de Autores y Caracterización de Autor, es determinar cierto patrón que nos ayude a reconocer nuevos textos como pertenecientes o excluyentes a dicha agrupación, el objetivo del *Author Masking* o *Obfuscation Evaluation*, es identificar un estilo de autor sin conocimiento de su identidad o perfil, logrando determinar si existen correlaciones validas dentro de textos o con otros documentos, relacionando sus similitudes y discriminando en base a esta información, la autoría del individuo.

Este trabajo se enfocará en la identificación de autores modificando el algoritmo estudiado en [Reyes, 2016] y con experimentos de verificación de autores de la competencia

[PAN, 2017] para entender su comportamiento y analizar sobre éste la utilización de los n-gramas. Nuestro objetivo principal será analizar los algoritmos en este problema y cuantificar el efecto de bigrama sobre unigrama.

1.3. Objetivos

En el marco de esta memoria, se ha definido una serie de objetivos generales y específicos, los cuales se presentan a continuación:

1. **Objetivo General**
   Analizar algoritmo en el problema de verificación de autores y cuantificar el uso de bigramas sobre unigrama.
2. **Objetivos Específicos**
   - Formar una base de conocimiento de los problemas, métodos y técnicas que permitan detectar la autoría de documentos desconocidos.
   - Adaptar los algoritmos expuestos por [Reyes, 2016] para el problema de verificación de autores.
   - Construir y formalizar una base de datos apropiada que permita la realización de diferentes pruebas entre textos.
   - Definir experimentos y métricas que nos permitan cuantificar los resultados obtenidos para comparar el uso de bigrama y unigrama con los mismos data sets.

CAPÍTULO 2
ESTADO DEL ARTE

La detección de autores es un problema que lleva una larga trayectoria, envolviendo a diferentes áreas de *expertise* y un gran número de aplicaciones[Rudman, 1997]. Se utiliza desde principio de siglo XIX, analizando libros y ha llegado a la actualidad, a la investigación de textos cortos en redes sociales, cumpliendo su propósito de, determinar la propiedad del texto involucrado.

El trabajo de [Mendenhall, 1887] es uno de los primeros estudios que utilizando métodos estadísticos intenta resolver un problema lingüístico para responder sobre la autoría de diferentes autores, como años más tarde deja en evidencia un artículo publicado en 1902 sobre la autoría de Shakespeare en *"Did Marlowe write Shakespeare?"*, problema cuyo argumento no es nuevo y presentaba de forma histórica, evidencia de que obras atribuidas a William Shakespeare podrían pertenecer a diferentes escritores de la época. Mendenhall propone un estudio a la curva de la frecuencia relativa de un número de letras por palabra para determinar la autoría de un escrito. Cada autor mantendría un *"word-spectra"* diferente y característico, el asumía que cada escritor usaba un vocabulario particular y propio, que persistía en el tiempo.

Esta curva, propia de cada estilo de autor y obtenida bajo un criterio cuantitativo llamado discriminadores, son los elementos propios que destacarían en la estilometría, la cual evidenciaría su capacidad para distinguir la autoría.

La *Stylometry* suele confundirse o utilizarse al referirse a *Authorship Atrribution*. La distinción se basa en que, mientras que la primera intenta medir aspectos del estilo de quien escribe, la segunda busca discernir sobre la autoría de dicho texto.

Utilizando métodos basados en estadísticas bayesianas, analizando frecuencias de pequeños grupos de palabras recurrentes como *"and"* y *"to"*, Mosteller y Wallace en 1964 [Mosteller y Wallace, 1964] pudieron discriminar la autoría de *"The Federalist Paper"*, la cual es una serie de 85 ensayos policiales anónimos para persuadir a los *New Yorkinos* de adoptar una nueva constitución en EEUU. Aquellos ensayos escritos con el seudónimo de *Publius*, se lograron comprobar que eran correspondientes a Hamilton, Jay y Madison, politicos ampliamente conocidos en su país. Usando la comparación de largo de palabras y repetición de palabras, pudieron determinar que Jay habría escrito 5, Hamilton 51 , Madison 26, mientras que existían 3 escritos en conjunto. Trabajos como el de [Bailey, 1979] seguiría por la línea de la caracterización y comenzaría a cuantificar y darle estructura a las variables utilizadas. [Mosteller y Wallace, 1964, Peng y Hengartner, 2002] utilizando el largo de la palabra, [Holmes, 1992] investigando la riqueza del vocabulario, [Williams, 1940, Morton, 1965] utilizando el largo de una frase. [Rudman, 1997] calculaba cerca de 1000 diferentes medidas propuestas.

Aunque ninguna de estas metodologías ha demostrado ser lo suficientemente exacta (una afirmación que es analizada a cabalidad en [Chaski, 2005, Li *et al.*, 2006]), su evolución ha logrado validar la búsqueda, a través de métodos estadísticos, de la autoría como algo único y propio de una persona. Su significancia se ha puesto en evidencia en su utilidad en la historia como lo hemos revisado anteriormente.

2.1. Variaciones del Problema

- Author Verification

  [Koppel y Schler, 2004] lo reconoce como un problema más difícil que el de identificación. Lo plantea de esta forma por la explicación de que, en el problema de *Author Attribution* se tienen ejemplos de escritura de un número determinado de autores y se pide determinar cual de dichos autores es quien escribe un texto desconocido. Conociendo desde el momento que el texto pertenecerá a uno de estos autores, en cambio el problema de Verificación de Autores, sólo tiene la información de un escritor y se desea resolver si el texto desconocido le pertenece. Si por ejemplo quisiéramos determinar si un texto desconocido fue escrito por *Shakespeare* o *Marlowe*, sería suficiente usar sus respectivos textos para identificar un modelo de escritura y determinar si el texto desconocido es más propicio para uno que para otro. Si en vez de eso quisiéramos sólo determinar si el mismo texto desconocido fuese de *Shakespeare*, ya no tendríamos como determinar los textos no-pertenecientes y se debería, de alguna forma, determinar el estilo del autor, suponiendo que este no cambia drásticamente en el tiempo y que el texto desconocido es relativamente diferente ya sea por el tema, el tipo de escritura dado por el tiempo del autor o su origen. Esta diferencia se ha resuelto por algunos autores como [Seidman, 2013] con el *Impostors Method*, en el cual al no existir textos con los cuales se pueda comparar la no-pertenecia, se realiza una búsqueda de textos en la *web* que permitan ser una muestra clara de textos no escritos por el autor y con la cual permita una mejor identificación. Dicha aproximación se revisa en la competencia [PAN, 2017] como la metodología ganadora en el año 2013 y su utilización se vuelve más predominante en los años siguientes.

- Plagiarism Detection

  [Reyes, 2016] lo reconoce como un problema que ha ido en aumento principalmente debido al fácil acceso a documentos electrónicos, se plantea como un problema actual, presente en ámbito académico, laboral y científico. Citando a [Reyes, 2016]:

  "*Se describe la acción de plagiar de diferentes formas: una de las más comunes consiste en copiar un fragmento textual sin incluir su fuente, otra forma un poco más elaborada consiste en copiar un fragmento cambiando el orden de las palabras y/o reemplazando palabras por sinónimos y una tercera forma es la inclusión de fragmentos traducidos desde otros documentos en otros idiomas. Finalmente existe el llamado plagio por referencia, el cuál se da cuando una referencia está en un documento y se incluye en otro sin haber leído el origen; Esta forma de plagio es muy difícil de detectar ya que si*

una referencia está bien hecha es casi imposible comprobar si el autor del documento leyó o no el origen de la referencia."

Si en este ejemplo no hubiésemos hecho hincapié en el plagio o no hubiésemos relacionado el texto de origen, el lector para poder identificar dicho plagio debiese estar no sólo en conocimiento del documento si no también recordar lo escrito tal y como lo expuso el autor, lo cual es sumamente complicado para una persona pero no para una maquina.

La identificación de plagio en los textos de [Alzahrani *et al.*, 2012] y [Zu Eissen y Stein, 2006] lo reconocen como el problema de identificar fragmentos apoyándose en la lingüística para resolver los cambios que un texto puede tener en otro y comparando dichas diferencias. Este último lo trabaja al igual que [Reyes, 2016] dentro del ámbito del plagio intrínseco, esta distinción la define [Barrón Cedeño, 2008] como: *detección intrínseca de plagio*, donde no es necesaria la comparación de un documento con otros, y la *detección de plagio con referencia* donde si existe la comparación entre documentos.

- Author Profiling o Characterization

  Se entiende como un problema donde existe un documento o fragmento de éste y se desea identificar al autor, una variación del problema principal que se asemeja a la estilometria ya que utiliza modelos y técnicas sobre el léxico, sintaxis y estructura de los escritos para determinar la caracterización de un referente. Algunas de estas características son exploradas por separados como variaciones de este problema, como por ejemplo determinar el sexo de un determinado autor, su idioma o su edad.

  Un ejemplo de esto se expone en [Akhabue y Lautenbach, 2010] en el cual se trabaja con documentos científicos debido a que estos poseen por lo general dos o más autores explícitamente identificados como contribuyentes por igual, dando el mismo crédito a los autores de la misma publicación e intentando identificar sus contribuciones personales. También se mencionan algunas técnicas especificas de este problema como el concepto de entropía, tratado más ampliamente en [Taylor *et al.*, 2008] donde se describe cada texto con un nivel de entropía propio y donde se utiliza una base de texto conocido del autor para poder comprender y caracterizar patrones de entropía dicha autoría.

2.2. Análisis en Texto

Una forma de entender y analizar un documento es sobre sus mismas reglas gramaticales, las cuales gobiernan el uso del lenguaje utilizado. Si bien cada lengua posee su propia gramática y entender dicho lenguaje (entendiendo sus modos y tiempos) sería elevar considerablemente la complejidad del problema, son estas misma distinciones las que puede dar luces a la identificación de un cierto autor, ayudando a normalizar ciertos patrones como las características estudiadas por *Author Profiling*, encaminando suposiciones de estilos y

caracterización, hasta el punto de acertar con el tiempo/localidad en la cual se ha generado el documento.

Una taxonomía propuesta para cuantificar/caracterizar el estilo del texto fue lo que trabajaron Holmes (1990), Stamatatos, Fakotakis y Kokkinakis (2000) y Zheng (2006). Su revisión de "Style Makers" se enfoca netamente a resolver y entender los requisitos fundamentales para el cálculo computacional que se puede llevar a cabo. Primero con el léxico y los caracteres utilizados por el autor, comenzando a ver un documento como una serie de "tokens" ("word-tokens" o "character-tokens"), seguidos por un análisis más profundo orientados a la asociación de sintaxis y semántica que se puedan determinar.

Entendiendo el tipo de análisis, se tienen las siguientes consideraciones en cada uno de estos aspectos:

1. **Vocabulario/Léxico**

   La primera gran agrupación que se puede utilizar en un análisis de textos es el vocabulario, el cual se define como el conjunto de palabras que conforman un idioma. Esta información puede presentarse en indicios sobre determinadas palabras empleadas provenientes de una región o actividad específica, y aún cuando este conjunto se va modificando con el tiempo, este cambio también suele tomarse con gran importancia ya que da directrices precisas sobre una persona en un momento en particular.

   Un ejemplo de esto podría ser el de "*Beale Ciphers*", la historia de que en 1822 un hombre llamado Thomas J. Beale entierra un tesoro que contenía oro, plata y joyas, y entrega a un hostelero en el condado de Virginia tres textos cifrados, donde indicaría la ubicación, el contenido y la lista con los nombres de los dueños del tesoro. La historia se hace pública cuando en 1960 el hombre que había recibido la carta, la da a conocer.

   La historia ha sido la base de múltiples documentales, libros y discusiones. Y aunque el misterio aún prevalece, en el ámbito académico no quedan muchas dudas. Revisando una de las cartas de 1822 se puede leer la correspondencia enviadas al hotelero diciendo:

   > "They determined to follow them, and secure as many as possible. Keeping well together, they followed their trail for two weeks or more, securing many and stampeding the rest."[BAE, 2017].

   Donde el análisis de [Nickell, 1982] nos revela que, la palabra *"stampede"* llama notoriamente la atención ya que es una palabra que aparece recién en 1844 incluida en el Diccionario de la Universidad de Oxford (principal punto de referencia del estudio etimológico) y que la primera variación de esta palabra (*"stompado"*) databa del año 1826, cuatro años después de la fecha de la carta. Similarmente palabras como *"improvised"* o *"appliance"*, también levantan fuertes dudas sobre su veracidad, tanto por su significado o por su uso (la última, utilizada por 1600 app. había sido catalogada como obsoleta, hasta que se volvió a utilizar por el año 1960).

[Kruh, 1982], en su estudio de estos documentos encriptados y sobre la riqueza de vocabulario empleado evidencia que el autor es extranjero y años después, [Hammer, 1988] concluye que el documento debe haber sido escrito entre 1940 a 1980. Por supuesto estos métodos no son del todo exactos, pero menciona una función que nos cuantificara la diversidad del vocabulario.

2. **Sintaxis**

   Este método se basa en la premisa de que, cada autor tiende a utilizar un orden lógico entre las palabras que utiliza. La relación que crea es una representación particular y única que enmarca el estilo y la comparación del léxico utilizado.

   Según lo expuesto por [Stamatatos *et al.*, 2000] [Stamatatos *et al.*, 2001] la información obtenida de la sintaxis, es más confiable que aquella que es obtenida directamente del léxico. [Baayen *et al.*, 2002], fueron los primeros en realizar medidas de esta información basándose en su anotación, una frecuencia de reglas estructurada y compleja que tenía un mayor resultado analizándola junto al léxico, que por si sola. Y que fue simplificada por [Stamatatos, 2009b] para ser utilizada con herramientas de NLP para detectar una sentencia completa de un fragmento de esta. Para este análisis se consideran las estructuras propias de un texto y se simplifican en una agrupación confiable a la hora de determinar un textos, para el autor esto sería la consideración de una frase nominal (grupo de palabras que tiene como característica poseer un núcleo, determinante y adyacente), una frase verbal (compuesto por verbos auxiliares o principal y nexos) y una frase preposicional (conjuntos de proposiciones, como por ejemplo, “a causa de” o “junto a”).

   Estructura que se definen para obtener los árboles de dependencia y ser utilizadas en herramientas como el lenguaje de programación Python y su NLTK (*Natural Lenguage Toolkit*) en los cuales revisaremos su configuración y su analizador de dependencia estadística (desde relaciones estructurales como de, dependencia de palabras dentro de una frase). Árboles de dependencia que se pueden analizar mediante su anchura(número máximo de nodos que existe en un nivel del árbol) y profundidad, relaciones individuales, como su frecuencia, brindan una valiosa información sobre cada texto ya que permiten identificar y medir la complejidad de la estructura de las oraciones utilizadas por el autor[Sidorov *et al.*, 2013].

3. **Semántica**

   Las reglas que establecen el significado de las palabras, se encuentran en un estado de desarrollo menos avanzado que los del análisis sintáctico. El significado de ciertas palabras o expresiones utilizadas son especialmente complicadas de tratar mediante algún método estructurado ya que no somos conscientes del tipo de conocimiento que esto implica, su representación en temas de desambiguación de sentidos, tratamiento de los fenómenos relacionados a la coherencia textual, etc. se limita a los aspectos más tratables del problema.

   En este sentido, uno de los mayores aportes al estudio de la semántica es de [Argamon *et al.*, 2007], en la cual define un set de características asociadas a ciertas

palabras o expresiones, determinando el sentido que se le asocia a una palabra en particular en la inmersión total del texto analizado.

- LSA *Latent Semantic Analysis* [Sebastiani, 2005], método automático que utiliza técnicas matemáticas y estadísticas para extraer e inferir relaciones de contextos, utilizando las palabras o partes del texto. No es *NLP* como tal, ya que no utiliza los tradicionales *inputs* de frecuencias de palabras o análisis previamente expuestos, sólo analiza texto sin procesamiento. Define palabras como cadenas de caracteres únicas y separa en pasajes significativos como párrafos o expresiones.

2.3. Modelos sobre Análisis en Textos

Para cada una de las definiciones previamente expuestas, se consideran diferentes tipos de modelos que enmarcan una o varias de estas características. En entre ellas las más utilizadas en el problema de detección de autores, son:

- Frecuencia de Palabras [Sebastiani, 2005]
  Si revisamos un texto desde un fragmento o su completitud, la frecuencia de ciertas palabras se verán totalmente expuestas a la dependencia de la muestra. Se podrá revisar la frecuencia de palabras como segmentos en un texto identificando cada palabra con un mismo valor o realizando un filtro sobre palabras predominantes, por ejemplo para un tipo de vocabulario se podría definir la siguiente definición:
  N: El número de palabras tokens en dicha muestra
  w_i: frecuencia de la palabra i
  $f_{(i,N)}$: frecuencia de w_i en una muestra de N tokens
  $V(N)$: número de tipos en la muestra de N tokens que nos dará el tamaño del vocabulario

- Caracteres
  Captura las marcas de puntuación de un texto (como por ejemplo, las comas, paréntesis, signos de exclamación, de interrogación,etc.). Estas características se calculan como la relación entre el número de apariciones de dichos caracteres en una muestra o su número total dentro del documento. También se observa como característica del autor el uso de mayúsculas o el de números dentro de un texto.

- POS
  Part-of-Speech, utilizado mucho en funciones de sintaxis que buscan la frecuencia de un fragmento de un texto o combinaciones de estas Partes-del-Discurso como una simple aproximación de características sintácticas.

- *Vocabulary Richness* [Holmes, 1994]

  Visualización sobre la diversidad de vocabulario utilizada, dando a entender el conjunto de palabras que domina el autor, y reflejando una parte única de su perfil.

  - *Type-token radio* (R) Se entiende como el tamaño de un vocabulario (tokens únicos) dividido por el número total de tokens en el documento (definiendo tokens como el número de ocurrencias de una palabra). Siendo por definición:
    V: Tamaño del vocabulario dentro del texto seleccionado
    N: Número de tokens

    $$R = V/N$$

  - *Simpson's D Index* (D)
    Es como se mide la diversidad, en autoría se define como la posibilidad de que dos miembros de un par de palabras tokens, escogidos arbitrariamente, pertenezcan al mismo tipo (al mismo autor o misma caracterización, según se analice). Donde se tiene la definición de:
    n: Número de tipos que se producen en una muestra de texto o de tokens.
    N: Total de muestra de texto o de tokens.

    $$D = \frac{\sum n(n-1)}{N(N-1)}$$

  - *Yule's Charactecteristic* (K)
    Es una medida basada en el supuesto de que la ocurrencia de una palabra dada arbitrariamente, se basa en el azar y puede considerarse como una distribución Poisson, en la cual exactamente k-eventos discretos tendrán lugar durante un intervalo de longitud t.

  - *Hapax Legomena (HL) & Dislegomena(D)*
    HL son palabras que ocurren una vez en el texto y *D* son aquellas que ocurren sólo dos veces en el texto. Su utilización se debe a la valoración de palabras dentro de un texto y se puede adoptar en conjunto a los vocabularios.

- N-gramas

  Sub-secuencia de n elementos los cuales pueden determinarse como párrafos, palabras o caracteres, según el método de aproximación, las más utilizadas son de palabras y caracteres ya que tiene como objetivo, determinar la siguiente

palabra/carácter a utilizar o analizar como es la secuencia con la cual un autor utiliza ciertas palabra/caracteres. La frecuencia de varios n-gramas se ha utilizado con buenos resultados para la captura de las preferencias léxicas de un autor, para determinar autorías de opiniones, identificando también similitudes de clasificación de textos o hasta determinando el lenguaje nativo del autor, pero no ha tenido buenos resultados al determinar las características de sintaxis de un autor.

Una de las definiciones más importantes es la decisión del n a trabajar. Un n muy grande, captura de mejor forma el léxico y el contexto de un documento, pero también incrementa considerablemente la dimensión de representación (aumenta las características a considerar), mientras que un n pequeño (2 o 3) podrá ser más capaz de representar información sobre sub-palabras (análisis sobre sílabas). La mejor selección de n se realiza según el tipo de lenguaje natural a utilizar, siendo un n-grande más apropiado para lenguajes como el Alemán o Griego, y un n-pequeño para el Inglés [Kešelj *et al.*, 2003].

2.4. Técnicas de Clasificación

1. **Multivariable**
   Para el estudio de detección de autores se revisan las técnicas de *Machine Learning*, *Discriminant Analysis*, *Cluster Analysis* y los *Principal Components Analysis*, estas técnicas son combinadas frecuentemente con los modelos de frecuencia para lograr una mayor efectividad.

   a) **Machine Learning** (ML)

   El objetivo de ML es la construcción de programas computacionales que aprendan y mejoren con la experiencia. Las técnicas más implementadas para la categorización de texto y detección de autores son *Naives Bayes*, *K-nearest Neighbour* y *Support Vector Machines*.

   - *Naive Bayes* (NB)
     Basado en NB, utiliza la probabilidad conjunta de palabras o categorías para estimar las probabilidades que tiene un documento para ser categorizado dentro de un conjunto específico (dimensiones como el tema, sentimiento y el lenguaje utilizado). Para mayor información revisar [Peng y Schuurmans, 2003, Peng *et al.*, 2004].
   - *K-Nearest Neighbour*
     Almacenan todos los ejemplos de entrenamiento, determinando la similitud del texto con aquellos que ha memorizado, entregando una puntuación

con la cual se puede categorizar dicho documento. Su fuerte es la identificación de lenguaje y la detección de autores. Lo que se puede entender mejor revisando [Diederich *et al.*, 2003].

- *Support Vector Machines*
  Intenta realizar una distinción entre los textos de ejemplo separándolos en los positivos como en los negativos, a través de *support vectors* [Diederich *et al.*, 2003]. Por ejemplo en el análisis de los *"Federalist Paper"*, mencionados anteriormente, obtendríamos una separación entre aquellos que son atribuidos a un autor determinado. Una de las fortalezas de esta técnica es que puede procesar documentos de largo significativo y bases de datos con un gran número de textos, es por esto que suele utilizarse para la detección de autores y la categorización de textos.

b) *Discriminant Analysis*
Esta técnica intenta maximizar la diferencia entre grupos y minimizar aquella diferencia que se encuentra dentro del grupo, así logra predecir la membresía a cierto grupo con un set básico de predictores.

c) *Cluster Analysis*
Técnica que intenta organizar la información mediante variables seleccionadas, tal que, la data pueda ser enmarcada dentro de un conjunto. Muy parecido a DA al intentar homogeneizar dentro de su conjunto y distinguir claramente fuera de estos. En los *cluster* una de las decisiones más importantes es la del cálculo sobre la cantidad de conjuntos a realizar, esto determinará significativamente el análisis realizado. Esta técnica se utiliza principalmente sobre el análisis de la frecuencia de palabras más utilizadas por un autor en específico y para distinguir textos de diferentes autores ([Hoover, 2001] habla de una exactitud de este método por sobre el 90 %).

d) *Principal Component Analysis*
El objetivo principal del PCA es entender las relaciones y correlaciones de cierto grupo de datos. Todas las variables son comparadas, o sea que los vectores son generados por cada par de variable, [Sering *et al.*, 2018] utiliza PCA para la comparación de palabras bases y sobre la sintaxis para la detección de autores, mientras que [Stamatatos, 2009b] la utiliza sobre el test corpus.

CAPÍTULO 3
PROPUESTA DE SOLUCIÓN

Este trabajo se basa en los algoritmos planteados en [Reyes, 2016] para la detección de plagio intrínseco, en este se reconoce que, aunque la mayoría de los algoritmos de detección se basan en la captura del estilo del autor existen ciertas distinciones propuestas para analizar los segmentos o palabras con respecto al documento, esto será interesante ya que los métodos propuestos podrán ser mejorados revisando algunos de los elementos mencionados anteriormente con los *Modelos sobre análisis en textos* (sección 2.3).

El autor nos presenta un análisis divido en tres grandes fases, las cuales dictaminaran la continuación de este trabajo:

1. La primera fase consiste en la segmentación del texto mediante algún criterio definido para cada uno de los algoritmos. Aquí podemos encontrar autores que realizan una segmentación por capítulos, secciones, párrafos o cantidad de palabras. En esta etapa también es donde se realiza en algunos casos ciertos preprocesamientos al texto, por ejemplo, eliminando *stopwords* o quitando símbolos y caracteres que no pertenezcan al conjunto [a-z].

2. La segunda fase, consiste en aplicar el algoritmo propiamente tal a los segmentos, buscando determinar el estilo del autor o la complejidad de éstos. Aquí podemos encontrar autores como [Stamatatos, 2009a] que propone en considerar un documento como una bolsa de n-gramas de palabras, creando, para un n definido, un vector de frecuencia de n-gramas normalizado al que llama *Perfil del Texto*, repitiendo la operación para cada uno de los segmentos. Por otro lado, [Funez y Errecalde, 2011] propone segmentar el texto para luego a cada uno de los segmentos calcular uno a uno los índices estilométricos, generando una cantidad de vectores igual al número de segmentos, donde cada vector es de dimensión d, dependiendo de la cantidad de índices considerados.

3. La tercera y última fase consiste en definir un criterio que permita clasificar cada uno de los segmentos definiendo cuales de éstos se consideran escritos por el autor del texto. Aquí la tendencia es comparar los valores obtenidos en la fase anterior, poniendo especial atención en los los *outlier* que se puedan presentar.

Los algoritmos aplicados por este autor son tres, **DOCODE** [Gallardo, 2013] con el cual comienza, **DOCODE NORMALIZADO** y **DOCODE NORMALIZADO POR SEGMENTO**, los cuales son mejoras del algoritmo principal. A continuación explicaremos brevemente cada uno de ellos y las diferencias utilizadas para actuar sobre el problema elegido.

3.1. Análisis de Algoritmos

Considerando sólo la segunda fase, el algoritmo de *DOCODE* se presenta como el cálculo de segmentos d_c que se obtienen de la sumatoria de la diferencia de magnitud de las palabras w contenidas en un segmento, con V como el vectores de frecuencia de palabras y v_c como el vector de frecuencia del segmento c.

Por cada w del vector v_c se calculará la diferencia con respecto a la misma palabra en el documento completo obteniendo de esta forma una caracterización de la homogeneidad de la escritura en dicho segmento.

Con todos los d_c se determina el estilo de escritura de documento y se resuelve un criterio (umbral) a partir del cual se podrán analizar qué segmentos son escritos por el autor y cuales no.

En los casos extremos se tendrá que, las palabras más utilizadas en el documento generará un cálculo de la diferencia de magnitudes de las componentes con tendencia a 1 y aquellas que menos se utilicen tenderá a 0 (siendo cero si w aparece sólo una vez en el documento o se encuentra concentrada en el segmento). Para mayor información, revisar [Gallardo, 2013].

Como en el algoritmo original se plantea la opción de distinguir los segmentos que contenían plagio, se realizará una modificación para determinar si el segmento es escrito por el autor o no (la modificación se podrá apreciar en color rojo). Se revisará en primera instancia el o los documentos conocido(s) como propio(s) del autor para sacar el estilo de su autoría, con esto se revisará el documento desconocido por segmento y se comparará con el umbral, adicionando la variable $Cantidad\_Aciertos$ que nos proporcionará información de la cantidad de segmentos analizados que fueron considerados como escritos por el autor, finalmente se comparará este porcentaje de pertenencia con una nueva variable de $Porcentaje\_Discriminatorio$, el cual nos ayudará a determinar cuando el documento y por ende la carpeta en revisión, es un posible escrito por el autor analizado. Su output será marcar la carpeta como posible escrito por el autor como un *Verdadero* si es que se cumple la condición y un *Falso* si no se cumple. El algoritmo descrito se presenta en pseudocódigo a continuación:

Algorithm 1 DOCODE

Require: $C, V, m, \delta, C_d, V_d$

for $c \in C$ **do**
  $d_c = 0$
  construir v_c usando los términos del segmento c
  for $w \in v_c$ **do**
    $d_c = d_c + \frac{|FREQ(w,V) - FREQ(w,v_c)|}{|FREQ(w,V) + FREQ(w,v_c)|}$
  end for
end for
$Estilo\_Documento\_Conocido = \frac{1}{|C|} \sum_{c \in C} d_c$
for $c_d \in C_d$ **do**
  $d_{cd} = 0$
  construir v_{dc} usando los términos del segmento c_d
  for $w_d \in v_{dc}$ **do**
    $d_{cd} = d_{cd} + \frac{|FREQ(w_d,V_d) - FREQ(w_d,v_{dc})|}{|FREQ(w_d,V_d) + FREQ(w_d,v_{dc})|}$
  end for
end for
for $c_d \in C_d$ **do**
  if $d_{cd} \geq Estilo\_Documento\_Conocido - \delta$ **then**
    $Cantidad\_Aciertos = Cantidad\_Aciertos + 1$
  end if
  $Cantidad\_Segmentos = Cantidad\_Segmentos + 1$
end for
$Porcentaje\_Pertenencia = \frac{Cantidad\_Aciertos}{Cantidad\_Segmentos}$
if $Porcentaje\_Pertenencia \geq Porcentaje\_Discriminatorio$ **then**
  Marcar carpeta como VERDADERO (posible escrito del autor)
else
  Marcar carpeta como FALSO
end if

Para el caso del algoritmo de *DOCODE* su autor plantea que su δ óptimo es $0{,}075$. Este δ se propone por [Reyes, 2016] para los siguientes algoritmos de DOCODE NORMALIZADO y DOCODE NORMALIZADO POR SEGMENTO como un $\lambda \cdot \sigma$, siendo σ la desviación estándar de los d_c calculados y nuestro umbral definido por:

$$Umbral = Estilo \pm \lambda \cdot \sigma$$

En el caso del algoritmo *DOCODE NORMALIZADO*, comparado con DOCODE, se plantea una mejora al problema de segmentos con palabras poco utilizadas o muy utilizadas que pueden generar ruido en las soluciones, su foco deja de estar en la cantidad de apariciones de una palabra en un segmento determinado y considera la distribución de esta misma en los segmentos. Su hipótesis, según la describe [Reyes, 2016] es:

> *"Si alguna de las palabras usadas en un documento son específicas de un autor, entonces se puede pensar que estas palabras se concentran en los párrafos o segmentos que dicho autor escribió."*

En este caso, las palabras más utilizadas tenderán a 0 y las menos utilizadas a 1. Por lo que el algoritmo con la modificación de cantidad de aciertos, se define como (la modificación del algoritmo principal al algoritmo mejorado se podrán apreciar en color azul, mientras que nuevamente la modificación para este problema se podrá apreciar en color rojo):

Algorithm 2 DOCODE Normalizado

Require: $C, V, m, \delta, C_d, V_d$

Normalizar V
for $c \in C$ **do**
  $d_c = 0$
  construir v_c **normalizado** usando los términos del segmento c
  for $w \in v_c$ **do**
    $d_c = d_c + \frac{|freq(w,V) - freq(w,v_c)|}{|freq(w,V) + fre(w,v_c)|}$
  end for
end for
$Estilo\_Documento\_Conocido = \frac{1}{|C|} \sum_{c \in C} d_c$
for $c_d \in C_d$ **do**
  $d_{dc} = 0$
  construir v_{cd} **normalizado** usando los términos del segmento c_d
  for $w_d \in v_{cd}$ **do**
    $d_{cd} = d_{cd} + \frac{|freq(w_d,V_d) - freq(w_d,v_{cd})|}{|freq(w_d,V_d) + fre(w_d,v_{cd})|}$
  end for
end for
for $c_d \in C_d$ **do**
  if $d_{cd} <= Estilo\_Documento\_Conocido + \lambda \cdot \sigma$ **then**
    $Cantidad\_Aciertos = Cantidad\_Aciertos + 1$
  end if
  $Cantidad\_Segmentos = Cantidad\_Segmentos + 1$
end for
$Porcentaje\_Pertenencia = \frac{Cantidad\_Aciertos}{Cantidad\_Segmentos}$
if $Porcentaje\_Pertenencia \geq Porcentaje\_Discriminatorio$ **then**
  Marcar carpeta como VERDADERO (posible escrito del autor)
else
  Marcar carpeta como FALSO
end if

Para el caso del algoritmo de *DOCODE NORMALIZADO* su autor utiliza el δ descompuesto por $\lambda \cdot \sigma$ y se plantea que su λ óptimo para una mejor métrica F_1 (sección 3.2.1) es de $0{,}2$.

Por último se define el algoritmo de *DOCODE NORMALIZADO POR SEGMENTO* como la mejora en función de analizar la distribución de las palabras a lo largo de los segmentos y si esta guarda relación con la distribuciones obtenidas del documento completo. Dicha mejora se encargará de eliminar el ruido provocado por las palabras que no están contenidas en los vectores v_c a la hora de comparar, palabras que aparecen en uno o pocos segmentos, y realizar los cálculos de cada segmento respecto al vector de frecuencias V del documento o documentos completos. Su algoritmo con la modificación de cantidad de aciertos se define como:

Algorithm 3 DOCODE Normalizado por Segmento

Require: $C, V, m, \delta, C_d, V_d$

for $c \in C$ **do**
  $d_c = 0$
  construir v_c **normalizado** usando los términos del segmento c
  construir $V|_{v_c}$ **tomando desde** V **solo las palabras contenidas en** v_c
  Normalizar V_{auxc}
  for $w \in v_c$ **do**
    $d_c = d_c + \frac{|freq(w,V|_{v_c}) - freq(w,v_c)|}{|freq(w,V|_{v_c}) + fre(w,v_c)|}$
  end for
end for
$Estilo\_Documento\_Conocido = \frac{1}{|C|} \sum_{c \in C} d_c$
for $c_d \in C_d$ **do**
  $d_{cd} = 0$
  construir v_{cd} **normalizado** usando los términos del segmento c_d
  construir $V|_{v_{cd}}$ **tomando desde** V_d **solo las palabras contenidas en** v_{cd}
  Normalizar V_{dauxc}
  for $w_d \in v_{cd}$ **do**
    $d_{cd} = d_{cd} + \frac{\left|freq(w_d,V_d|_{v_{cd}}) - freq(w_d,v_{cd})\right|}{\left|freq(w_d,V_d|_{v_{cd}}) + fre(w_d,v_{cd})\right|}$
  end for
end for
for $c_d \in C_d$ **do**
  if $d_{cd} < Estilo\_Documento\_Conocido + \lambda \cdot \sigma$ **then**
    $Cantidad\_Aciertos = Cantidad\_Aciertos + 1$
  end if
  $Cantidad\_Segmentos = Cantidad\_Segmentos + 1$
end for
$Porcentaje\_Pertenencia = \frac{Cantidad\_Aciertos}{Cantidad\_Segmentos}$
if $Porcentaje\_Pertenencia \geq Porcentaje\_Discriminatorio$ **then**
  Marcar carpeta como VERDADERO (posible escrito del autor)
else
  Marcar carpeta como FALSO
end if

Para el caso del algoritmo de *DOCODE NORMALIZADO POR SEGMENTO* su autor plantea que su λ óptimo para una mejor métrica de F_1 (sección 3.2.1) es de $0{,}8$.

3.2. Criterio de Evaluación

A continuación se describe una serie de métricas que serán utilizadas para medir la efectividad de los algoritmos y las mejoras propuestas en la sección anterior.

3.2.1. Matriz de Confusión

Una matriz de confusión es una herramienta que permite la visualización del desempeño de un algoritmo. Cada columna de la matriz representa el número de predicciones de cada clase, mientras que cada fila representa a las instancias en la clase real.

	Predicción Negativa	Predicción Positiva
Real Negativo	TN	FP
Real Positivo	FN	TP

Tabla 1: Matriz de confusión

A partir de la matriz de confusión existe una serie de métricas que permiten analizar el desempeño del algoritmo, algunas de estas utilizadas en los experimentos son:

- Precision, indica el grado o frecuencia en el que, el algoritmo predice positivo correctamente.
$$precision = \frac{TP}{TP + FP}$$
- Recall, establece el porcentaje de textos que se clasifican correctamente
$$recall = \frac{TP}{TP + FN}$$
- F_1, media armónica entre *precision* y *recall*
$$F_1 = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$

3.3. Entorno de Desarrollo y Scripts

En el transcurso de este trabajo se desarrollaron una serie de *scripts* que permitieron llevar a cabo los experimentos descritos más adelante en la sección 4.2. A continuación se presentan detalles del entorno de desarrollo, librerías y detalles de los *scripts* desarrollados.

Entorno de desarrollo

El entorno de desarrollo y herramientas utilizadas durante el desarrollo de esta memoria se describe a continuación:

Características de Hardware		
Equipo	Procesador	Intel core i5
	Memoria RAM	4 Gb
Características de Software		
Sistema Operativo	Windows 10	
Lenguajes de Programación	Python 2.7	
Librerías Importantes	Plotly 1.37.1	
	xlwt 1.3.0	

Scripts

Para llevar a cabo los experimentos se desarrollo una serie de scripts en Python, agrupados en el directorio DOCODEX4 disponible en los recursos anexos de este trabajo y en el repositorio github `https://github.com/loretoiii/DOCODEx4`. A continuación se presenta la estructura del directorio y una breve explicación de cada uno de los scripts desarrollados.

- **Source/**
  Directorio principal en la cual se dispondrá de todas las carpetas a analizar y el archivo de texto **truth.txt** si es que se desea analizar las respuestas de dicho análisis.
- **Result/**
  Directorio principal que se utilizará para dejar la carpeta de resultados obtenida de una ejecución. Como cada carpeta se identifica con el n utilizado en la prueba, el día con la hora, minutos y segundos de creación, no existirá problema de sobre-escritura al ejecutar los mismos experimentos.
- **Fases/**
  Directorio que contiene las fuentes de los experimentos descritos en la sección 3.5 y

sus correspondientes resultados.

- **VerificacionAutores.py/**
  Script principal que toma las carpetas de **Source** y los analiza. De este archivo ejecutaremos todos los demás script. Sus principales características son que, define las direcciones de **Source** y **Result**, toma las variables más importantes ($m, n,$,λ's, lista de porcentajes y segmento limitante si se desea utilizar uno en específico) definidas al comienzo del archivo y con la cual se realiza el trabajo entre los textos conocidos y desconocidos.

- **text2.py/**
  Clase que realiza el preprocesamiento al texto a analizar y contiene los métodos *docode*, *docode_normalizado* y *docode_normalizado_segmento*.

- **graficame4.py/**
  Script que genera un gráfico de distribución con cada punto obtenido de cada segmento analizado, tanto de los documentos conocidos como de los desconocidos, y por cada algoritmo analizado: *docode*, *docode_normalizado* y *docode_normalizado_segmento*.

- **umbrales1.py/**
  Script que genera 3 archivos excels, uno por cada algoritmo analizado, en los cuales se revisa cada carpeta con cada λ para obtener el porcentaje de pertenencia de cada una de ellas.

- **GraficarResultados.py/**
  Script que utiliza los archivos generados por **umbrales1** para generar la gráfica de cada porcentaje de pertenencia obtenido e identificarlo si se trata de una respuesta positiva (el autor del texto desconocido era quien había escrito los textos conocidos) o no.

- **MatrizConfusion.py/**
  Script que genera un archivo en excel que dispone de todas las combinaciones de algoritmos, λ's y porcentaje de pertenencia utilizado, señalando el F_1, *Precision* y *Recall* obtenido por cada una de estas combinaciones. También genera los histogramas de F_1 y de aciertos, el primero con cada F_1 obtenido con el objetivo de identificar claramente las mejores métricas de F_1 por algoritmo y el segundo, con cada acierto obtenido (*True Positive* y *True Negative*) con el objetivo de identificar claramente la configuración que obtuvo una mayor cantidad de aciertos considerando todas las carpetas analizadas.

- **Histograma.py/**
  Script que genera gráficos de histograma, por cada algoritmo analizado, con la cantidad de carpetas clasificadas como verdaderas y cantidad de carpetas clasificadas como falsas para cada combinación del archivo generado por **MatrizCondusion**.

- **Histograma2.py/**
  Script que genera por algoritmo un gráfico que representa por porcentaje de pertenencia y cada λ's la cantidad de aciertos y rechazos obtenidos.

- **B0B1.py/**
  Script que con carpetas conocidas muestra la mejor configuración de λ y segmento limitante que genera un F_1 óptimo.Esto lo gráfica mostrando un histograma donde por cada λ y segmento limitante nos dirán si el segmento debe considerarse como verdadero(0) o falso(1), considerando verdadero como escrito por el autor. Para carpetas desconocidas se revisará la configuración realizada con el segmento limitante ya fijo y se podrá revisar los valores obtenidos para F_1 y el total de aciertos.

3.4. Descripción del Data Set

Se crea un data set con los textos obtenidos de [GUT, 2017], plataforma *online* en la que se ofrecen más de 56.000 *eBooks* gratis (sin *copyright* o con *copyright* vencido). Se seleccionan 25 autores, la mayoría conocidos pero principalmente se opta por seleccionar autores que posean por lo menos 4 libros en el data set sin ningún otro criterio en particular (resumen de autores en Tabla 2).

De la selección se puede revisar que son en su mayoría autores con escritos originales en idioma inglés, pero también se utilizaron obras traducidas a este idioma. Existen algunos en tanto que por ser obras muy antiguas poseen una revisión por un traductor externo, pero finalmente todos los libros a analizar dispondrán del mismo idioma.

Tabla 2: Lista Autores Data Set

Autores	Nacionalidad
Mark Twain	EEUU
Jane Austen	Reino Unido
Charles Dickens	Reino Unido
Leon Toltoi	Rusia
Edgar Alan Poe	EEUU
Virginia Woolf	Reino Unido
Homero	Jonia
Arthur Conan Doyle	Reino Unido
Platon	Atenas
F.Nietzsche	Alemania
Shakespear	Reino Unido
Fyodor Dostoyevsky	Rusia
Charlotte Brontë	Reino Unido
Lewis Carroll	Reino Unido
Alexandre Dumas	Francia
Gustave Flaubert	Francia
Oscar Wilde	Irlanda
Eleanor Hallowell Abbott	EEUU
L. Frank (Lyman Frank) Baum	EEUU
Aldous Huxley	Reino Unido
H.L.Sayler	EEUU
Felix Dahn	Alemania
Dante Alighieri	Italia
Richard Harding Davis	EEUU
Philip K.Dick	EEUU

El origen de los autores es principalmente Reino Unido, con un total de 8 autores, y Estados Unidos con un total de 7 autores. Los libros poseen un promedio de 115.797 palabras y 626.866 caracteres. Y sus años de publicación varían entre 1308 [1], con La Divina Comedia de Dante Alieghieri, y 1954, con *Beyond the Door* de Philip K.Dick, por lo cual se tienen una diferencia de 646 años. También hay que considerar que existen dos autores que no se consideran en esta diferencia por situarse por lejos como predecesores en esta lista, ellos son Homero (griego antiguo del siglo VIII a.C) y Platon (griego del siglo 427 a.C.). Su selección se basa en la idea de que, tanto sus temas como su estilo de escritura debiesen ser lo suficientemente diferentes de un autor contemporáneo y esto podría servir para definir algunas caracterizaciones sobre este data set.

[1] Aproximación de cuando se cree que se terminó de escribir esta obra ya que no se posee una certeza real sobre la fecha de su publicación real

El data set se divide en dos niveles, donde el elemento principal es una carpeta que contiene uno o más archivos de textos con el titulo 'knownXX', con XX referenciando al número del archivo y el cual tiene un texto conocido por el autor a analizar de la carpeta y un archivo con el titulo de 'unknown', el cual contiene un texto que se analizará para decidir si pertenece o no al autor.

3.4.1. Nivel 0

Este nivel posee dos carpetas, una de **Training** que servirá para realizar un análisis sobre los mejores resultados que podemos obtener realizando modificaciones a las variables a utilizar, y una de carpeta de **Test**, que nos ayudará a probar las hipótesis.

Las carpetas se han creado, desde un conjunto original sin ninguna prioridad ni orden. El conjunto original contenía 200 carpetas con 2 archivos cada una, uno conocido y otro desconocido. 100 carpetas con resultados positivos y 100 carpetas con resultados negativos, esto significa que en los resultados positivos el texto desconocido corresponde al autor del texto conocido y que en los negativos, no lo hace.

Cada autor generará entonces 8 carpetas, 4 de ellas serán carpetas positivas y 4 negativas, el orden de esta generación se podrá revisar en la lista de anexos. Las carpetas positivas se generaron creando el archivo desconocido con un fragmento sacado del texto conocido del autor de dicha carpeta, este fragmento genera este nuevo archivo y es eliminado desde el archivo original. El largo del fragmento es seleccionado de forma aleatoria y corresponde a un porcentaje de entre el 1,1 % del texto original al 83,1 %.

Por otra parte, en el caso de las carpetas con resultados falsos (que el texto desconocido no es del mismo autor que los textos conocidos), carpetas que llamaremos negativas, se utiliza el fragmento de otros autores: 50 carpetas se han creado variando el autor desconocido y la otra mitad con un mismo autor: Fyodor Dostoyevsky, esta división se emplea para determinar si existe alguna característica que nos ayude a divisar una diferencia al comparar estilos, mientras que los primeros textos son comparados con diferentes autores que pueden tener o no ciertas características parecidas al autor original, en la segunda parte podemos analizar con una constante particular, el estilo del autor ruso. Cabe destacar que la elección del autor es arbitraria y se podría variar a cualquier otro estilo de la lista de autores, sólo se implementa por poseer textos extensos y un vocabulario especial que se cree pueda ser aprovechado para realizar comparaciones efectivas.

Para la carpeta de **Training** se tiene que el el 35 % de sus autores no escribieron sus obras en inglés, el año promedio de esta carpeta es de 1872 (se hace la misma excepción con los autores predecesores mencionados anteriormente y sus años son desde 1308 a 1954), el promedio de palabras de las carpetas positivas considerando sólo los textos conocidos es de 107.515 y el promedio de caracteres de estas mismas es de 605.471. Para los textos desconocidos se tiene un promedio de 18.924 palabras y 106.924 caracteres, por lo que

se tendrá que los textos desconocidos, en promedio serán de un fragmento del 23 % del libro original. Para las carpetas negativas se tendrá que la mayor diferencia entre años de publicación de los textos a comparar será de 70 años, que el promedio de palabras será de 122.591 en los textos conocidos y de un promedio de 20.424 para los textos desconocidos, teniendo un promedio de un 46 % de texto desconocido por texto conocido. En el caso de la cantidad de caracteres su promedio en los textos conocidos será de 691.053 y el promedio de los textos desconocidos será de 111.754.

En la carpeta de **Test** se tendrá también que el 35 % de sus autores no escribieron sus textos en inglés, pero su año promedio entre las carpetas es de 1818 (desde 1308 a 1954). Teniendo también menor porcentaje en la cantidad de fragmento desconocido por texto conocido, este será de un 20 % para las carpetas positivas, teniendo entre éstas un promedio de palabras de los textos conocidos de 86.560 y un promedio de caracteres de 446.180. Mientras que sus textos desconocidos tendrán un promedio de palabras de 18.594 y 95.156 de promedio de caracteres. Entre las carpetas negativas la diferencia aumenta abruptamente a un 84 %, esto se debe porque los textos desconocidos, aunque su promedio de palabras de textos conocidos es de 110.233 contra un promedio de 24.915 en textos desconocidos, hay una diferencia muy grande en algunas carpetas que poseen textos conocidos medianamente cortos y son comparados con textos desconocidos de un largo significativo.

Tabla 3: Resumen de Palabras/Caracteres por carpetas - Nivel 0

Carpeta de:		Training		Test	
Calificadas como		Verdaderas	Falsas	Verdaderas	Falsas
Texto Conocido	Cantidad Prom. Palabras	107.515	122.590,8	86.560,12	110.233
	Cantidad Prom. Caracteres	605.471	691.053	446.180	561.969
Texto Desconocido	Cantidad Prom. Palabras	18.924	20.424	18.594	24.915
	Cantidad Prom. Caracteres	106.924	111.754	95.156	137.691
	Porcentaje promedio de la cantidad de palabras en texto desconocidas por palabras en texto conocido	23 %	46 %	20 %	84 %

3.4.2. Nivel 1

Este nivel posee 100 carpetas, 4 carpetas de cada autor de nuestra lista. Cada una de estas cuatro carpetas tiene 2 archivos conocidos del autor y uno desconocido, dividiéndose de esta forma: las primeras dos son carpetas con resultados positivos, o sea que su texto desconocido es también del autor y dos carpetas con resultados negativos, perteneciendo el texto desconocido a otro autor.

Para la primera y segunda carpeta el texto elegido serán obras del mismo autor que, cuando sea posible, su fecha de publicación será parecida a los textos-libros conocidos del autor. Para la tercera carpeta, se utilizará un autor, nuevamente en lo posible, de idioma y/o fecha

de publicación parecido al autor, para finalizar con una cuarta carpeta que sea un texto desconocido, en lo posible, suficientemente diferente al autor. Como esta diferencia no puede ser medida o caracterizada se utilizan libros aleatorios y hasta se podrá encontrar en algunos ejemplos específicos, algunos textos que no compartan el mismo idioma siendo textos en español. La idea de esto es poder realizar comparaciones evidentes que podamos observar y analizar.

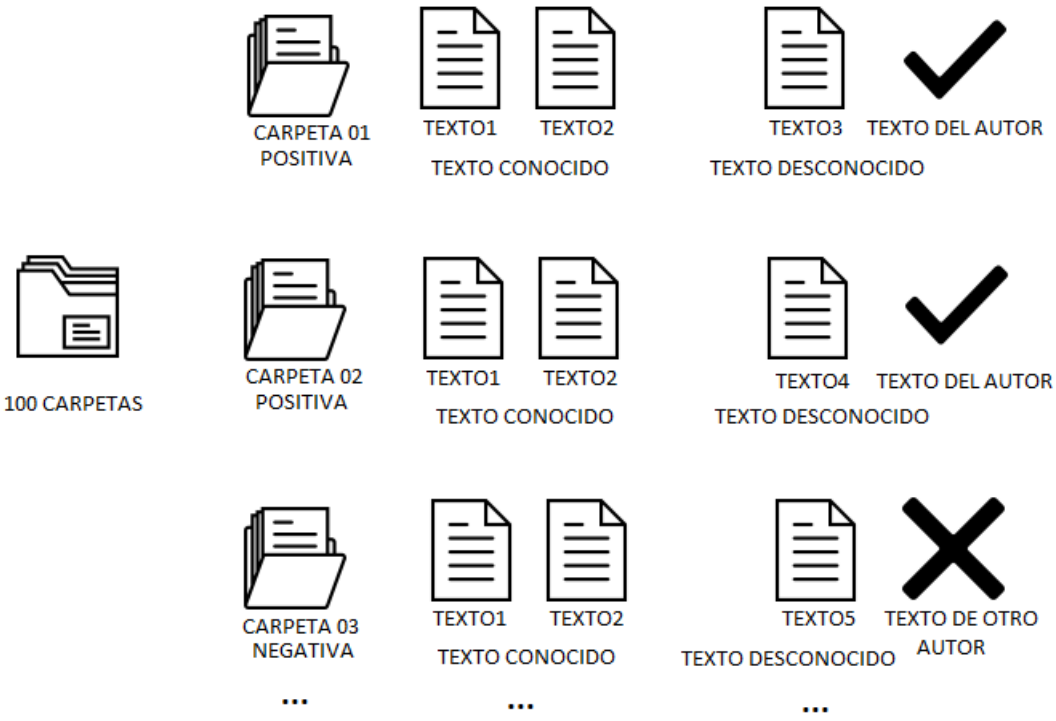


Figura 1: Imagen: División Carpetas

Para estas carpetas se tiene un porcentaje de palabras de los textos-libros desconocidos, por cantidad de palabras de los textos-libros conocidos de un 41 % , teniendo como mínimo un porcentaje de un 2 %. Si sólo analizamos los textos pertenecientes al autor (primeras dos carpetas de cada autor) el porcentaje de cantidad de palabras en los textos conocidos por cantidad de palabras en textos desconocidos crece a un 72 % teniendo como porcentaje mínimo también un 2 %. Con un promedio de 307.017 palabras en los textos conocidos y un promedio de 1.581.557 de caracteres.

Tabla 4: Resumen de Palabras/Caracteres por carpetas - Nivel 1

Carpeta de:		Nivel1	
Calificadas como		Verdaderas	Falsas
Texto Conocido	Cantidad Prom. Palabras	158.556	158.556
	Cantidad Prom. Caracteres	794.487	794.487
Texto Desconocido	Cantidad Prom. Palabras	72.737	190.838
	Cantidad Prom. Caracteres	372.968	1.010.878
	Porcentaje promedio de la cantidad de palabras en texto desconocidas por palabras en texto conocido	59 %	19 %

La cantidad de palabras, caracteres y años de publicación se podrán revisar en mayor detalle en la tabla de anexo sección 5.3.

3.5. Experimentos Numéricos

Siguiendo las fases descritas en el comienzo de esta sección, se procede a determinar en cada una de ellas los criterios de evaluación que aplicarán para este problema en particular.

Se comenzará revisando experimentos que nos validen la correcta obtención de resultados, comparándonos con el algoritmo inicial. A continuación, se procederá a definir el criterio que determinará las respuesta de nuestra función. Finalmente se analizarán las constantes en escena y las mejoras propuestas.

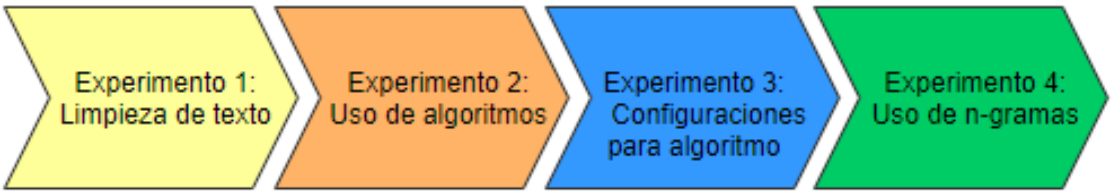


Figura 2: Imagen: Etapas para explicación de experimentos

3.5.1. Experimento 0

En esta fase se revisará el manejo de texto, el cual utiliza la misma base que *DOCODE* considerando las palabras como unidad fundamental y realizando un preprocesamiento del documento quitando todo carácter que no pertenezca al conjunto [a-z], y se comparará con los resultados obtenidos por [Reyes, 2016], también en su fase inicial. Se explicará paso a paso la forma más rápida encontrada de como comparar los resultados.

Primero, se deberá seleccionar un texto cualquiera, para nuestro ejemplo se toma el documento propuesto en la memoria de Detección de Plagio (como llamaremos a la memoria de [Reyes, 2016]), libro de " *Little Woman* " descargado de [GUT, 2017]. Para utilizarlo en la ejecución del algoritmo Verificación de Autores (como llamaremos al cambio de algoritmos propuesto en esta memoria), se deberá crear una carpeta y agregar en ella el texto generando. Se deberán dejar los archivos en la carpeta *Source*, tal como se ha dejado en la carpeta de prueba Fase 0 descrita en la sección de script 3.3.

Como el archivo base de la memoria anterior utiliza los λ's para un mejor Puntaje Final y Verificación de Autores sólo posee Mejor F_1 [2], se deberá realizar la modificación de los λ's en el archivo *app.py*, obteniendo los siguientes resultados que se encuentran comparados con los valores obtenidos en nuestro archivo generado *Trace.txt* :

Tabla 5: Comparación de resultados en Experimento 0 - Manejo de texto

		Detección Plagio	Verificación Autores
Palabras Analizadas		194.256	194.256
DOCODE	Estilo de Escritura	0.495	0.495
	Umbral	0.42	0.420
DOCODE Normalizado	Estilo de Escritura	0.578	0.578
	Umbral	0.585	0.585
DOCODE Norm por Segmento	Estilo de Escritura	0.329	0.329
	Umbral	0.363	0.363

Cabe destacar que se podrán comparar también los segmentos generados. En detección de plagio se podrán revisar en la pestaña "Segmentos" y en verificación de autores en el archivo *Trace2.txt* (en la carpeta de resultados) o de manera visual en el gráfico generado en la carpeta "GraficosPorCarpeta" [3], ejemplo de esto se visualiza a continuación en la siguiente Figura 3. En ella se podremos tener información de la carpeta revisada con cada uno de los algoritmos, DOCODE, DOCODE NORMALIZADO y DOCODE NORMALIZADO POR SEGMENTO, respectivamente, por cada algoritmo veremos tres análisis, los dos primeros representan

[2] Esto se debe a que la memoria anterior utiliza el principio de la Granularidad para obtener Puntaje Final, este principio es utilizado para darle un mayor énfasis a un plagio encontrado dentro del documento y no es equiparable con alguna función del algoritmo presentado en este documento.

[3] La carpeta de Resultados generados para esta prueba se encuentra en la carpeta *Fase0/Resultados*

a los documentos definidos como conocidos y el último será del documento desconocido. Revisando, podremos ver un boxplot, el cual nos permite ver la dispersión de los puntos (segmentos) con la mediana y observar rápidamente los *outliers*, y un gráfico de puntos, el cual representa a cada segmento construido con el algoritmo, por último se tienen los puntos del documento desconocido y una linea en paralelo que nos muestra el umbral obtenido[4]. Para DOCODE todos los puntos que se encuentren arriba del umbral son posibles segmentos escritos por el autor, para DOCODE NORMALIZADO y DOCODE NORMALIZADO POR SEGMENTO serán los puntos bajo el umbral los que relacionarán a los segmentos escritos por el autor.

[4]En la imagen nos posicionamos en el umbral del algoritmo DOCODE revisando que nos ha dado el valor de $0{,}42047$, si nos posicionáramos en cada punto obtenido del documento desconocido (en el ejemplo, los puntos en color verde) podríamos revisar uno a uno los valores obtenidos.

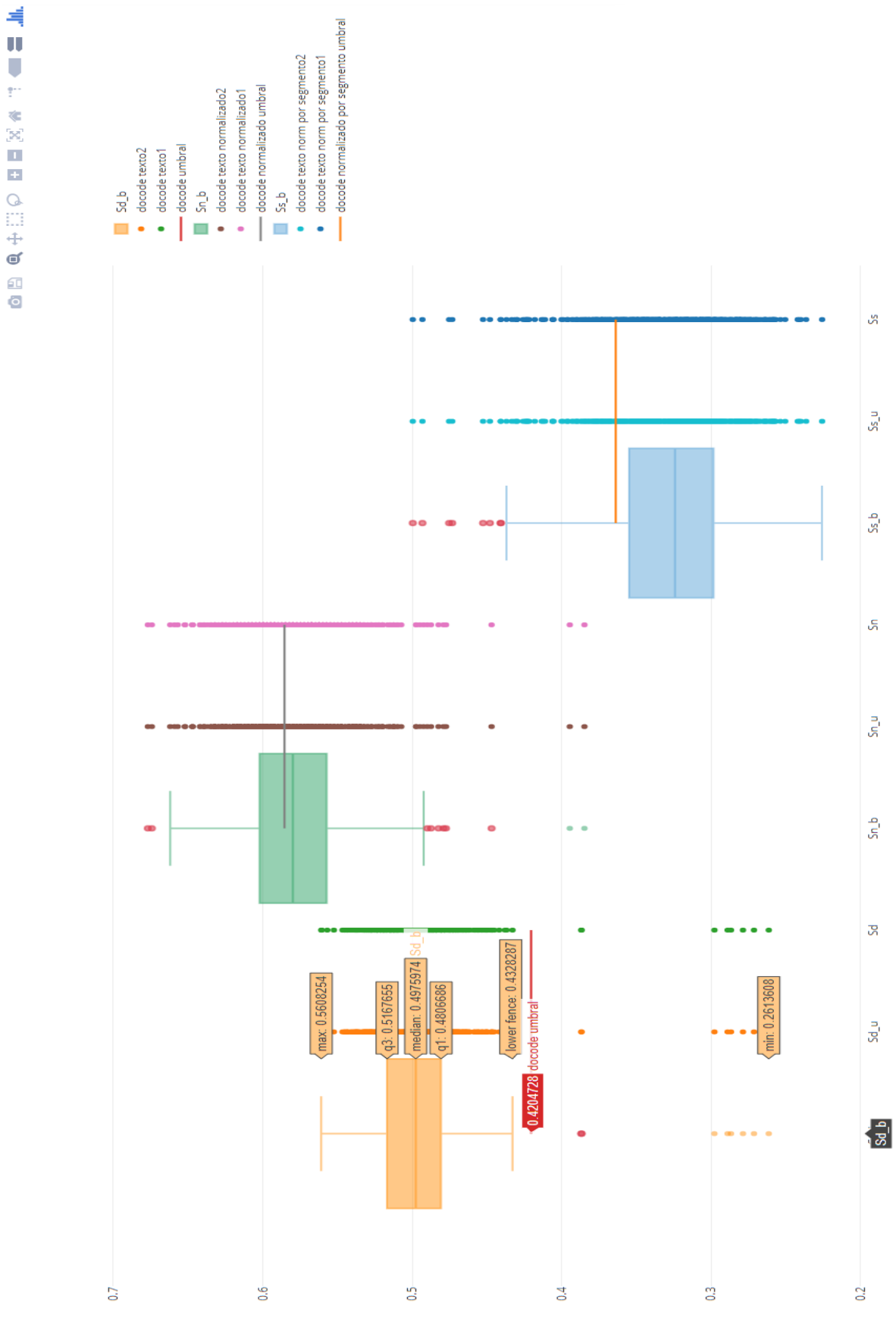


Figura 3: Distribución con porcentaje de pertenencia en cada segmento. Se representan los valores obtenidos por cada algoritmo (Sd: DOCODE, Sn: DOCODE NORMALIZADO y Ss: DOCODE NORMALIZADO POR SEGMENTO).

Con esto también se logra validar que nuestro algoritmo, utilizando el archivo *gramas.py* (explicado en 3.3) y con la configuración base de n=1, obtiene los mismos resultados para los segmentos analizados. Por lo que se valida que la diferencia en el manejo de texto no afecta los resultados y se comprueba su uso.

3.5.2. Experimento 1

Esta fase tiene su principio en la aplicación del algoritmo de por si, con los cambios realizados para resolver el nuevo problema de Verificación de Autores. Si bien en esta fase ya no es posible compararse con [Reyes, 2016], si se pueden generar resultados que nos permitan establecer nuestra posición inicial frente al problema. Para esto utilizaremos una cantidad pequeña de documentos y generaremos una prueba controlada para poder explicar los resultados obtenidos.

Para esta prueba se utilizaron las carpetas de Fase1 descrita en la sección de script 3.3. En este experimento se tienen 8 carpetas de los cuales queremos analizar los estilos de los autores: Jane Austen y Leon Tolstoi por lo cual cada autor tiene 4 carpetas de prueba, las 2 primeras corresponden a la revisión de textos desconocidos que sí corresponden a su autoría y las 2 ultimas carpetas corresponderán a la revisión de textos desconocidos que no correspondan a su autoría. La elección de autores no es relevante y los resultados son analizados en base al funcionamiento de los algoritmos y no sobre las particularidades de los textos o autores.

Si revisamos el excel generado en nuestra carpeta de *Result*, documento: *Resultados.xls*, obtendremos la siguiente tabla, en donde podremos observar los porcentajes de pertenencia γ [5] obtenidos en cada carpeta con cada algoritmo:

Tabla 6: Resultados Experimentos 1 - Comparación de porcentajes de pertenencia por cada carpeta

Solución: Carpetas	Real	DOC	DOC NOR	DOC SEG	Pertenencia γ con DOC	Pertenencia γ con DOC NOR	Pertenencia γ con NOR SEG
EN01	V	V	V	V	0,9680	0,6164	0,8127
EN02	V	V	V	V	0,9830	0,6286	0,8398
EN03	F	V	F	F	0,9635	0,2914	0,4939
EN04	F	V	F	F	0,9184	0,4131	0,3500
EN05	V	V	F	V	0,8197	0,4334	0,7811
EN06	V	F	V	V	0,0625	0,6500	0,9750
EN07	F	V	F	V	0,9716	0,1457	0,6477
EN08	F	V	V	V	0,9789	0,5526	0,6184
Total		3	6	6			

[5] Porcentajes obtenidos de la cantidad de segmentos del documento desconocido que son clasificados como posibles escritos del autor, su definición se formal se detallará en el Experimento 2

En ella podemos revisar los porcentajes de pertenencia (γ) obtenidos al analizar cada segmento del archivo desconocido y compararlo con el umbral generado por los documentos conocidos, obteniendo así, aciertos que serán sumados para poder compararse con la cantidad de segmentos del documento desconocido. En esta prueba podemos hacer uso del archivo *Trace2.txt*, en el cual se podrá revisar, por algoritmo utilizado, como es obtenido el estilo y umbral [6] del documento conocido y como es comparado cada segmento del documento desconocidos, obteniendo por ejemplo en nuestra primera carpeta con el algoritmo de *DOCODE*, un umbral de $0{,}393$ app lo cual es comparado en 219 segmentos del documento desconocido, obteniendo 212 aciertos lo que nos da un porcentaje de que el documento pertenece al autor de un $0{,}968$ app, para *DOCODE NORMALIZADO*, tendremos un umbral de 0.551 app. lo cual es comparado en estos 219 segmentos obteniendo 135 aciertos lo que nos da un porcentaje de pertenencia de un $0{,}616$ app. y para *DOCODE NORMALIZADO POR SEGMENTO*, con un umbral de $0{,}325$ app, revisado en los 219 segmentos se obtienen 178 aciertos, lo que nos da un porcentaje de pertenencia de un $0{,}812$ app.

Dichos resultados dependen de nuestras elecciones fundamentadas en [Gallardo, 2013] y [Reyes, 2016] sobre el mejor λ para cada algoritmo, y la definición que resolvimos dar a los porcentajes, cuando γ es menor a $0{,}5$ se indicará como un resultado Falso, que indicará que la probabilidad de pertenencia obtenido fue menor que $0{,}5$ y se cree que el documento no pertenece al autor. Por el contrario si el γ es mayor o igual a $0{,}5$ se calificará como Verdadero y se determinará que es probable que el documento pertenezca al autor de los textos conocidos. Bajo esta condición y para esta prueba obtenemos que, para el algoritmo de *DOCODE*, 3 de 8 carpetas se han calificado correctamente, para *DOCODE NORMALIZADO* 6 de 8 carpetas se han calificado correctamente y para *DOCODE NORMALIZADO POR SEGMENTO* lo han hecho, nuevamente y no con la mismas soluciones, 6 de 8 carpetas.

De forma visual se podrá revisar los gráficos por carpetas, donde en el ejemplo anterior obtendremos la Figura 4:

[6] Los λ's utilizados son los entregados por [Reyes, 2016] para obtener Mejor F_1.

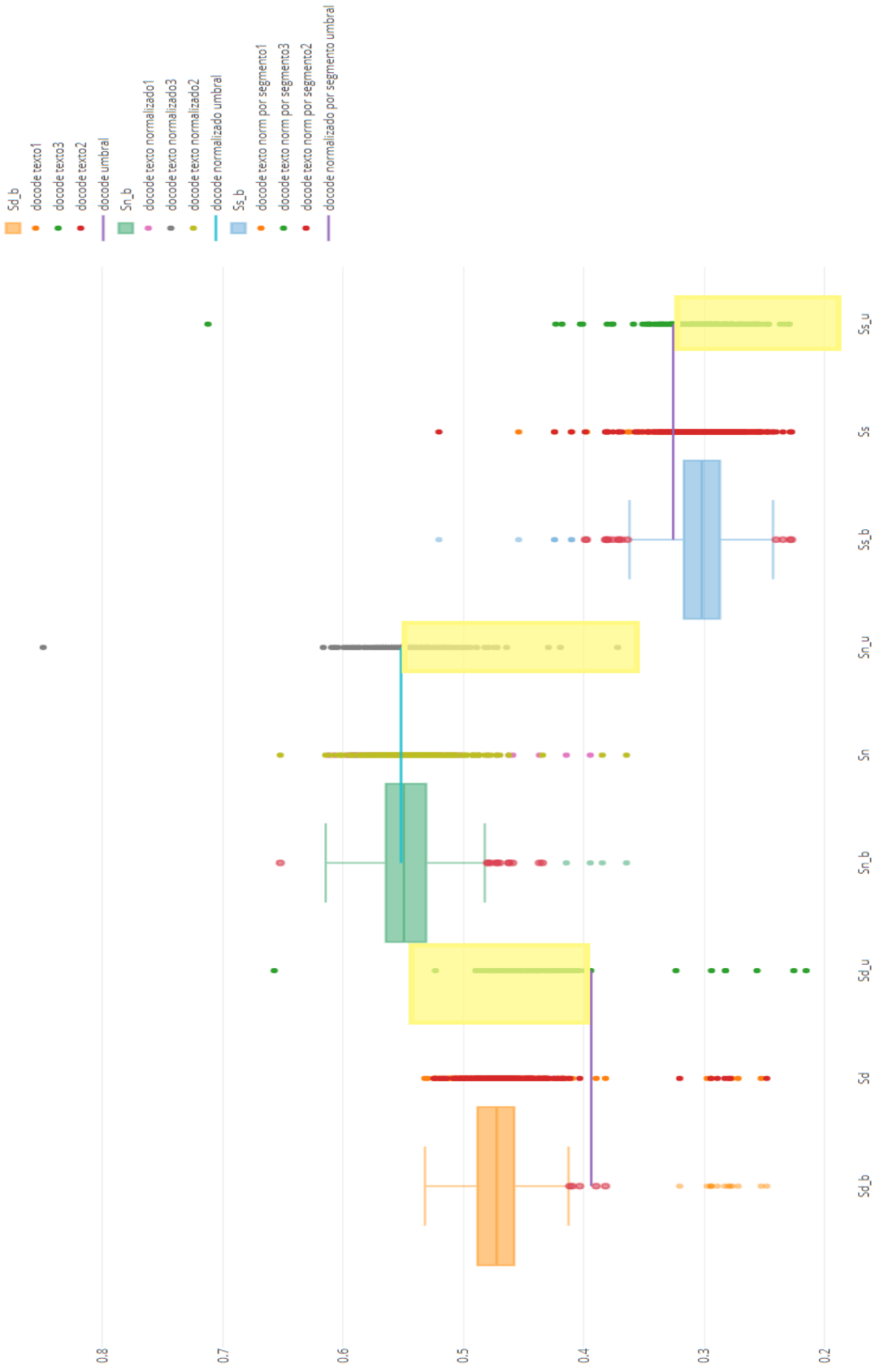


Figura 4: Gráfico de distribución con carpetas de Fase 1

Gráfico en el cual se podrá observar que (se marca en amarillo los segmentos descritos), para *DOCODE*, la mayor cantidad de segmentos del documento desconocido (en nuestro ejemplo, documento 3, en color verde) se encuentran arriba de su umbral (línea morada), para *DOCODE NORMALIZADO* la mayoría de segmentos el color gris que representan el texto desconocido se encuentra a bajo de su umbral (línea calipso) y que para *DOCODE NORMALIZADO POR SEGMENTO* la mayoría de sus segmentos del documento desconocido (puntos en color verde) están a bajo de su umbral (línea calipso). Por lo que será una forma visual simple y rápida de analizar las carpetas por separado y determinar si será catalogado como perteneciente al autor o no.

3.5.3. Experimento 2

En esta última fase el objetivo principal del análisis se enfoca en el criterio de clasificación. En esta, observaremos los valores obtenidos y determinaremos las propuestas a utilizar.

Como pudimos observar, en nuestra fase anterior tenemos ciertas atribuciones fundamentadas en experimentos con el Problema de Plagio Intrínseco y aquellas que hemos definido arbitrariamente para poder generar resultados comparables para nuestro problema.

Es en este punto es donde podremos analizar la influencia del λ en cada uno de nuestros algoritmos y como, generando los diferentes resultados con cada uno de estos λ's, podremos caracterizar la importancia de la constante de calificación para cada uno de ellos.

Si por ejemplo, utilizando las mismas carpetas de la fase anterior, observamos los resultados de los archivos *Hist-pertenencia-TODOS.html* y *Hist-aciertos-TODOS.html*, podremos ver que para el algoritmo de *DOCODE NORMALIZADO* con $\lambda = 0{,}2$ y un porcentaje de discriminación de mayor o igual al $\gamma \geq 0{,}5$ obtenemos una cantidad de 6 aciertos (3 aciertos catalogados como verdaderos y 3 aciertos catalogados como falsos, revisando la Figura 5). Pero si revisamos los gráficos, esta cantidad de aciertos no es la mayor cantidad que se puede obtener en estos algoritmos, la mayor cantidad de aciertos es de 7 para *DOCODE NORMALIZADO* y de 8 aciertos para *DOCODE NORMALIZADO POR SEGMENTOS* los cuales se logran con múltiples combinaciones de λ y porcentajes discriminatorios, por ejemplificar se puede mencionar los valores de $\lambda = 0{,}8$, $\gamma \geq 0{,}6$ y $\lambda = 0{,}3$, $\gamma \geq 0{,}5$ respectivamente (6).

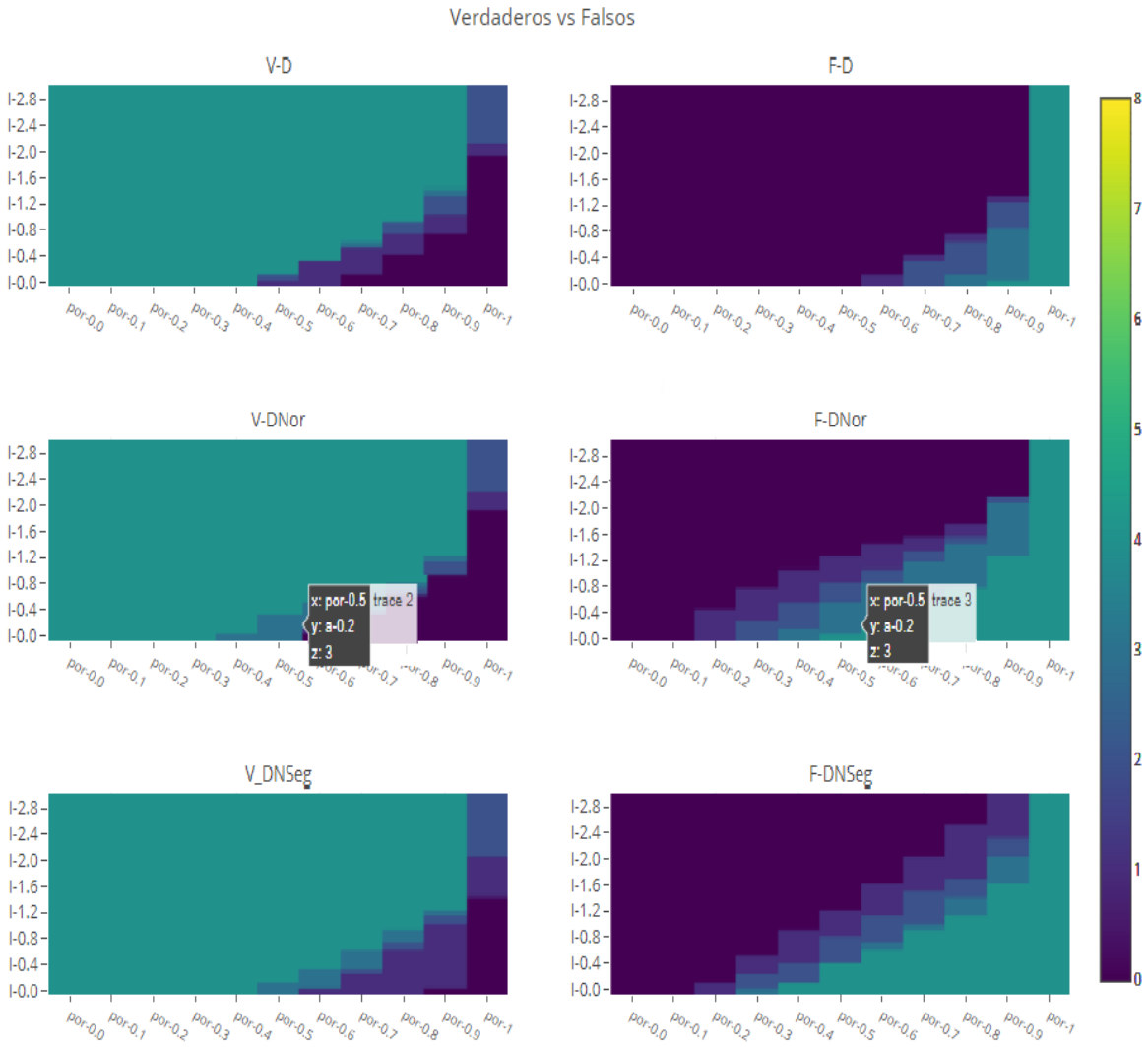


Figura 5: Histograma de la cantidad de aciertos verdaderos y la cantidad de aciertos falsos por algoritmo. Donde se revisa por cada eje y el λ y por cada eje x el γ utilizados.

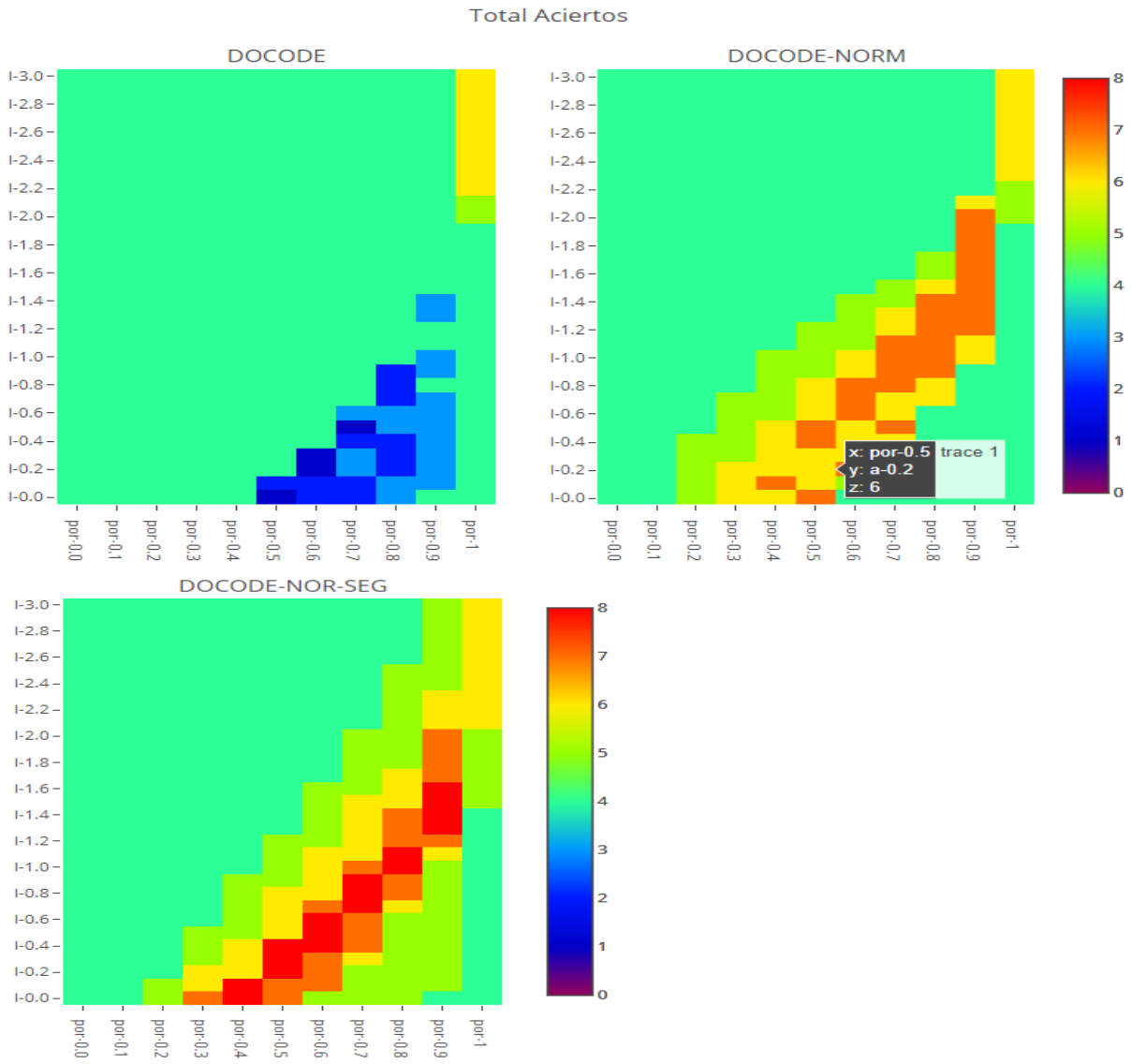


Figura 6: Histograma de la cantidad de aciertos por algoritmo. Donde se revisa por cada eje y el λ y por cada eje x el γ utilizados.

Esto nos presenta que si bien los textos desconocidos se están catalogando de acuerdo a los algoritmos presentado por [Reyes, 2016], estos deben analizarse con respecto a las características propias del problema de Verificación de Autores. Comprobando que cada respuesta, tanto positiva como negativa, generará un valor importante en este problema.

Para determinar un λ óptimo, se puede recurrir al mejor F_1 (revisado en la sección 3.2.1) el cual nos ayudará a determinar nuestros valores óptimos tanto para λ como para el porcentajes discriminatorios. De esta forma se construye el siguiente gráfico de calor, que puede ser revisado en el archivo *Hist-Mejor-F1-TODOS.html* y en el cual podremos asegurar que si bien podríamos tener una cantidad de aciertos considerables no significaría que nuestro F_1 sea óptimo, ya que podría deberse a aciertos aleatorios más que a una configuración entrenada del algoritmo.

En nuestro ejemplo anterior, considerando para el *DOCODE NORMALIZADO* los valores de $\lambda = 0{,}8$, $\gamma \geq 0{,}6$, se tendrá un $F_1 = 0{,}888$, por lo cual validaremos que es el mejor F_1 posible (Figura 7) y los valores determinados si son óptimos dentro de este escenario.

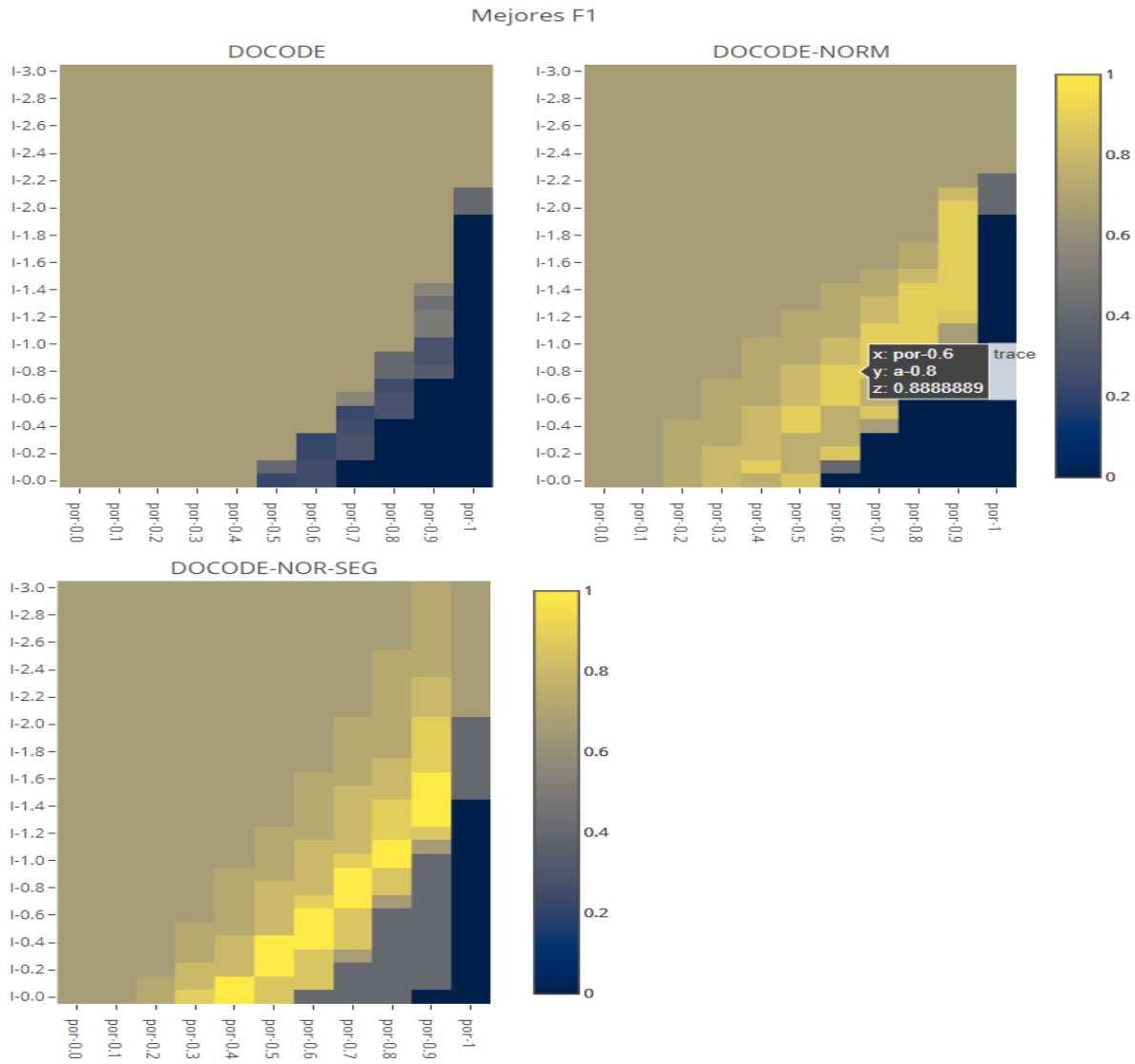


Figura 7: Histograma de los F_1 obtenidos por algoritmo. Donde se revisa por cada eje y el λ y por cada eje x el γ utilizados.

Ya discutido la importancia del λ , se debe generar la discusión sobre nuestro porcentaje discriminatorio, si bien ahora vemos que se encuentra arbitrariamente fijo, se podrá revisar que este no depende de un sólo pivote.

Si realizamos una prueba sobre el comportamiento que tienen los aciertos, ya sean verdaderos o falsos en su definición, deberemos en principio definir una estructura para dicho porcentaje.

Esta estructura se considerará entre $[0, 1]$, ya que se tiene para una cantidad n de segmentos d_c del documento desconocido que, la posibilidad de no acertar con ningún segmento es de 0 aciertos y la posibilidad de acertarlos todos será de $n/n = 1$ (recordando que este acierto se define según su algoritmo como $d_c \geq Umbral$ para el caso de *DOCODE* y $d_c <= Umbral$ para los otros algoritmos. En este ejemplo vemos 7 segmentos, la posibilidad de acertarlos todos es 1, de no acertar ninguno es 0 y de acertar 3 de 7 es $0{,}43$.

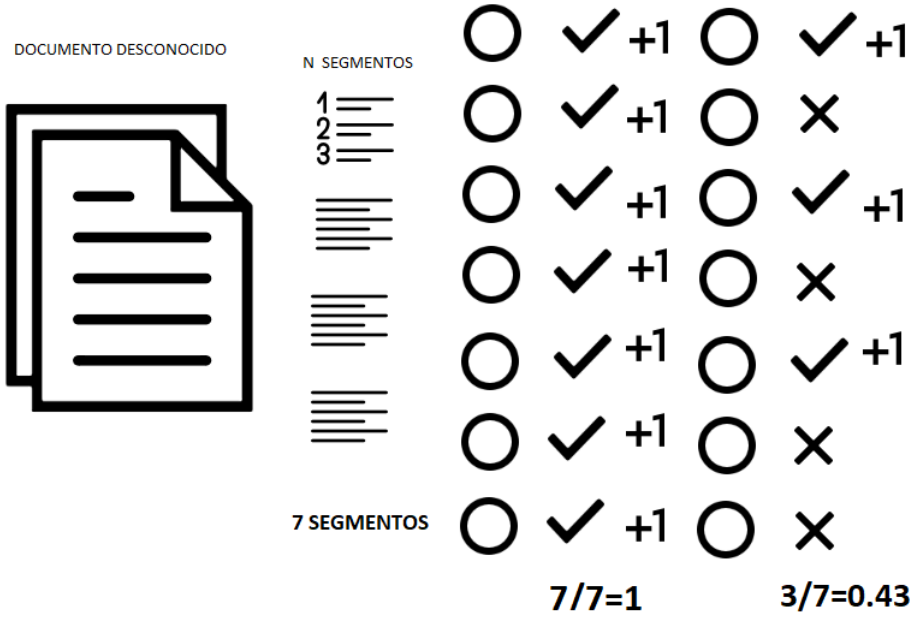


Figura 8: Imagen: Segmentos y aciertos

Para definir entonces los cortes se debe considerar que para una cantidad de aciertos m que sea muy pequeña con respecto a la cantidad total de segmentos n, se deberán analizar segmentos pequeños, pero si m es cercano a n sus segmentos podrá ser relativamente más grandes. Como en este punto sabemos que los experimentos a realizar no serán con variables fijas se opta por una segmentación arbitraria de 10. Obteniendo que en estos segmentos y este escenario, la cantidad de aciertos verdaderos y aciertos falsos se distribuye de esta forma (Figura 9):

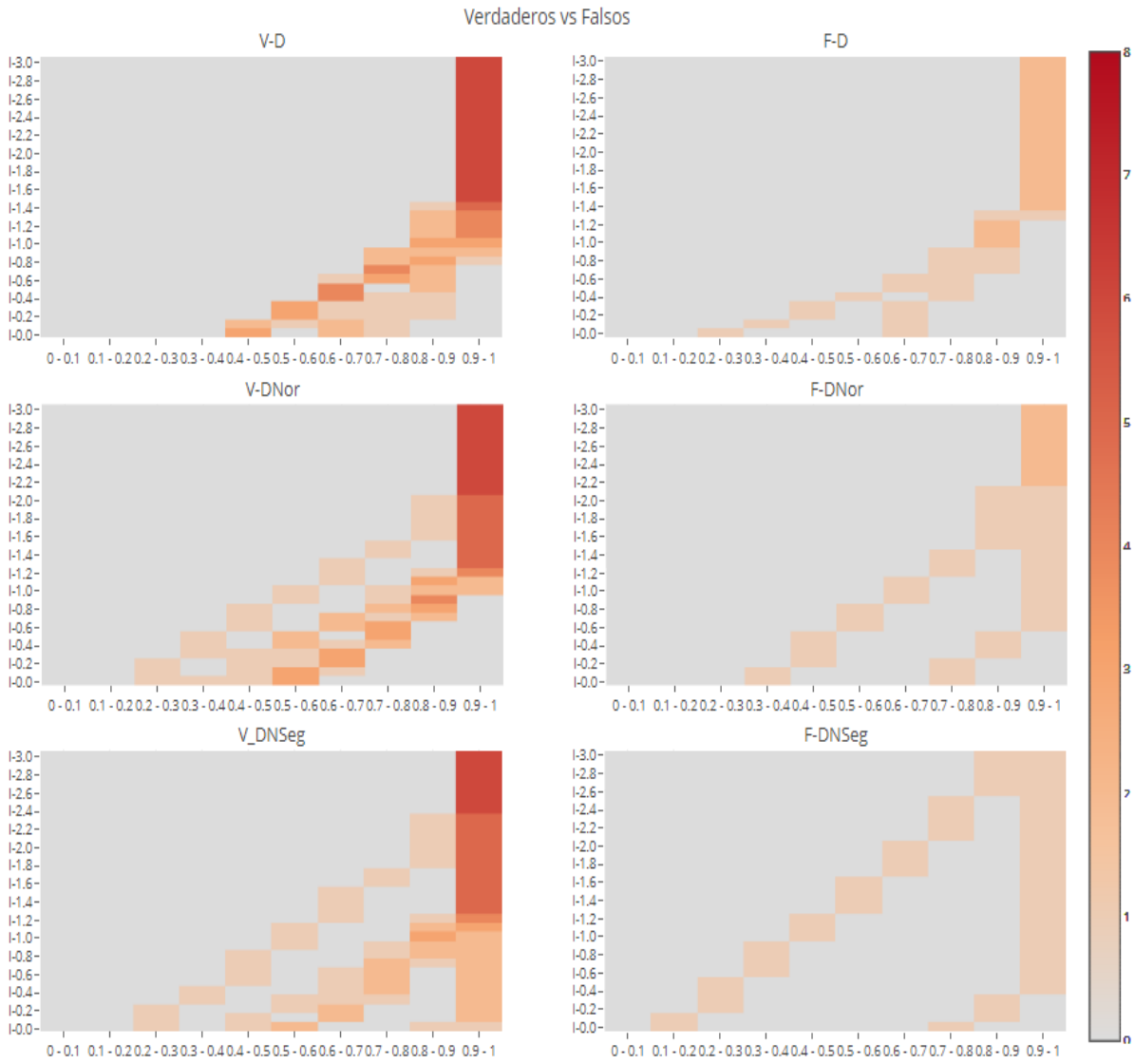


Figura 9: Histograma de la cantidad de aciertos reales verdaderos y la cantidad de aciertos reales falsos por algoritmo. Donde se revisa por cada eje y el λ y por cada eje x los aciertos de los valores obtenidos por cada segmento.

Lo que nos demuestra que el porcentaje discriminatorio no es necesariamente un valor fijo que límite los verdaderos de los falsos si no que, posiblemente sean conjunto de valores limitantes, lo que llamaremos *segmento limitante* para describir en cada punto si el segmento es preferiblemente verdadero o falso por el F_1 obtenidos en los experimentos de prueba. Para entender esta característica se crea una prueba donde se revisará cada segmento de porcentaje discriminatorio con un segmento limitante per se. Por lo que si bien un segmento limitante óptimo podría darse sólo con un punto que este entre $[0, 1]$ podría también darse el caso que mi segmento limitante tenga 9 puntos entre $[0, 1]$, lo cual no sería ni óptimo ni concluyente, pero si podría ser posible.

La cantidad de aciertos obtenidos con un resultado verdadero o con un resultado falso se puede revisar tanto en la figura anterior 9 como en la Figura siguiente 11,si es que no hemos elegido un segmento limitante en nuestro script de VerificacionAutores.py (3.3). Si no hemos elegido segmento, la Figura siguiente (10) nos mostrará para cada λ propuesto la combinación óptima de segmento limitante, la cual si revisamos en la figura anterior podremos conocer la cantidad de aciertos obtenidos con cada variable. El segmento se leerá como un conjunto de verdaderos (1) y falsos(0) en cada uno de los diez cortes definidos anteriormente de $[0, 1]$. Su lectura se realizará en el gráfico de "Mejores V o F" de cada algoritmos, identificando el mejor λ por el mejor valor obtenido de F_1 (toda la información se saca del mismo gráfico). Por ejemplo para el ejercicio anterior se obtiene la siguiente información para DOCODE:

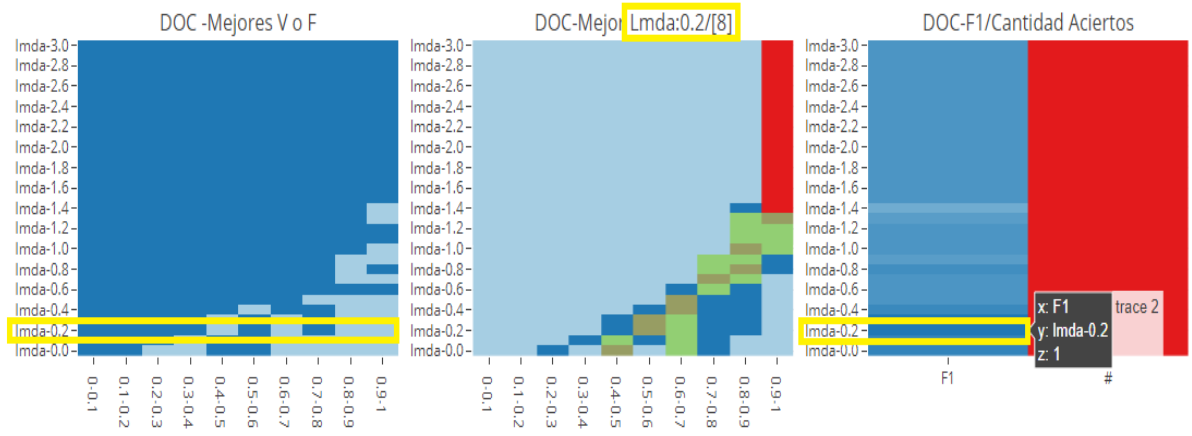


Figura 10: Revisión de segmento limitante, λ y cantidad de aciertos.

Donde se puede revisar rápidamente cual es el mejor λ, en este caso $0{,}2$, con 8 aciertos, y en el gráfico "Mejores V o F" se obtiene por este λ el conjunto [1,1,1,0,1,0,1,0,0], lo cual nos dice que, en los aciertos obtenidos entre 0 a $0{,}3$ es mejor considerarlos como verdaderos (1) ya que se obtendrá un mejor F_1, entre $0{,}4$ a $0{,}5$ es mejor considerarlos como falso (0), entre $0{,}5 - 0{,}6$ serán verdadero, entre $0{,}6 - 0{,}7$ nuevamente falso, $0{,}7 - 0{,}8$ verdadero y entre $0{,}8 - 1$ es mejor considerarlo como falso [7].

[7]Existen resultados en los segmentos que no son propios de aciertos y se dan de forma automática, por

Obtenido el valor de los segmentos limitantes y ejecutando el script con esta configuración en una carpeta de *test*, la cual nos ayudará en experimentos posteriores a comparar valores obtenidos, se genera la siguiente Figura 11 que puede ser revisada en el archivo *B0B1.html*. En este nos mostrará en "Cantidad Aciertos" el total de aciertos obtenidos y el F_1 obtenido con esta configuración.

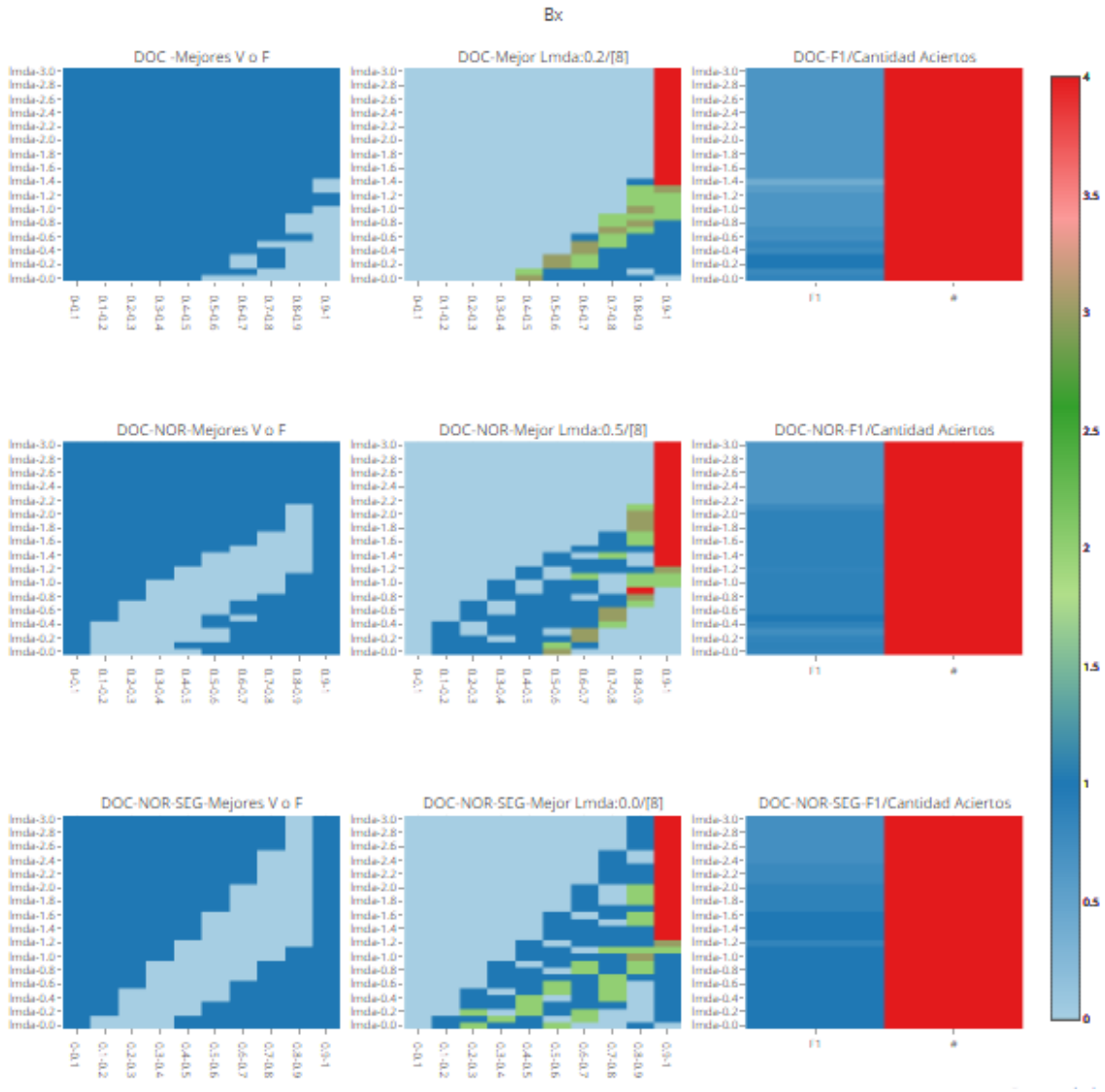


Figura 11: Histogramas del segmento limitante, Histograma con cantidad de aciertos obtenido con el segmento limitante e histograma que muestra el F_1 obtenido y su cantidad de aciertos totales.

ejemplo, cuando existe la misma cantidad de aciertos para carpetas verdaderas y falsas o si no existe ningún acierto en ambas, el algoritmo utilizad la última designación del segmento anterior para hacer la designación en el segmento actual. Esto explica por qué se puede asignar un segmento $0{,}9 - 1$ como falso.

Y en el cual se puede observar que para el algoritmo de *DOCODE* la mejor configuración viene dada por un $\lambda = 0{,}5$, con un segmento limitante de un punto en $0{,}7$, con el cual se obtiene un $F_1 = 0{,}857$ y una cantidad de aciertos de 7 aciertos de 8.

Esto se puede revisar para cada configuración, obteniendo tanto el segmento limitante como el λ óptimo para este segmento. Estas configuraciones se podrán estudiar en los entrenamientos y ser utilizadas en las pruebas como veremos en el capitulo de Validación de la Solución 4.2.

3.5.4. Experimento 3

Finalmente, un punto importante de la configuración es revisar el n-grama a utilizar. En este trabajo sólo nos enfocaremos en realizar experimentos con unigrama ($n = 1$) y bigrama ($n = 2$) para poder analizar los resultados obtenidos.

El n-grama es una subsecuencia de n elementos, palabra o carácter, que se utiliza como núcleo principal del documento. Para el problema de Atribución de Autores lo más común es tomar los n-gramas del documento, calcular la frecuencia o el conjunto de n-gramas y finalmente calcular la distancia con alguna función. En nuestro caso se utilizará el valor de n para definir como se trabajará el texto en lo que hemos explicado como Experimento 0 3.5.1.

Su importancia radica principalmente en que, siendo el núcleo a trabajar (al realizar el preprocesamiento del texto), dependerá del algoritmo a utilizar para determinar si le conviene analizar la frecuencia de ciertas palabras o es mucho más importante revisar el conjunto de palabras que la misma palabra como tal, lo siguiente lo podremos revisar con el ejemplo a continuación, donde de forma básica comparamos una frase de un texto 1 con un texto 2, dejando en evidencia la cantidad de aciertos que tienen. Para este ejemplo se entenderá como el uso de bigrama podría influenciar en la comparación de textos, si bien existen palabras que por si solas tienen una gran cantidad de aciertos ya que son ampliamente utilizadas en nuestro lenguaje, si uno las revisa con las palabras de su contexto pueden darnos una mayor información en su comparación. **Cabe destacar que su uso dependerá también del elemento a utilizar, el algoritmo con el cual se trabaje y los documentos a revisar, ya que no siempre bigrama es mejor que unigrama o viceversa, y por distintas características el n óptimo a utilizar puede cambiar.**

Figura 12: Ejemplo de palabras en unigrama y bigrama

En nuestros experimentos el uso de n irá cambiando y generando nuevas configuraciones tanto para unigrama como para bigrama donde se irá comparando con los experimentos realizados, en esta sección sólo revisaremos como afecta el n en nuestros algoritmo. Tomando el mismo ejemplo de la fase anterior y ejecutando el *script* con sólo cambiar de $n = 1$ a $n = 2$, podremos notar que si bien el número de aciertos es el mismo, este dispone de λ's óptimos distintos debido a que también los valores de sus segmentos son distintos y por ende, sus umbrales y luego su configuración también podrá ser distinta.

Gráficamente la mejor forma de compararlo es revisando el gráfico por carpeta (figuras 13,14 y 15), tomaremos la primera carpeta en cuestión y revisaremos los resultados obtenidos en los dos experimentos, para cada algoritmo, lo que se muestra a continuación. En estos gráficos podemos ver ciertas diferencias por ejemplo en el caso de DOCODE y DOCODE NORMALIZADO, con bigrama, la cantidad de segmentos que están bajo y sobre el umbral respectivamente son mayores que en su situaciones con unigrama. Para el caso de DOCODE NORMALIZADO POR SEGMENTO esta afirmación no es cierta y es en el caso de unigrama donde logra tener una mayor cantidad de segmento bajo el umbral.

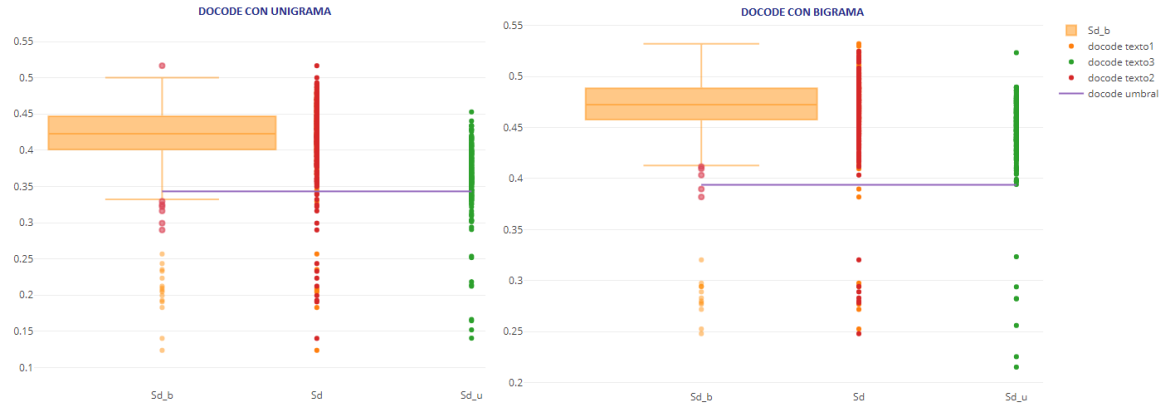


Figura 13: DOCODE con unigrama y bigrama

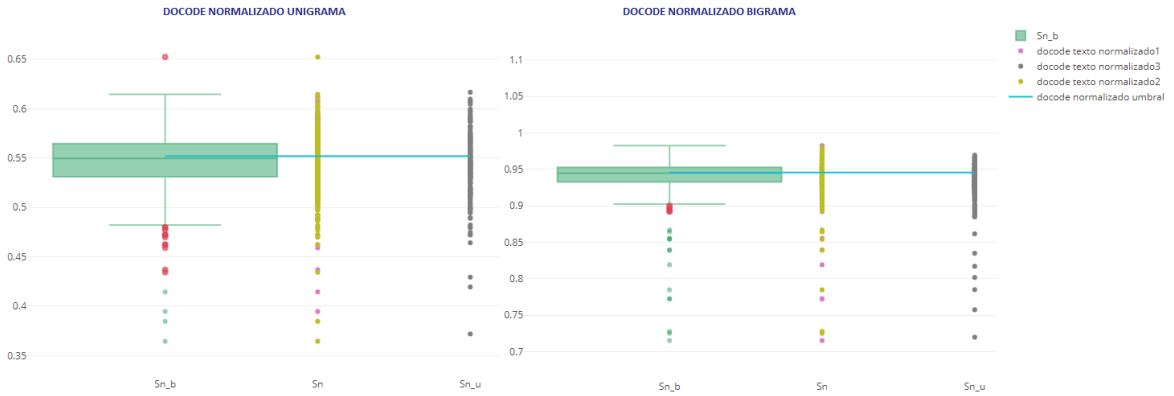


Figura 14: DOCODE NORMALIZADO con unigrama y bigrama

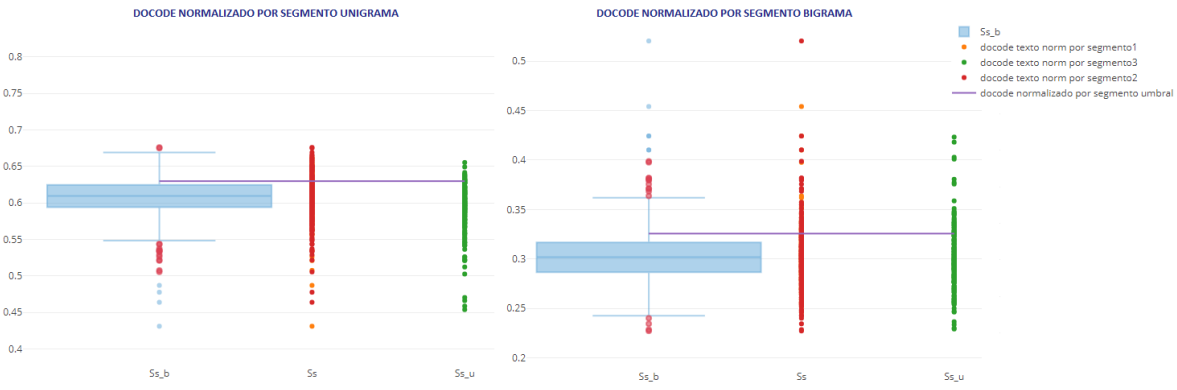


Figura 15: DOCODE NORMALIZADO POR SEGMENTO con unigrama y bigrama

CAPÍTULO 4
VALIDACIÓN DE LA SOLUCIÓN

CLEF [CLE, 2018] es una conferencia anual que contribuye a la evaluación sistemática de diferentes campos de evaluación y un conjunto de laboratorios y talleres. PAN [PAN, 2017] es una de las competencias que se ha integrado a esta conferencia hace 10 años y que lleva 18 años realizando evaluaciones forenses en textos digitales, tiempo en el cual han revisado problema de *Authorship Attribution*, *Plagiarism Detection* y *Credibility Analysis* [8].

El problema de Verificación de Autores se ha llevado a cabo los años: 2013, 2014 y 2015, siempre con pequeñas variaciones. Para el 2013 se consideró un número pequeño para el *set* de datos y sólo se consideraron 3 idiomas, inglés, español y griego, para el año 2014 se consideraron 6 idiomas y un *set* de datos mucho más amplio, finalmente para el 2015 se bajo la cantidad de idiomas utilizados a 4 como también se redujo el set de datos de *training* y se amplio el *set* de datos de *test*. Para poder analizar los experimentos realizados con estas carpetas se realizará una descripción de cada una para luego describir las pruebas y los resultados obtenidos en cada una de ellas.

4.1. Descripción de Carpetas de la Competencia

- PAN 2013

  Se tienen 3 carpetas en idioma inglés, una de *training* (de 10 autores) y dos de *test*, corpus 1 (de 20 autores) y corpus 2 (de 30 autores). Con un promedio de 5192 palabras por documento. Se tiene la siguiente tabla:

Tabla 7: Resumen Carpetas PAN 2013

	Training Corpus		Test Corpus 1		Test Corpus 2	
Cantidad de Autores	10		20		30	
Cantidad de:	Palabra	Carácter	Palabra	Carácter	Palabra	Carácter
En Texto Conocido	33399	203598	93063	557671	132922	794224
En Texto Desconocido	10350	63317	21198	128980	31765	191359
Promedio en Texto Conocido	3339	20359	4653	27883	4430	26474
Promedio en Texto Desconocido	334	6331	1059	6449	1059	6379
Promedio por Documentos	4374	26692	5713	68665	5490	32853

[8]Este último, compone una nueva categoría, no muy bien definida, que intenta resolver problemas como, de vandalismo en textos, autores que comprometen un texto original quitando, cambiando o agregando texto que comprometen la integridad del documento original, o *Deception Detection*, la cual se concibió como el problema de identificar individuos que buscan favores sexuales dentro de conversaciones en linea.

- PAN 2014

  Se tienen 6 carpetas en idioma inglés, dos de *training* divididas en novela y ensayo y dos de *test* por cada una de esta división, corpus 1 y corpus 2 para novelas y para ensayos. Con un promedio de 10.226 palabras para novelas y un promedio de 3.174 palabras para los documentos de ensayos. Se tiene la siguiente tabla:

Tabla 8: Resumen Carpetas PAN 2014 - Novelas

Novelas	Training Corpus		Test Corpus 1		Test Corpus 2	
Cantidad de Autores	100		100		200	
Cantidad de:	Palabra	Carácter	Palabra	Carácter	Palabra	Carácter
En Texto Conocido	450977	2540970	512624	2869961	1041743	5822170
En Texto Desconocido	181269	1020878	693917	3918792	1416297	7999111
Promedio en Texto Conocido	4509	25409	5126	28699	5208	29110
Promedio en Texto Desconocido	1812	10208	6939	39187	7081	39995
Promedio por Documentos	6322	35618	12065	67887	12290	69106

Tabla 9: Resumen Carpetas PAN 2014 - Ensayos

Ensayos	Training Corpus		Test Corpus 1		Test Corpus 2	
Cantidad de Autores	100		100		200	
Cantidad de:	Palabra	Carácter	Palabra	Carácter	Palabra	Carácter
En Texto Conocido	472488	2537624	238567	1282865	461552	2478902
En Texto Desconocido	170574	914837	81032	435577	161091	865207
Promedio en Texto Conocido	2362	12688	2385	12828	2307	12394
Promedio en Texto Desconocido	852	4574	810	4355	805	4326
Promedio por Documentos	3215	17262	3195	17184	3113	16720

- PAN 2015

  Se tienen 2 carpetas en idioma inglés, una de *training*(de 100 autores) y una de *test* (de 500 autores). Con un promedio de 1.095 palabras por documento. Se tiene la siguiente tabla:

Tabla 10: Resumen Carpetas PAN 2015

	Training Corpus		Test Corpus	
Cantidad de Autores	100		500	
Cantidad de:	Palabra	Carácter	Palabra	Carácter
En Texto Conocido	52849	199739	314016	1541458
En Texto Desconocido	52489	195768	254226	1146086
Promedio en Texto Conocido	528	1997	628	3082
Promedio en Texto Desconocido	524	1957	508	2292
Promedio por Documentos	1053	3955	1136	5375

4.2. Descripción de Experimentos y Resultados

Para cada experimento, se utilizarán las siguientes carpetas de forma aislada:

- PAN 2013
- PAN 2014
- PAN 2015
- Nivel 0

Dentro de las cuales se resolverá tomar una carpeta contenida dentro de ellas, ya sea *training* o *test*, como carpeta de entrenamiento. Esto se debe a la suposición de que carpetas con poca o mucha cantidad de palabras por documentos, podrán impactar al probar carpeta con un mayor o un menor promedio de palabras por documento, por lo que la elección de carpetas de entrenamientos será en base a esta suposición y se elegirá, en medida de lo posible, la carpeta con un promedio de palabras en la mediana de nuestros resultados.

Al aplicar el algoritmo se utilizará un segmento de cantidad de palabras, $m = 400$, fijo para todos los experimentos, y un número de n-gramas de uno y dos $n = 1, 2$. Para cada prueba realizada en la carpeta de *training* se obtendrá un λ óptimo y un segmento limitante que vendrán determinados por el F_1 óptimo.

Con esta configuración se procederá a probar la(s) carpeta(s) definida(s) como parte del *test*, y obtener los resultados. También, con el propósito de realizar comparaciones sobre esta, se dejará la información de los resultados con la configuración de λ óptimo dada por los autores de los algoritmos y detallada en la sección de Análisis de Algoritmos 3.1.

Los resultados obtenidos para cada carpeta son:

- PAN 2013

  Se utiliza la carpeta de Test Corpus 2 como *training* (30 autores) y se obtienen los siguientes resultados:

Tabla 11: Resultados PAM13 - Test Corpus 1 (20 autores)

PAN 2013	DOCODE			DOCODE NORM			DOCODE NOR SEG		
Test Corpus 1	lmda	aciertos	F_1	lmda	aciertos	F_1	lmda	aciertos	F_1
Con λ de autores anteriores	0.075	9	0.444	0.2	8	0.454	0.8	9	0.592
n=1	0.3	13	0.500	0.6	15	**0.705**	0.8	14	0.625
n=2	0.4	15	**0.705**	1.3	14	0.461	0.0	14	0.700

Tabla 12: Resultados PAM13 - Training (10 autores)

PAN 2013	DOCODE			DOCODE NORM			DOCODE NOR SEG		
Test Corpus 1	lmda	aciertos	F_1	lmda	aciertos	F_1	lmda	aciertos	F_1
Con λ de autores anteriores	0.075	5	0.250	0.2	6	0.250	0.8	5	0.333
n=1	0.3	6	0.400	0.6	6	0.285	0.8	7	0.363
n=2	0.4	6	0.500	1.3	7	0.444	0.0	6	**0.615**

- PAN 2014

  Novelas

  Se utiliza la carpeta de Test Corpus 1 como carpeta de *training*(100 autores) y se obtienen los siguientes resultados:

Tabla 13: Resultados PAM14 - Novela - Test Corpus 2 (100 autores)

PAN 2014	DOCODE			DOCODE NORM			DOCODE NOR SEG		
Test Corpus 2	lmda	aciertos	F_1	lmda	aciertos	F_1	lmda	aciertos	F_1
Con λ de autores anteriores	0.075	99	0.542	0.2	86	0.462	0.8	96	0.547
n=1	0.4	130	0.694	1.2	116	0.428	1.5	135	0.636
n=2	0.2	121	0.663	1.2	142	**0.718**	0.2	113	0.536

Tabla 14: Resultados PAM14 - Novela - Training (100 autores)

PAN 2014	DOCODE			DOCODE NORM			DOCODE NOR SEG		
Training	lmda	aciertos	F_1	lmda	aciertos	F_1	lmda	aciertos	F_1
Con λ de autores anteriores	0.075	39	0.600	0.2	48	0.540	0.8	55	**0.661**
n=1	0.4	68	0.360	1.2	57	0.307	1.5	61	0.385
n=2	0.2	73	0.390	1.2	62	0.418	0.2	65	0.265

Ensayos

Se utiliza la carpeta de Test Corpus 2 como *training* (200 autores) y se obtienen los siguientes resultados:

Tabla 15: Resultados PAM14 - Ensayo - Test Corpus 1 (100 autores)

PAN 2014	DOCODE			DOCODE NORM			DOCODE NOR SEG		
Test Corpus 2	lmda	aciertos	F_1	lmda	aciertos	F_1	lmda	aciertos	F_1
Con λ de autores anteriores	0.075	50	0.509	0.2	51	0.625	0.8	51	0.657
n=1	1.6	54	0.651	2.7	55	0.689	2.2	53	0.680
n=2	0.0	61	0.589	2.3	56	**0.694**	0.1	57	0.576

Tabla 16: Resultados PAM14 - Ensayo - Training (200 autores)

PAN 2014	DOCODE			DOCODE NORM			DOCODE NOR SEG		
Training	lmda	aciertos	F_1	lmda	aciertos	F_1	lmda	aciertos	F_1
Con λ de autores anteriores	0.075	102	0.435	0.2	105	0.639	0.8	105	0.664
n=1	1.6	119	0.447	2.7	112	**0.657**	2.2	111	0.664
n=2	0.0	112	0.505	2.3	103	0.645	0.1	108	0.619

- PAN 2015

  Se utiliza la carpeta de *training*(100 autores) y se obtienen los siguientes resultados:

Tabla 17: Resultados PAM15 - Test (500 autores)

Nivel 0	DOCODE			DOCODE NORM			DOCODE NOR SEG		
Test	lmda	aciertos	F_1	lmda	aciertos	F_1	lmda	aciertos	F_1
Con λ de autores anteriores	0.075	250	0.419	0.2	264	0.587	0.8	294	0.637
n=1	1.6	276	0.558	1.4	283	**0.657**	0.0	292	0.632
n=2	0.0	287	0.431	0.0	292	0.414	0.3	280	0.598

- Nivel 0

  Se utiliza la carpeta de *training*(100 autores) y se obtienen los siguientes resultados:

Tabla 18: Resultados Nivel 0 - Test (100 autores)

Nivel 0	DOCODE			DOCODE NORM			DOCODE NOR SEG		
Test	lmda	aciertos	F_1	lmda	aciertos	F_1	lmda	aciertos	F_1
Con λ de autores anteriores	0.075	32	0.529	0.2	50	0.468	0.8	59	0.709
n=1	0.5	68	0.645	0.4	71	0.666	0.7	69	0.724
n=2	0.2	71	0.694	0.0	73	0.722	0.1	67	**0.748**

4.3. Comparación con Competencia

Como se ha discutido anteriormente el objetivo de nuestros experimentos es revisar el efecto que tiene el promedio de palabras y la cantidad de autores a analizar lo cual, si queremos analizar con la competencia, deberá variar significativamente. Para tener una idea de esto utilizaremos las carpetas de la competencia PAN 2013 ya que en las siguientes competencias comienza a aplicar un factor significante para sus pruebas, que considera las carpetas que no pueden ser definidas como propias del autor o no, y en nuestro caso esto no sucedería.

Para poder revisar esta carpeta y poder compararnos con los resultados de la competencia deberemos ejecutar los algoritmos sobre la carpeta de *test*, obtener las configuraciones óptimas y realizar una ejecución sobre las carpetas de validación, corpus de la competencia con las cuales revisaremos los resultados obtenidos.

En la ejecución de la carpeta de prueba, con unigrama ($n = 1$) obtenemos:

- DOCODE
  Se realizan 7 aciertos de 10, con un $F_1 = 0{,}769$, un $\lambda = 0{,}7$ y un segmento limitante de [1,1,1,0,0,0,0,0,0,1]

- DOCODE NORMALIZADO
  Se realizan 7 aciertos de 10, con un $F_1 = 0{,}769$, un $\lambda = 1{,}0$ y un segmento limitante de [0,0,0,0,0,0,1,1,1,1]

- DOCODE NORMALIZADO POR SEGMENTO
  Se realizan 8 aciertos de 10, con un $F_1 = 0{,}800$, un $\lambda = 1{,}0$ y un segmento limitante de [1,1,1,1,1,1,0,0,0,1]

En la ejecución de la carpeta de prueba, con bigrama ($n = 2$) obtenemos:

- DOCODE
  Se realizan 7 aciertos de 10, con un $F_1 = 0{,}727$, un $\lambda = 0{,}0$ y un segmento limitante de [1,1,1,1,1,1,0,0,0,1]
- DOCODE NORMALIZADO
  Se realizan 7 aciertos de 10, con un $F_1 = 0{,}769$, un $\lambda = 1{,}0$ y un segmento limitante de [1,1,1,1,1,1,0,0,0,1]
- DOCODE NORMALIZADO POR SEGMENTO
  Se realizan 6 aciertos de 10, con un $F_1 = 0{,}666$, un $\lambda = 0{,}1$ y un segmento limitante de [1,1,1,1,1,1,0,0,0,1]

Lo cual se procederá a utilizar con el corpus y sacar sus F_1, *Precision* y *Recall*. En esta parte se advierte una diferencia importante con lo que nosotros obtenemos de nuestra Matriz de Confusión, explicada anteriormente (sección 3.2.1), ya que si bien F_1 se mantiene la competencia obtiene las otras dos variables de la siguiente forma:

$$Recall = \frac{respuestas\_correctas}{cantidad\_de\_problemas}$$

$$Precision = \frac{respuestas\_correctas}{cantidad\_de\_respuestas}$$

Por lo que le da énfasis tanto a los *True Positive* como a los *True Negative* que se estarían considerando como parte de las respuestas correctas. Aparte se observa que, si se resuelven todos los problemas (no existen carpetas en donde no se sepa si el autor ha escrito o no el documento) las dos variables tendrán el mismo valor.

Con esto y resolviendo los problemas con la configuración obtenida se obtiene la siguiente comparación:

Tabla 19: Comparación Competencia PAM 2013

	F_1	Precision	Recall
Seidman	0.800	0.800	0.800
Layton	0.767	0.800	0.800
Jankowska	0.733	0.733	0.733
Vilariño	0.733	0.733	0.733
Halvani	0.700	0.700	0.700
Feng & Hirst	0.700	0.700	0.700
Ghaeini	0.691	0.760	0.633
DOCODE n=1	**0.683**	0.683	0.683
DOCODE NOR SEG n=2	**0.667**	0.667	0.667
Van Dam	0.600	0.600	0.600
DOCODE NOR SEG n=1	**0.600**	0.600	0.600
DOCODE NOR n=1	**0.558**	0.558	0.558
DOCODE n=2	**0.558**	0.558	0.558
DOCODE NOR n=2	**0.558**	0.558	0.558
Kern	0.533	0.533	0.533
Vartapetiance	0.500	0.500	0.500
Ledesma	0.467	0.467	0.467
Grozca	0.400	0.400	0.400

También destacar que si bien nuestro mejor resultado es de $F_1 = 0{,}683$ la configuración con la cual marcamos nuestra ejecución se debe a la primera combinación óptima encontrada y si bien pueden haber más (la configuración puede cambiar ya que varios resultados nos pueden dar el mismo F_1), este análisis es mucho más complicado y se debe realizar revisando las configuraciones de otras pruebas pero de ser considerado, se puede mencionar que con algunas modificaciones pudimos llegar rápidamente a otra combinación óptima con DOCODE NORMALIZADO y obtener un $F_1 = 0{,}700$. Lo que nos lleva a pensar que existen otras mejoras en este punto del algoritmo que si bien, nuestros resultados óptimos nos darán cuenta de los mejores resultados con el primer λ que disponga del mejor F_1, se debe considerar que aún es posible mejorar nuestras configuraciones.

4.4. Tiempo Computacional

El tiempo utilizado al pasar de unigrama ($n = 1$) a bigrama ($n = 2$) se calcula con el comando *time* que traen los sistemas unix, y del cual obtenemos el tiempo del sistema, tiempo de usuario y tiempo real transcurrido.

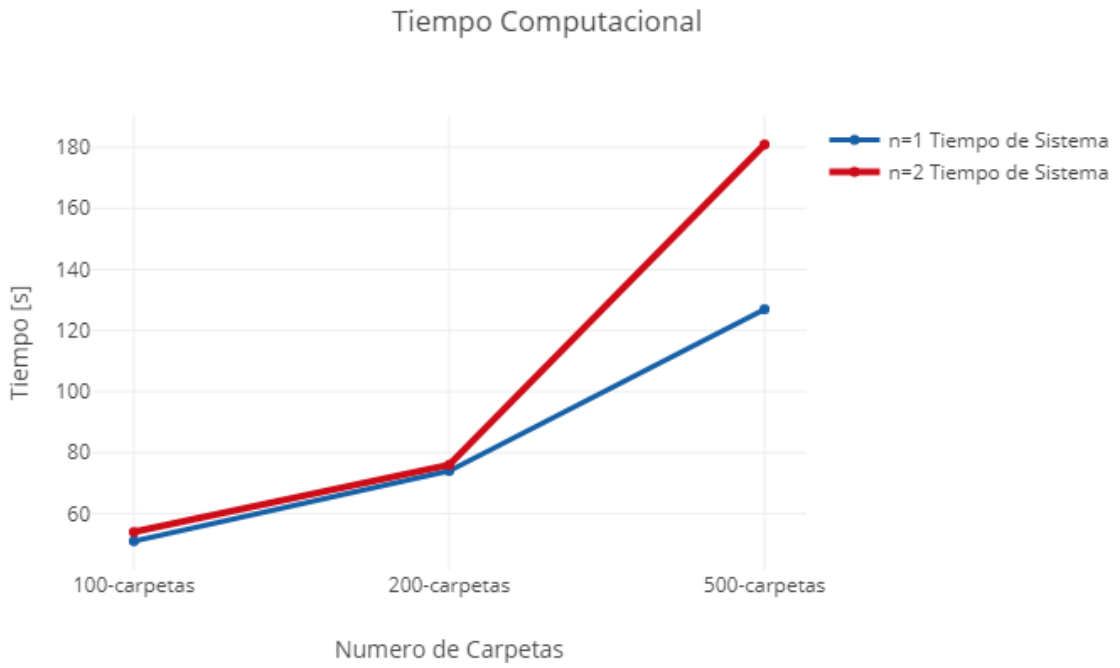


Figura 16: Gráfico de tiempo computacional (sys en segundos)

Tabla 20: Tiempos obtenidos con n=1 (unigrama) y n=2 (bigrama)

	100 Carpetas		200 Carpetas		500 Carpetas	
	n=1	n=2	n=1	n=2	n=1	n=2
real	31m58s	38m4s	40m53s	40m29s	185m22s	188m46s
user	25m36s	31m31s	32m28s	32m34s	173m45s	174m25s
sys	0m51s	0m54s	1m14s	1m16s	2m7s	3m1s

Como se observa la diferencia en tiempo es despreciable y debiese aumentar exponencialmente bajo esa proporción, por lo que no sería problema si es que se desease trabajar con n más grandes.

4.5. Configuración Final y Resultados

Analizando los resultados obtenidos en todos nuestros experimentos de la sección 4.2 se obtendrá una configuración *óptima* para cualquier prueba a realizar, la cual se validará probando dicha configuración con las carpetas del data set generado en **Nivel 1** 3.4 y observando sus resultados.

Para esta configuración se analizaron todos los λ y segmentos generados en cada una de las pruebas realizadas. Para cada algoritmo se resolvió utilizar el λ que hubiese tenido un mejor F_1, tanto en unigrama como en bigrama, esto es debido a que los valores obtenidos por λ no se repetían por experimento y el promedio de ellos no consideraba el factor de los valores obtenidos por los F_1. Para el caso de los segmentos limitantes, se analizaron por segmento y en conjunto de los F_1 obtenidos, dándole un valor por experimento y resolviendo dejar los que hayan tenido mayor cantidad de aciertos en cada uno de sus segmentos.

Finalmente la configuración obtenida es:

Tabla 21: DOCODE - Configuración Óptima

	DOCODE	
	n=1	**n=2**
lambda	0.4	0.3
segmento limitante	[0,1,1,0,0,1,1,1,1,0]	[0,0,0,0,1,1,1,1,1,0]

Tabla 22: DOCODE NORMALIZADO - Configuración Óptima

	DOCODE NORMALIZADO	
	n=1	**n=2**
lambda	0.6	0.4
segmento limitante	[0,0,0,0,0,1,1,1,0,1]	[0,1,1,1,1,1,1,1,1,0]

Tabla 23: DOCODE NORMALIZADO POR SEGMENTO - Configuración Óptima

	DOCODE NORMALIZADO POR SEGMENTO	
	n=1	**n=2**
lambda	0.7	0.7
segmento limitante	[0,0,0,0,0,1,1,1,0,0]	[0,1,1,1,1,1,1,1,1,0]

Y probada en la carpeta de **Nivel 1** se obtienen los siguientes resultados:

Tabla 24: Resultados Nivel 1 (100 autores)

Nivel 0	DOCODE			DOCODE NORM			DOCODE NOR SEG		
Test	lmda	aciertos	F_1	lmda	aciertos	F_1	lmda	aciertos	F_1
n=1	0.4	68	0,685	0.6	68	0,586	0.7	68	0,614
Resultados Con Mejor F_1	0.2	70	0,703	0.5	73	0,606	0.3	71	0,621
n=2	0.3	69	0,700	0.4	68	0,692	0.7	69	0,45
Resultados Con Mejor F_1	0.1	72	0.730	0.2	66	0.702	0.5	69	0.500

Analizando los resultados generados y comparándolo con el mejor F_1 obtenido bajo esta configuración, podemos decir que, aunque los resultados no generan los mejores F_1 si están cerca de estos. También se observa que los F_1 generados por bigrama son mejores que los generados por unigrama para todos los algoritmos.

CAPÍTULO 5
CONCLUSIONES

5.1. Conclusiones Generales

Bajo los resultados obtenidos podemos resolver que, los algoritmos utilizados por [Reyes, 2016] pudieron ser modificados para obtener buenos resultados dentro del problema de **Verificación de Autores**. Es más, se pudo demostrar que se puede trabajar sobre las configuraciones de los λ's entregados (óptimos para F_1) para que generen mejores resultados, ya que en comparación con los F_1 obtenidos de los $\lambda = 0{,}075, 0{,}2, 0{,}8$ para los algoritmos de DOCODE, DOCODE NORMALIZADO y DOCODE NORMALIZADO POR SEGMENTO estos mejoraron en un 87.5 % de las veces[9]. También se puede observar que nuestro peor resultado de F_1 esta en DOCODE NORMALIZADO POR SEGMENTO con bigrama y el mejor es con DOCODE NORMALIZADO con bigrama. Veremos en este ámbito que el paso de unigrama a bigrama funciona mejor para DOCODE NORMALIZADO con un 75 % de mejora en sus resultados de F_1, un 50 % para DOCODE y sólo un 37.5 % para DOCODE NORMALIZADO POR SEGMENTO, lo cual nos dice que para este primer algoritmo si se sugeriría utilizar bigramas en su configuración mientras que en el último, no se aconsejaría. Para el caso de DOCODE, se ve una relación en el algoritmo con λ's pequeños que mejoran al pasar a bigrama, lo cual no ocurre al tener λ's grandes (mayores a 1-1.5) lo cual se podría atribuir a data set de *training* pequeños en relación a los data set de *test*.

En cuanto a los aciertos obtenidos, estos tienen un promedio del 62 % de aciertos por experimento, con un mínimo de un 53 % de acierto (obtenido por los 3 algoritmos con unigrama) y un máximo de 75 % de aciertos obtenido con DOCODE NORMALIZADO con unigrama. Si nos enfocamos en los resultados de unigramas, obtenemos, al igual que [Reyes, 2016], que los mejores resultados de F_1 son entregados por DOCODE NORMALIZADO POR SEGMENTO, mientras que los mejores aciertos son realizados por DOCODE NORMALIZADO. Para bigrama los mejores aciertos son realizados por el algoritmo de DOCODE.

Cabe destacar que si bien se han realizado pruebas sobre una configuración óptima según F_1 que pueda ser utilizada para cualquier data set, la poca homogeneidad de sus resultados tanto en el λ y segmento limitante, hacen suponer que data sets que sean armados con alguna particularidad especifica, como por ejemplo que sus data set de *training* sean mejores que los data set de *test* o que la cantidad promedio de palabras en los data set de *training* sean menores a los data set de *test*, impactaría en los resultados obtenidos.

Finalmente podemos concluir que este aporte podrá ser utilizado en la resolución de este problema en específico, pero se debe entender que cada pequeña contribución en cada uno de estos problemas de autoría es una muestra de que, es posible caracterizar un estilo

[9] Porcentaje obtenido de los resultados entregados en los experimentos de la sección de experimentos 4.2

de escritura y, sí, podemos determinar al autor de cada documento. Y es que, el problema de autoría es complejo pero se debe tener presente que, no analizamos una personalidad cambiante como lo hacen la psicología si no que lo hacemos en un momento dado, en un punto fijo, el documento será una fiel imagen de la persona en ese preciso momento, de su estilo de escritura, y eso hace que el problema sea posible.

5.2. Cumplimiento de Objetivos

En la sección 1.3 se definieron los objetivos específicos que se buscaban cumplir con este trabajo. A continuación se detalla sobre el cumplimiento de cada uno de ellos.

- Formar una base de conocimiento de los problemas, métodos y técnicas que permitan detectar la autoría de documentos desconocidos.

  Este objetivo se expone en la sección 1.3 donde se presentan varios técnicas utilizadas para el problema general de *Authorship Attribution*. En esta sección se analizan las bases de detección de autoría como también sus métodos y aplicaciones del problema.

- Adaptar los algoritmos expuestos por [Reyes, 2016] para el problema de verificación de autores.

  En el sección 3.1 de análisis de algoritmo se plantean los algoritmos utilizados por [Reyes, 2016] y los cambios realizados para su utilización en el problema de **Verificación de Autores**.

- Definir experimentos y métricas que nos permitan cuantificar los resultados obtenidos para comparar el uso de bigrama y unigrama con los mismos data sets.

  Este objetivo se presenta en la sección 3.4 donde se describe el data set a utilizar en algunos experimentos. En esta sección se expone un detalle de su construcción y sus componentes.

- Definir métricas y experimentos que permitan analizar el uso de n-gramas.

  En la sección 4.2 se presenta una serie de experimentos realizados sobre unigrama y bigrama, sus tiempos computacionales y los resultados obtenidos. Información por la cual se pudo comparar el uso de n-grama en el problema y se obtuvieron conclusiones al respecto.

5.3. Trabajo Futuro

Como se comentó en las conclusiones la configuración óptima para F_1 podría tener relación con ciertas particularidades de los data set a trabajar, uno de los puntos que se propone trabajar es entender la relación que mantienen los data set dentro de su configuración (promedio de palabras, caracteres, vocabulario utilizado, idioma,etc.) con el estilo de escritura otorgado por los algoritmos. Esto también se puede plantear como la relación en la configuración y los problemas a desarrollar, experimentar con la configuración otorgada por [Reyes, 2016] en el problema de **Detección de Plagio Intrínseco** o la generada en la sección 4.5 para el problema de **Validación de Autores**, considerando ventanas m de largo variable para ambos algoritmos y su relación con dicho problema, *¿la configuración lograda es producto de los data set utilizados o tiene relación con el problema a resolver?*.

En este punto se podría estudiar el análisis de largo del texto en los *data-set* de cada problema. La afirmación de que, poco texto puede tener relación con malos resultados, podría estudiarse utilizando una ventana deslizante en el texto conocido, con la cual se genere una mayor cantidad de segmentos y permita obtener mejores resultados. Lo cual podría confundirse con la idea comentada anteriormente en la cual se propone analizar el largo de la ventana m en los algoritmos, pero no lo es, la idea anteriormente descrita se basa en la diferencia que tiene un m fijo y la cantidad de segmentos generados obtenidos del total de palabras dividido el valor de m elegido, como lo hemos tratado en los algoritmos. En cambio con la propuesta de considerar una ventana no-fija si no que se vaya deslizando palabra por palabra, se podrá ir generando más segmentos del mismo documento ya que, en vez de tener un documento segmentado desde la palabra inicial w_1 hasta la palabra 400 del documento (w_{400}), de la w_{400} a w_{800} (lo cual corresponde inmediatamente al segundo segmento de largo 400) hasta finalizar el documento, tendríamos el nuevo escenario que consideraría desde la palabra inicial w_1 hasta la palabra 400 del documento (w_{400}), desde la segunda palabra de nuestro documento (w_2) hasta la palabra (401) y así sucesivamente hasta terminar el documento. Con lo que se espera que este mayor número de segmentos generados ayuden a obtener mejores comparaciones y por ende, mejores resultados.

En cuanto a la utilización de n-grama se espera que sus resultados se comporten como una curva donde el mejor n nos dará un mejor F_1, un punto interesante sería experimentar con diferente n por algoritmo y por problema, se esperaría que cada uno diera un n óptimo que no dependería de los data set si no más bien del algoritmo y tal vez, de alguna medida, del problema a resolver.

REFERENCIAS BIBLIOGRÁFICAS

[AIC, 2017] (2017). Aicbt. http://www.aicbt.com/authorship-attribution/. Accedido: 2017-11-17.

[BAE, 2017] (2017). Beaper papers. http://www.unmuseum.org/bealepap.htm. Accedido: 2017-11-17.

[GUT, 2017] (2017). Gutenberg. https://www.gutenberg.org/. Accedido: 2017-11-17.

[GIT, 2017] (2017). Jgaap. https://github.com/evllabs/JGAAP. Accedido: 2017-11-17.

[PAN, 2017] (2017). Pan. http://pan.webis.de/. Accedido: 2017-11-17.

[PLA, 2017] (2017). Plagiarisma. http://plagiarisma.net/es/. Accedido: 2017-11-17.

[PS-, 2017] (2017). Plagscam. https://www.plagscan.com. Accedido: 2017-11-17.

[CLE, 2018] (2018). Clef. http://clef2018.clef-initiative.eu/. Accedido: 2018-04-28.

[Akhabue y Lautenbach, 2010] Akhabue, E. y Lautenbach, E. (2010). "equal" contributions and credit: an emerging trend in the characterization of authorship. *Annals of epidemiology*, 20(11):868–871.

[Alzahrani *et al.*, 2012] Alzahrani, S. M., Salim, N., y Abraham, A. (2012). Understanding plagiarism linguistic patterns, textual features, and detection methods. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 42(2):133–149.

[Argamon *et al.*, 2007] Argamon, S., Koppel, M., Pennebaker, J. W., y Schler, J. (2007). Mining the blogosphere: Age, gender and the varieties of self-expression. *First Monday*, 12(9).

[Baayen *et al.*, 2002] Baayen, H., van Halteren, H., Neijt, A., y Tweedie, F. (2002). An experiment in authorship attribution. En *6th JADT*, pp. 29–37.

[Bailey, 1979] Bailey, R. W. (1979). Authorship attribution in a forensic setting.

[Barrón Cedeño, 2008] Barrón Cedeño, L. A. (2008). Detección automática de plagio en texto. Tesis de m□ster, Universidad Politécnica de Valencia, Valencia, España. Tesis desarrollada dentro del Máster en Inteligencia Artificial, Reconocimiento de Formas e Imagen Digital.

[Chaski, 2005] Chaski, C. E. (2005). Who's at the keyboard? authorship attribution in digital evidence investigations.

[Diederich *et al.*, 2003] Diederich, J., Kindermann, J., Leopold, E., y Paass, G. (2003). Authorship attribution with support vector machines. *Applied intelligence*, 19(1-2):109–123.

[Funez y Errecalde, 2011] Funez, D. G. y Errecalde, M. L. (2011). Detección de plagio intrínseco usando la segmentación de texto.

[Gallardo, 2013] Gallardo, G. I. L. O. (2013). Diseño e implementación de una técnica para la detección intrínseca de plagio en documentos digitales. Tesis de m?ster, Universidad de Chile, Santiago, Chile. Tesis para optar al grado de magíster en gestión de operaciones.

[Ginsburg, 2002] Ginsburg, J. C. (2002). The concept of authorship in comparative copyright law. *DePaul L. Rev.*, 52:1063.

[Hammer, 1988] Hammer, C. (1988). Second order homophonic ciphers. *Cryptologia*, 12(1):11–20.

[Holmes, 1992] Holmes, D. I. (1992). A stylometric analysis of mormon scripture and related texts. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, pp. 91–120.

[Holmes, 1994] Holmes, D. I. (1994). Authorship attribution. *Computers and the Humanities*, 28(2):87–106.

[Hoover, 2001] Hoover, D. L. (2001). Statistical stylistics and authorship attribution: an empirical investigation. *Literary and linguistic computing*, 16(4):421–444.

[Howard, 2008] Howard, B. S. (2008). Authorship attribution under the rules of evidence: empirical approaches–a layperson's legal system. *International Journal of Speech, Language & the Law*, 15(2).

[Kešelj *et al.*, 2003] Kešelj, V., Peng, F., Cercone, N., y Thomas, C. (2003). N-gram-based author profiles for authorship attribution. En *Proceedings of the conference pacific association for computational linguistics, PACLING*, volumen 3, pp. 255–264.

[Koppel y Schler, 2004] Koppel, M. y Schler, J. (2004). Authorship verification as a one-class classification problem. En *Proceedings of the twenty-first international conference on Machine learning*, p. 62. ACM.

[Koppel *et al.*, 2011] Koppel, M., Schler, J., y Argamon, S. (2011). Authorship attribution in the wild. *Language Resources and Evaluation*, 45(1):83–94.

[Kruh, 1982] Kruh, L. (1982). A basic probe of the beale cipher as a bamboozlement. *Cryptologia*, 6(4):378–382.

[Li *et al.*, 2006] Li, J., Zheng, R., y Chen, H. (2006). From fingerprint to writeprint. *Communications of the ACM*, 49(4):76–82.

[Luyckx y Daelemans, 2008] Luyckx, K. y Daelemans, W. (2008). Authorship attribution and verification with many authors and limited data. En *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1*, pp. 513–520. Association for Computational Linguistics.

[Mendenhall, 1887] Mendenhall, T. C. (1887). The characteristic curves of composition. *Science*, 9(214):237–249.

[Morton, 1965] Morton, A. Q. (1965). The authorship of greek prose. *Journal of the Royal Statistical Society. Series A (General)*, 128(2):169–233.

[Mosteller y Wallace, 1964] Mosteller, F. y Wallace, D. (1964). Inference and disputed authorship: The federalist.

[Nickell, 1982] Nickell, J. (1982). Discovered: The secret of beale's treasure. *The Virginia Magazine of History and Biography*, 90(3):310–324.

[Peng y Schuurmans, 2003] Peng, F. y Schuurmans, D. (2003). Combining naive bayes and n-gram language models for text classification. En *European Conference on Information Retrieval*, pp. 335–350. Springer.

[Peng *et al.*, 2004] Peng, F., Schuurmans, D., y Wang, S. (2004). Augmenting naive bayes classifiers with statistical language models. *Information Retrieval*, 7(3-4):317–345.

[Peng y Hengartner, 2002] Peng, R. D. y Hengartner, N. W. (2002). Quantitative analysis of literary styles. *The American Statistician*, 56(3):175–185.

[Rangel Pardo *et al.*, 2017] Rangel Pardo, F., Rosso, P., Potthast, M., y Stein, B. (2017). Overview of the 5th Author Profiling Task at PAN 2017: Gender and Language Variety Identification in Twitter. En Cappellato, L., Ferro, N., Goeuriot, L., y Mandl, T., editores, *Working Notes Papers of the CLEF 2017 Evaluation Labs*, volumen 1866 de *CEUR Workshop Proceedings*. CLEF and CEUR-WS.org.

[Reyes, 2016] Reyes, P. L. (2016). *Análisis y propuesta de mejoras a algoritmos de detección intrínseca de plagio*. Valparaíso, Chile. Memoria de titulación para optar al título de ingeniero civil en informática.

[Rudman, 1997] Rudman, J. (1997). The state of authorship attribution studies: Some problems and solutions. *Computers and the Humanities*, 31(4):351–365.

[Sebastiani, 2005] Sebastiani, F. (2005). Text categorization. En *Encyclopedia of Database Technologies and Applications*, pp. 683–687. IGI Global.

[Seidman, 2013] Seidman, S. (2013). Authorship verification using the impostors method. En *CLEF 2013 Evaluation Labs and Workshop-Online Working Notes*. Citeseer.

[Sering *et al.*, 2018] Sering, K., Milin, P., y Baayen, R. H. (2018). Language comprehension as a multi-label classification problem. *Statistica Neerlandica*.

[Sidorov *et al.*, 2013] Sidorov, G., Velasquez, F., Stamatatos, E., Gelbukh, A., y Chanona-Hernández, L. (2013). Syntactic dependency-based n-grams: More evidence of usefulness in classification. En *International Conference on Intelligent Text Processing and Computational Linguistics*, pp. 13–24. Springer.

[Stamatatos, 2009a] Stamatatos, E. (2009a). Intrinsic plagiarism detection using character n-gram profiles. *CEUR-WS*, 502(8):1613–0073.

[Stamatatos, 2009b] Stamatatos, E. (2009b). A survey of modern authorship attribution methods. *Journal of the Association for Information Science and Technology*, 60(3):538–556.

[Stamatatos *et al.*, 2000] Stamatatos, E., Fakotakis, N., y Kokkinakis, G. (2000). Automatic text categorization in terms of genre and author. *Computational linguistics*, 26(4):471–495.

[Stamatatos *et al.*, 2001] Stamatatos, E., Fakotakis, N., y Kokkinakis, G. (2001). Computer-based authorship attribution without lexical measures. *Computers and the Humanities*, 35(2):193–214.

[Taylor *et al.*, 2008] Taylor, Q. C., Stevenson, J. E., Delorey, D. P., y Knutson, C. D. (2008). Author entropy: A metric for characterization of software authorship patterns. En *Third International Workshop on Public Data about Software Development (WoPDaSD08)*, p. 6.

[Williams, 1940] Williams, C. B. (1940). A note on the statistical analysis of sentence-length as a criterion of literary style. *Biometrika*, 31(3/4):356–361.

[Zheng *et al.*, 2003] Zheng, R., Qin, Y., Huang, Z., y Chen, H. (2003). Authorship analysis in cybercrime investigation. En *International Conference on Intelligence and Security Informatics*, pp. 59–73. Springer.

[Zu Eissen y Stein, 2006] Zu Eissen, S. M. y Stein, B. (2006). Intrinsic plagiarism detection. En *European Conference on Information Retrieval*, pp. 565–569. Springer.

ANEXOS

Tabla 25: Lista Autores y descripción

Autores	Nacionalidad	Años en vida	Idioma
Mark Twain	EEUU	1835-1910	Inglés
Jane Austen	Reino Unido	1775-1817	Inglés
Charles Dickens	Reino Unido	1812-1870	Inglés
Leon Toltoi	Rusia	1828-1910	Ruso
Edgar Alan Poe	EEUU	1809-1849	Inglés
Virginia Woolf	Reino Unido	1882-1941	Inglés
Homero	Jonia	VIII a.C.	Griego
Arthur Conan Doyle	Reino Unido	1859-1930	Inglés
Platon	Atenas	427-347 a.C.	Griego
F.Nietzsche	Alemania	1844-1900	Alemán
Shakespear	Reino Unido	1564-1616	Inglés
Fyodor Dostoyevsky	Rusia	1821-1881	Ruso
Charlotte Brontë	Reino Unido	1816-1855	Inglés
Lewis Carroll	Reino Unido	1832-1898	Inglés
Alexandre Dumas	Francia	1802-1870	Francés
Gustave Flaubert	Francia	1821-1880	Francés
Oscar Wilde	Irlanda	1854-1900	Inglés
Eleanor Hallowell Abbott	EEUU	1872-1958	Inglés
L. Frank (Lyman Frank) Baum	EEUU	1856-1919	Inglés
Aldous Huxley	Reino Unido	1894-1963	Inglés
H.L.Sayler	EEUU	1863-1913	Inglés
Felix Dahn	Alemania	1834-1912	Alemán
Dante Alighieri	Italia	1265-132	Italiano
Richard Harding Davis	EEUU	1864-1916	Inglés
Philip K.Dick	EEUU	1928-1982	Inglés

Tabla 26: Nivel 0 - Training

Carpeta	Archivo	Autor	Libro	Año	Cant Palabras	Cant Caracteres	% DeTC[10]
EN01	known	Mark Twain	Adventures of Huckleberry Finn	1884	107434	566646	
	unknown	Mark Twain	Adventures of Huckleberry Finn		9509	49662	8,9
EN02	known	Mark Twain	The Adventures of Tom Sawyer	1876	68110	382996	
	unknown	Mark Twain	The Adventures of Tom Sawyer		8032	45106	11,8
EN03	known	Mark Twain	The Prince and The Pauper	1881	63086	361679	
	unknown	Mark Twain	The Prince and The Pauper		12093	70011	19,2
EN04	known	Mark Twain	Life On The Mississippi	1883	110803	626996	
	unknown	Mark Twain	Life On The Mississippi		39077	216285	35,3
EN05	known	Jane Austen	Pride and Prejudice	1813	113692	647317	
	unknown	Jane Austen	Pride and Prejudice		13677	77408	12,0
EN06	known	Jane Austen	Sense and Sensibility	1811	109662	605908	
	unknown	Jane Austen	Sense and Sensibility		17875	100244	16,3
EN07	known	Jane Austen	Persuasion	1818	50432	275713	
	unknown	Jane Austen	Persuasion		39725	219309	78,8
EN08	known	Jane Austen	Mansfield Park	1814	139770	779662	
	unknown	Jane Austen	Mansfield Park		25142	138714	18,0
EN09	known	Charles Dickens	Oliver Twist	1838	148120	817330	
	unknown	Charles Dickens	Oliver Twist		21746	118951	14,7
EN10	known	Charles Dickens	The Life And Adventures Of Nicholas Nickleby	1839	315507	1850337	
	unknown	Charles Dickens	The Life And Adventures Of Nicholas Nickleby		17798	100923	5,6
EN11	known	Charles Dickens	The Old Curiosity Shop	1840	228254	1282845	
	unknown	Charles Dickens	The Old Curiosity Shop		22569	126851	9,9
EN12	known	Charles Dickens	A Tale of Two Cities	1859	106905	594270	
	unknown	Charles Dickens	A Tale of Two Cities		36335	198715	34,0
EN13	known	Leon Toltoi	War and Peace	1865	491035	2830754	
	unknown	Leon Toltoi	War and Peace		92042	528796	18,7
EN14	known	Leon Toltoi	Anna Karenina	1877	330294	1847281	
	unknown	Leon Toltoi	Anna Karenina		31930	179325	9,7
EN15	known	Leon Toltoi	The Kreutzer Sonata and Other Stories	1890	65007	362227	
	unknown	Leon Toltoi	The Kreutzer Sonata and Other Stories		3857	21920	5,9

[10]Porcentaje de texto desconocido en comparación al texto conocido

EN16	known	Leon Toltoi	Childhood	1852	38091	211266	
	unknown	Leon Toltoi	Childhood		4506	24685	11,8
EN17	known	Edgar Alan Poe	The Works of Edgar Allan Poe Vol 2	1840	93539	540916	
	unknown	Edgar Alan Poe	The Works of Edgar Allan Poe Vol 2		7137	42130	7,6
EN18	known	Edgar Alan Poe	The Works of Edgar Allan Poe Vol 3	1842	104042	598631	
	unknown	Edgar Alan Poe	The Works of Edgar Allan Poe Vol 3		49082	296881	47,2
EN19	known	Edgar Alan Poe	The Works of Edgar Allan Poe Vol 4	1844	91406	537725	
	unknown	Edgar Alan Poe	The Works of Edgar Allan Poe Vol 4		18709	105966	20,5
EN20	known	Edgar Alan Poe	Eureka	1848	42443	255120	
	unknown	Edgar Alan Poe	Eureka		6640	41761	15,6
EN21	known	Virginia Woolf	Night and Day	1919	150073	869564	
	unknown	Virginia Woolf	Night and Day		22613	131778	15,1
EN22	known	Virginia Woolf	Jacob's Room	1922	51456	296704	
	unknown	Virginia Woolf	Jacob's Room		7663	43388	14,9
EN23	known	Virginia Woolf	The Voyage Out	1915	131896	738889	
	unknown	Virginia Woolf	The Voyage Out		10770	60925	8,2
EN24	known	Virginia Woolf	Monday or Tuesday	1921	21961	125059	
	unknown	Virginia Woolf	Monday or Tuesday		5216	29025	23,8
EN25	known	Homero	The Iliad		224743	1079874	
	unknown	Homero	The Iliad		27623	134024	12,3
EN51	known	Charlotte Brontë	Shirley	1811	228913	1274038	
	unknown	Charlotte Brontë	Shirley		33389	189546	14,6
EN52	known	Charlotte Brontë	The Professor	1857	78743	451920	
	unknown	Charlotte Brontë	The Professor		14923	84753	19,0
EN53	known	Lewis Carroll	Alice's Adventures in Wonderland	1865	28467	157169	
	unknown	Lewis Carroll	Alice's Adventures in Wonderland		3243	16426	11,4
EN54	known	Lewis Carroll	Through the Looking-Glass	1871	28949	155488	
	unknown	Lewis Carroll	Through the Looking-Glass		7057	38119	24,4
EN55	known	Lewis Carroll	Tangled Tale	1885	35205	183983	
	unknown	Lewis Carroll	Tangled Tale		7883	45024	22,4
EN56	known	Lewis Carroll	Sylvie and Bruno	1889	54252	300845	
	unknown	Lewis Carroll	Sylvie and Bruno		17845	102128	32,9
EN57	known	Alexandre Dumas	The Count of Monte Cristo	1844	441631	2558015	
	unknown	Alexandre Dumas	The Count of Monte Cristo		38931	225684	8,8

EN58	known	Alexandre Dumas	The Three Musketeers	1844	204835	1172672	
	unknown	Alexandre Dumas	The Three Musketeers		36897	213158	18,0
EN59	known	Alexandre Dumas	The Man in the Iron Mask	1857	161941	945223	
	unknown	Alexandre Dumas	The Man in the Iron Mask		20692	120912	12,8
EN60	known	Alexandre Dumas	Camille (La Dame aux Camilias)	1848	61871	328692	
	unknown	Alexandre Dumas	Camille (La Dame aux Camilias)		10087	54280	16,3
EN61	known	Gustave Flaubert	Madame Bovary	1856	102277	583780	
	unknown	Gustave Flaubert	Madame Bovary		19755	110892	19,3
EN62	known	Gustave Flaubert	A Simple Soul	1877	11503	66173	
	unknown	Gustave Flaubert	A Simple Soul		4159	23845	36,2
EN63	known	Gustave Flaubert	Salammbô	1862	60617	352479	
	unknown	Gustave Flaubert	Salammbô		50358	293163	83,1
EN64	known	Gustave Flaubert	The Temptation	1874	56269	331620	
	unknown	Gustave Flaubert	The Temptation		16664	97711	29,6
EN65	known	Oscar Wilde	The Happy Prince and Other Tales	1888	17950	93793	
	unknown	Oscar Wilde	The Happy Prince and Other Tales		3300	17347	18,4
EN66	known	Oscar Wilde	The Picture of Dorian Gray	1890	85387	465976	
	unknown	Oscar Wilde	The Picture of Dorian Gray		14909	81650	17,5
EN67	known	Oscar Wilde	The Canterville Ghost	1887	15282	86711	
	unknown	Oscar Wilde	The Canterville Ghost		3046	17402	19,9
EN68	known	Oscar Wilde	For Love of the King	1926	5854	33648	
	unknown	Oscar Wilde	For Love of the King		1933	11274	33,0
EN69	known	Eleanor Hallowell Abbott	The Stingy Receiver	1917	23850	136996	
	unknown	Eleanor Hallowell Abbott	The Stingy Receiver		7749	44584	32,5
EN70	known	Eleanor Hallowell Abbott	Old-Dad	1923	28233	164620	
	unknown	Eleanor Hallowell Abbott	Old-Dad		17107	100011	60,6
EN71	known	Eleanor Hallowell Abbott	Rainy Week	1921	44887	263207	
	unknown	Eleanor Hallowell Abbott	Rainy Week		10533	61672	23,5
EN72	known	Eleanor Hallowell Abbott	The Sick-a-Bed Lady	1911	80529	457784	
	unknown	Eleanor Hallowell Abbott	The Sick-a-Bed Lady		16050	90094	19,9
EN73	known	L. Frank Baum	The Land of Oz	1900	36311	198114	
	unknown	L. Frank Baum	The Land of Oz		14469	79031	39,8
EN74	known	L. Frank Baum	The Magic of Oz	1919	40150	200764	
	unknown	L. Frank Baum	The Magic of Oz		12717	65049	31,7

EN75	known	L. Frank Baum	The Scarecrow of Oz	1915	44084	224114	
	unknown	L. Frank Baum	The Scarecrow of Oz		11162	59667	25,3
EN9901	known	Mark Twain	Adventures of Huckleberry Finn	1884	116949	616320	
	unknown	Philip K.Dick	The Crystal Crypt	1954	1300	7218	1,1
EN9902	known	Mark Twain	The Adventures of Tom Sawyer	1876	76142	428104	
	unknown	Philip K.Dick	The Defenders	1953	1903	10114	2,5
EN9903	known	Mark Twain	The Prince and The Pauper	1881	75179	431692	
	unknown	Philip K.Dick	Beyond Lies the Wub	1952	1318	6770	1,8
EN9904	known	Mark Twain	Life On The Mississippi	1883	149879	843281	
	unknown	Philip K.Dick	Beyond the Door	1954	1117	5865	0,7
EN9905	known	Jane Austen	Pride and Prejudice	1813	127368	724725	
	unknown	Richard Harding Davis	In the Fog	1897	3356	18922	2,6
EN9906	known	Jane Austen	Sense and Sensibility	1811	127536	706152	
	unknown	Richard Harding Davis	Captain Macklin	1902	14673	78431	11,5
EN9907	known	Jane Austen	Persuasion	1818	90156	495022	
	unknown	Richard Harding Davis	A Charmed Life	1891	2125	11728	2,4
EN9908	known	Jane Austen	Mansfield Park	1814	164911	918376	
	unknown	Richard Harding Davis	Cinderella	1891	6231	34309	3,8
EN9909	known	Charles Dickens	Oliver Twist	1838	169865	936281	
	unknown	Dante Alighieri	The New Life (La Vita Nuova)	1320	3194	15381	1,9
EN9910	known	Charles Dickens	The Life And Adventures Of Nicholas Nickleby	1839	333304	1951260	
	unknown	Dante Alighieri	The Vision of Hell	1320	4595	25179	1,4
EN9911	known	Charles Dickens	The Old Curiosity Shop	1840	228254	1282845	
	unknown	Dante Alighieri	The Divine Comedy of Dante Alighieri	1308	24030	136056	10,5
EN9912	known	Charles Dickens	A Tale of Two Cities	1859	143239	792985	
	unknown	Dante Alighieri	The Divine Comedy	1308	13670	76758	9,5
EN9913	known	Leon Toltoi	War and Peace	1865	583077	3359550	
	unknown	Felix Dahan	Felicitas	1892	10388	57627	1,8
EN9914	known	Leon Toltoi	Anna Karenina	1877	362223	2026606	
	unknown	Felix Dahan	A Struggle for Rome	1876	18319	101475	5,1
EN9915	known	Leon Toltoi	The Kreutzer Sonata and Other Stories	1890	68864	384147	
	unknown	Felix Dahan	The Scarlet Banner	1861	11287	62865	16,4
EN9916	known	Leon Toltoi	Childhood	1852	42596	235951	
	unknown	H.L.Sayler	The Stolen Aeroplane	1911	27446	156670	64,4

EN9917	known	Edgar Alan Poe	The Works of Edgar Allan Poe Vol 2	1840	100675	583046	
	unknown	H.L.Sayler	The Aeroplane Express	1911	6447	37155	6,4
EN9918	known	Edgar Alan Poe	The Works of Edgar Allan Poe Vol 3	1842	104042	598631	
	unknown	H.L.Sayler	The Airship Boys' Ocean Flyer	1909	3761	21121	3,6
EN9919	known	Edgar Alan Poe	The Works of Edgar Allan Poe Vol 4	1844	91406	537725	
	unknown	H.L.Sayler	In the Clouds for Uncle Sam	1910	9686	55717	10,6
EN9920	known	Edgar Alan Poe	Eureka	1848	49082	296881	
	unknown	Aldous Huxley	The Burning Wheel	1920	10470	60894	21,3
EN9921	known	Virginia Woolf	Night and Day	1919	172685	1001342	
	unknown	Aldous Huxley	Leda	1920	1559	5920	0,9
EN9922	known	Virginia Woolf	Jacob's Room	1922	59118	340092	
	unknown	Aldous Huxley	Limbo	1920	3992	18604	6,8
EN9923	known	Virginia Woolf	The Voyage Out	1915	142665	799814	
	unknown	Aldous Huxley	Crome Yellow	1921	9125	52624	6,4
EN9924	known	Virginia Woolf	Monday or Tuesday	1921	21961	125059	
	unknown	L. Frank Baum	Tik-Tok of Oz	1914	8932	51574	40,7
EN9925	known	Homero	The Iliad		252365	1213898	
	unknown	L. Frank Baum	The Scarecrow of Oz	1915	14217	75313	5,6
EN9951	known	Charlotte Brontë	Shirley	1811	228913	1274038	
	unknown	Fyodor Dostoyevsky	Poor Folk	1853	32322	176136	14,1
EN9952	known	Charlotte Brontë	The Professor	1857	93665	536673	
	unknown	Fyodor Dostoyevsky	Poor Folk	1844	32322	176136	34,5
EN9953	known	Lewis Carroll	Alice's Adventures in Wonderland	1865	31709	173595	
	unknown	Fyodor Dostoyevsky	Poor Folk	1844	32322	176136	101,9
EN9954	known	Lewis Carroll	Through the Looking-Glass	1871	36005	193607	
	unknown	Fyodor Dostoyevsky	Poor Folk	1844	32322	176136	89,8
EN9955	known	Lewis Carroll	Tangled Tale	1885	35205	183983	
	unknown	Fyodor Dostoyevsky	Poor Folk	1844	32322	176136	91,8
EN9956	known	Lewis Carroll	Sylvie and Bruno	1889	72096	402973	
	unknown	Fyodor Dostoyevsky	Poor Folk	1844	32322	176136	44,8
EN9957	known	Alexandre Dumas	The Count of Monte Cristo	1844	480561	2783699	
	unknown	Fyodor Dostoyevsky	Poor Folk	1844	32322	176136	6,7
EN9958	known	Alexandre Dumas	The Three Musketeers	1844	241731	1385830	
	unknown	Fyodor Dostoyevsky	Poor Folk	1844	32322	176136	13,4

EN9959	known	Alexandre Dumas	The Man in the Iron Mask	1857	182632	1066135	
	unknown	Fyodor Dostoyevsky	Poor Folk	1844	32322	176136	17,7
EN9960	known	Alexandre Dumas	Camille (La Dame aux Camilias)	1848	71957	382972	
	unknown	Fyodor Dostoyevsky	Poor Folk	1844	32322	176136	44,9
EN9961	known	Gustave Flaubert	Madame Bovary	1856	122031	694672	
	unknown	Fyodor Dostoyevsky	Poor Folk	1844	32322	176136	26,5
EN9962	known	Gustave Flaubert	A Simple Soul	1877	15661	90018	
	unknown	Fyodor Dostoyevsky	Poor Folk	1844	32322	176136	206,4
EN9963	known	Gustave Flaubert	Salammbô	1862	110974	645642	
	unknown	Fyodor Dostoyevsky	Poor Folk	1844	32322	176136	29,1
EN9964	known	Gustave Flaubert	The Temptation	1874	56269	331620	
	unknown	Fyodor Dostoyevsky	Poor Folk	1844	32322	176136	57,4
EN9965	known	Oscar Wilde	The Happy Prince and Other Tales	1888	21249	111140	
	unknown	Fyodor Dostoyevsky	Poor Folk	1844	32322	176136	152,1
EN9966	known	Oscar Wilde	The Picture of Dorian Gray	1890	85387	465976	
	unknown	Fyodor Dostoyevsky	Poor Folk	1844	32322	176136	37,9
EN9967	known	Oscar Wilde	The Canterville Ghost	1887	15282	86711	
	unknown	Fyodor Dostoyevsky	Poor Folk	1844	32322	176136	211,5
EN9968	known	Oscar Wilde	For Love of the King	1926	7786	44922	
	unknown	Fyodor Dostoyevsky	Poor Folk	1844	32322	176136	415,1
EN9969	known	Eleanor Hallowell Abbott	The Stingy Receiver	1917	31598	181580	
	unknown	Fyodor Dostoyevsky	Poor Folk	1844	32322	176136	102,3
EN9970	known	Eleanor Hallowell Abbott	Old-Dad	1923	45339	264631	
	unknown	Fyodor Dostoyevsky	Poor Folk	1844	32322	176136	71,3
EN9971	known	Eleanor Hallowell Abbott	Rainy Week	1921	44887	263207	
	unknown	Fyodor Dostoyevsky	Poor Folk	1844	32322	176136	72,0
EN9972	known	Eleanor Hallowell Abbott	The Sick-a-Bed Lady	1911	80529	457784	
	unknown	Fyodor Dostoyevsky	Poor Folk	1844	32322	176136	40,1
EN9973	known	L. Frank Baum	The Land of Oz	1900	50779	277145	
	unknown	Fyodor Dostoyevsky	Poor Folk	1844	32322	176136	63,7
EN9974	known	L. Frank Baum	The Magic of Oz	1919	52866	265813	
	unknown	Fyodor Dostoyevsky	Poor Folk	1844	32322	176136	61,1
EN9975	known	L. Frank Baum	The Scarecrow of Oz	1915	55245	283781	
	unknown	Fyodor Dostoyevsky	Poor Folk	1844	32322	176136	58,5

Tabla 27: Nivel 0 - Test

Carpeta	Archivo	Autor	Libro	Año	Cant Palabras	Cant Caracteres	% DeTC
EN26	known	Homero	The Odyssey		121081	639639	
	unknown	Homero	The Odyssey		13194	69781	10,9
EN27	known	Homero	The Adventures of Ulysses the Wanderer		23118	119952	
	unknown	Homero	The Adventures of Ulysses the Wanderer		6119	32433	26,5
EN28	known	Homero	The Hero of Ithaca		45519	239804	
	unknown	Homero	The Hero of Ithaca		8758	47258	19,2
EN29	known	Arthur Conan Doyle	The Adventures of Sherlock Holmes	1892	103024	553442	
	unknown	Arthur Conan Doyle	The Adventures of Sherlock Holmes		7659	41509	7,4
EN30	known	Arthur Conan Doyle	Memoirs of Sherlock Holmes	1894	77992	427951	
	unknown	Arthur Conan Doyle	Memoirs of Sherlock Holmes		15155	82650	19,4
EN31	known	Arthur Conan Doyle	The Return of Sherlock Holmes	1905	92093	508850	
	unknown	Arthur Conan Doyle	The Return of Sherlock Holmes		26181	144737	28,4
EN32	known	Arthur Conan Doyle	The Hound of the Baskervilles	1902	57847	315831	
	unknown	Arthur Conan Doyle	The Hound of the Baskervilles		8156	37055	14,1
EN33	known	Platon	The Republic		157821	860437	
	unknown	Platon	The Republic		66518	378672	42,1
EN34	known	Platon	Symposium		29732	164156	
	unknown	Platon	Symposium		6413	36696	21,6
EN35	known	Platon	Apology		16046	87302	
	unknown	Platon	Apology		3474	19602	21,7
EN36	known	Platon	Protagoras		24904	136438	
	unknown	Platon	Protagoras		6918	39643	27,8
EN37	known	F.Nietzsche	Beyond Good and Evil	1886	60052	362958	
	unknown	F.Nietzsche	Beyond Good and Evil		7437	45851	12,4
EN38	known	F.Nietzsche	Thus Spake Zarathustra	1896	101844	575316	
	unknown	F.Nietzsche	Thus Spake Zarathustra		19185	105448	18,8
EN39	known	F.Nietzsche	The Antichrist	1895	37900	226301	
	unknown	F.Nietzsche	The Antichrist		5749	34831	15,2
EN40	known	F.Nietzsche	Ecce Homo	1908	47091	242509	
	unknown	F.Nietzsche	Ecce Homo		10601	61613	22,5
EN41	known	Shakespear	The Complete Works of William Shakespeare	1600	1062792	4447975	
	unknown	Shakespear	The Complete Works of William Shakespeare		325949	1410817	30,7

EN42	known	Shakespear	Romeo and Juliet	1597	36575	144166		
	unknown	Shakespear	Romeo and Juliet		9058	34844	24,8	
EN43	known	Shakespear	Hamlet	1599	30376	169231		
	unknown	Shakespear	Hamlet		6446	35654	21,2	
EN44	known	Shakespear	The Tragedie of Macbeth	1623	19378	99742		
	unknown	Shakespear	The Tragedie of Macbeth		3765	20103	19,4	
EN45	known	Fyodor Dostoyevsky	The Brothers Karamazov	1879	315824	1779728		
	unknown	Fyodor Dostoyevsky	The Brothers Karamazov		46754	264463	14,8	
EN46	known	Fyodor Dostoyevsky	The Idiot	1868	215065	1225656		
	unknown	Fyodor Dostoyevsky	The Idiot		35305	199653	16,4	
EN47	known	Fyodor Dostoyevsky	The Grand Inquisitor	1879	13070	75448		
	unknown	Fyodor Dostoyevsky	The Grand Inquisitor		3373	19480	25,8	
EN48	known	Fyodor Dostoyevsky	Poor Folk	1844	43170	236471		
	unknown	Fyodor Dostoyevsky	Poor Folk		14819	81283	34,3	
EN49	known	Charlotte Brontë	Jane Eyre	1847	166133	894194		
	unknown	Charlotte Brontë	Jane Eyre		32322	176136	19,5	
EN50	known	Charlotte Brontë	Villette	1853	200872	1137607		
	unknown	Charlotte Brontë	Villette		27249	153306	13,6	
EN76	known	L. Frank (Lyman Frank) Baum	Tik-Tok of Oz	1914	42089	222954		
	unknown	L. Frank (Lyman Frank) Baum	Tik-Tok of Oz		14217	75313	33,8	
EN77	known	Aldous Huxley	Crome Yellow	1921	52625	306180		
	unknown	Aldous Huxley	Crome Yellow		8932	51574	17,0	
EN78	known	Aldous Huxley	Limbo	1920	54613	309961		
	unknown	Aldous Huxley	Limbo		9125	52624	16,7	
EN79	known	Aldous Huxley	Leda	1920	23146	94652		
	unknown	Aldous Huxley	Leda		3992	18604	17,2	
EN80	known	Aldous Huxley	The Burning Wheel	1920	10403	46356		
	unknown	Aldous Huxley	The Burning Wheel		1559	5920	15,0	
EN81	known	H.L.Sayler	In the Clouds for Uncle Sam	1910	35294	194010		
	unknown	H.L.Sayler	In the Clouds for Uncle Sam		10470	60894	29,7	
EN82	known	H.L.Sayler	The Airship Boys' Ocean Flyer	1909	63430	350746		
	unknown	H.L.Sayler	The Airship Boys' Ocean Flyer		9686	55717	15,3	
EN83	known	H.L.Sayler	The Aeroplane Express	1910	46994	260149		
	unknown	H.L.Sayler	The Aeroplane Express		3761	21121	8,0	

EN84	known	H.L.Sayler	The Stolen Aeroplane	1911	40509	224696	
	unknown	H.L.Sayler	The Stolen Aeroplane		6447	37155	15,9
EN85	known	Felix Dahn	The Scarlet Banner	1861	98432	541289	
	unknown	Felix Dahn	The Scarlet Banner		27446	156670	27,9
EN86	known	Felix Dahn	A Struggle for Rome v1	1876	111084	623201	
	unknown	Felix Dahn	A Struggle for Rome v1		11287	62865	10,2
EN87	known	Felix Dahn	A Struggle for Rome v2	1876	117371	652843	
	unknown	Felix Dahn	A Struggle for Rome v2		18319	101475	15,6
EN88	known	Felix Dahn	Felicitas	1892	43179	239791	
	unknown	Felix Dahn	Felicitas		10388	57627	24,1
EN89	known	Dante Alighieri	The Divine Comedy	1308	101469	564684	
	unknown	Dante Alighieri	The Divine Comedy		13670	76758	13,5
EN90	known	Dante Alighieri	The Divine Comedy of Dante Alighieri	1308	150098	715517	
	unknown	Dante Alighieri	The Divine Comedy of Dante Alighieri		24030	136056	16,0
EN91	known	Dante Alighieri	The Vision of Hell	1320	36545	203323	
	unknown	Dante Alighieri	The Vision of Hell		4595	25179	12,6
EN92	known	Dante Alighieri	The New Life (La Vita Nuova)	1320	30652	147920	
	unknown	Dante Alighieri	The New Life (La Vita Nuova)		3194	15381	10,4
EN93	known	Richard Harding Davis	Cinderella	1891	45236	247090	
	unknown	Richard Harding Davis	Cinderella		6231	34309	13,8
EN94	known	Richard Harding Davis	A Charmed Life	1891	6090	35212	
	unknown	Richard Harding Davis	A Charmed Life		2125	11728	34,9
EN95	known	Richard Harding Davis	Captain Macklin	1902	65538	358938	
	unknown	Richard Harding Davis	Captain Macklin		14673	78431	22,4
EN96	known	Richard Harding Davis	In the Fog	1897	22712	124808	
	unknown	Richard Harding Davis	In the Fog		3356	18922	14,8
EN97	known	Philip K.Dick	Beyond the Door	1954	4920	28006	
	unknown	Philip K.Dick	Beyond the Door		1117	5865	22,7
EN98	known	Philip K.Dick	Beyond Lies the Wub	1952	5089	28997	
	unknown	Philip K.Dick	Beyond Lies the Wub		1318	6770	25,9
EN99	known	Philip K.Dick	The Defenders	1953	11113	60596	
	unknown	Philip K.Dick	The Defenders		1903	10114	17,1
EN100	known	Philip K.Dick	The Crystal Crypt	1954	10056	55982	
	unknown	Philip K.Dick	The Crystal Crypt		1300	7218	12,9

EN9926	known	Homero	The Odyssey		134274	709420	
	unknown		The Magic of Oz	1919	11162	59667	8,3
EN9927	known	Homero	The Adventures of Ulysses the Wanderer		252365	1213898	
	unknown		The Land of Oz	1919	12717	65049	5,0
EN9928	known	Homero	The Hero of Ithaca		252365	1213898	
	unknown		The Sick-a-Bed Lady	1911	14469	79031	5,7
EN9929	known	Arthur Conan Doyle	The Adventures of Sherlock Holmes	1892	110682	594951	
	unknown		Rainy Week	1921	16050	90094	14,5
EN9930	known	Arthur Conan Doyle	Memoirs of Sherlock Holmes	1894	93146	510601	
	unknown		Old-Dad	1923	10533	61672	11,3
EN9931	known	Arthur Conan Doyle	The Return of Sherlock Holmes	1905	118273	653587	
	unknown		The Stingy Receiver	1917	17107	100011	14,5
EN9932	known	Arthur Conan Doyle	The Hound of the Baskervilles	1902	66002	352886	
	unknown		For Love of the King	1926	7749	44584	11,7
EN9933	known	Platon	The Republic		224338	1239109	
	unknown		The Canterville Ghost	1887	1933	11274	0,9
EN9934	known	Platon	Symposium		36144	200852	
	unknown		The Picture of Dorian Gray	1890	3046	17402	8,4
EN9935	known	Platon	Apology		19520	106904	
	unknown		The Happy Prince and Other Tales	1888	14909	81650	76,4
EN9936	known	Platon	Protagoras		31821	176081	
	unknown		The Temptation	1874	3300	17347	10,4
EN9937	known	F.Nietzsche	Beyond Good and Evil	1886	67488	408809	
	unknown				16664	97711	24,7
EN9938	known	F.Nietzsche	Thus Spake Zarathustra	1896	121028	680764	
	unknown		Salammbô	1862	50358	293163	41,6
EN9939	known	F.Nietzsche	The Antichrist	1895	37900	226301	
	unknown		A Simple Soul	1877	4159	23845	11,0
EN9940	known	F.Nietzsche	Ecce Homo	1908	57691	304122	
	unknown		Madame Bovary	1856	19755	110892	34,2
EN9941	known	Shakespear	The Complete Works of William Shakespeare	1600	1388740	5858792	
	unknown		Camille (La Dame aux Camilias)	1848	10087	54280	0,7
EN9942	known	Shakespear	Romeo and Juliet	1597	45632	179010	
	unknown		The Man in the Iron Mask	1857	20692	120912	45,3

EN9943	known	Shakespear	Hamlet	1599	36821	204885	
	unknown		The Three Musketeers	1844	36897	213158	100,2
EN9944	known	Shakespear	The Tragedie of Macbeth	1623	23142	119845	
	unknown		The Count of Monte Cristo	1844	38931	225684	168,2
EN9945	known	Fyodor Dostoyevsky	The Brothers Karamazov	1879	362578	2044191	
	unknown		Sylvie and Bruno	1889	17845	102128	4,9
EN9946	known	Fyodor Dostoyevsky	The Idiot	1868	250369	1425309	
	unknown		Tangled Tale	1885	7883	45024	3,1
EN9947	known	Fyodor Dostoyevsky	The Grand Inquisitor	1879	13070	75448	
	unknown		Through the Looking-Glass	1871	7057	38119	54,0
EN9948	known	Fyodor Dostoyevsky	Poor Folk	1844	57988	317754	
	unknown		The Professor	1865	14923	84753	25,7
EN9949	known	Charlotte Brontë	Jane Eyre	1847	198454	1070330	
	unknown	Fyodor Dostoyevsky	Poor Folk	1857	33389	189546	16,8
EN9950	known	Charlotte Brontë	Villette	1853	200872	1137607	
	unknown	Fyodor Dostoyevsky	Poor Folk	1811	32322	176136	16,1
EN9976	known	L. Frank (Lyman Frank) Baum	Tik-Tok of Oz	1914	56305	298267	
	unknown	Fyodor Dostoyevsky	Poor Folk	1844	32322	176136	57,4
EN9977	known	Aldous Huxley	Crome Yellow	1921	61556	357754	
	unknown	Fyodor Dostoyevsky	Poor Folk	1844	32322	176136	52,5
EN9978	known	Aldous Huxley	Limbo	1920	54613	309961	
	unknown	Fyodor Dostoyevsky	Poor Folk	1844	32322	176136	59,2
EN9979	known	Aldous Huxley	Leda	1920	27137	113256	
	unknown	Fyodor Dostoyevsky	Poor Folk	1844	32322	176136	119,1
EN9980	known	Aldous Huxley	The Burning Wheel	1920	11961	52276	
	unknown	Fyodor Dostoyevsky	Poor Folk	1844	32322	176136	270,2
EN9981	known	H.L.Sayler	In the Clouds for Uncle Sam	1910	45764	254904	
	unknown	Fyodor Dostoyevsky	Poor Folk	1844	32322	176136	70,6
EN9982	known	H.L.Sayler	The Airship Boys' Ocean Flyer	1909	73115	406463	
	unknown	Fyodor Dostoyevsky	Poor Folk	1844	32322	176136	44,2
EN9983	known	H.L.Sayler	The Aeroplane Express	1910	50754	281270	
	unknown	Fyodor Dostoyevsky	Poor Folk	1844	32322	176136	63,7
EN9984	known	H.L.Sayler	The Stolen Aeroplane	1911	46955	261851	
	unknown	Fyodor Dostoyevsky	Poor Folk	1844	32322	176136	68,8

EN9985	known	Felix Dahn	The Scarlet Banner	1861	98432	541289	
	unknown	Fyodor Dostoyevsky	Poor Folk	1844	32322	176136	32,8
EN9986	known	Felix Dahn	A Struggle for Rome	1876	111084	623201	
	unknown	Fyodor Dostoyevsky	Poor Folk	1844	32322	176136	29,1
EN9987	known	Felix Dahn	A Struggle for Rome	1876	117371	652843	
	unknown	Fyodor Dostoyevsky	Poor Folk	1844	32322	176136	27,5
EN9988	known	Felix Dahn	Felicitas	1892	43179	239791	
	unknown	Fyodor Dostoyevsky	Poor Folk	1844	32322	176136	74,9
EN9989	known	Dante Alighieri	The Divine Comedy	1308	115138	641442	
	unknown	Fyodor Dostoyevsky	Poor Folk	1844	32322	176136	28,1
EN9990	known	Dante Alighieri	The Divine Comedy of Dante Alighieri	1308	150098	715517	
	unknown	Fyodor Dostoyevsky	Poor Folk	1844	32322	176136	21,5
EN9991	known	Dante Alighieri	The Vision of Hell	1320	41139	228502	
	unknown	Fyodor Dostoyevsky	Poor Folk	1844	32322	176136	78,6
EN9992	known	Dante Alighieri	The New Life (La Vita Nuova)	1320	33845	163301	
	unknown	Fyodor Dostoyevsky	Poor Folk	1844	32322	176136	95,5
EN9993	known	Richard Harding Davis	Cinderella	1891	45236	247090	
	unknown	Fyodor Dostoyevsky	Poor Folk	1844	32322	176136	71,5
EN9994	known	Richard Harding Davis	A Charmed Life	1891	8214	46940	
	unknown	Fyodor Dostoyevsky	Poor Folk	1844	32322	176136	393,5
EN9995	known	Richard Harding Davis	Captain Macklin	1902	80210	437369	
	unknown	Fyodor Dostoyevsky	Poor Folk	1844	32322	176136	40,3
EN9996	known	Richard Harding Davis	In the Fog	1897	26067	143730	
	unknown	Fyodor Dostoyevsky	Poor Folk	1844	32322	176136	124,0
EN9997	known	Philip K.Dick	Beyond the Door	1954	6037	33871	
	unknown	Fyodor Dostoyevsky	Poor Folk	1844	32322	176136	535,4
EN9998	known	Philip K.Dick	Beyond Lies the Wub	1952	6407	35767	
	unknown	Fyodor Dostoyevsky	Poor Folk	1844	32322	176136	504,5
EN9999	known	Philip K.Dick	The Defenders	1953	13015	70710	
	unknown	Fyodor Dostoyevsky	Poor Folk	1844	32322	176136	248,3
EN99991	known	Philip K.Dick	The Crystal Crypt	1954	11355	63200	
	unknown	Fyodor Dostoyevsky	Poor Folk	1844	32322	176136	284,6

Tabla 28: Nivel 1

Carpeta	Archivo	Autor	Libro	Año	Cant Palabras	Cant Caracteres	% DeTC
EN01	known	Mark Twain	Adventures of Huckleberry Finn	1884	116949	616320	
	known	Mark Twain	The Adventures of Tom Sawyer	1876	76142	428104	
	unknown	Mark Twain	The Prince and The Pauper	1881	149879	843281	77,6
EN02	known	Mark Twain	Adventures of Huckleberry Finn	1884	116949	616320	
	known	Mark Twain	The Adventures of Tom Sawyer	1876	76142	428104	
	unknown	Mark Twain	Life On The Mississippi	1883	75179	431692	38,9
EN03	known	Mark Twain	Adventures of Huckleberry Finn	1884	116949	616320	
	known	Mark Twain	The Adventures of Tom Sawyer	1876	76142	428104	
	unknown	Herman Melville	Moby Dick	1851	219992	1270330	113,9
EN04	known	Mark Twain	Adventures of Huckleberry Finn	1884	116949	616320	
	known	Mark Twain	The Adventures of Tom Sawyer	1876	76142	428104	
	unknown	Pedro de Angelis	Colección de viajes y expediciónes	1837	58144	320284	30,1
EN05	known	Jane Austen	Pride and Prejudice	1813	127368	724725	
	known	Jane Austen	Sense and Sensibility	1811	127536	706152	
	unknown	Jane Austen	Persuasion	1818	90156	495022	35,4
EN06	known	Jane Austen	Pride and Prejudice	1813	127368	724725	
	known	Jane Austen	Sense and Sensibility	1811	127536	706152	
	unknown	Jane Austen	Mansfield Park	1814	164911	918376	64,7
EN07	known	Jane Austen	Pride and Prejudice	1813	127368	724725	
	known	Jane Austen	Sense and Sensibility	1811	127536	706152	
	unknown	Georgette Heyer	The Black Moth	1921	101061	554036	39,6
EN08	known	Jane Austen	Pride and Prejudice	1813	127368	724725	
	known	Jane Austen	Sense and Sensibility	1811	127536	706152	
	unknown	Mark Twain	Life On The Mississippi	1883	149879	843281	58,8
EN09	known	Charles Dickens	Oliver Twist	1838	169865	936281	
	known	Charles Dickens	The Life And Adventures Of Nicholas Nickleby	1839	333304	1951260	
	unknown	Charles Dickens	The Old Curiosity Shop	1840	228254	1282845	45,4
EN10	known	Charles Dickens	Oliver Twist	1838	169865	936281	
	known	Charles Dickens	The Life And Adventures Of Nicholas Nickleby	1839	333304	1951260	
	unknown	Charles Dickens	A Tale of Two Cities	1859	143239	792985	28,5
EN11	known	Charles Dickens	Oliver Twist	1838	169865	936281	
	known	Charles Dickens	The Life And Adventures Of Nicholas Nickleby	1839	333304	1951260	
	unknown	Mark Twain	Adventures of Huckleberry Finn	1884	116949	616320	23,2

EN12	known	Charles Dickens	Oliver Twist	1838	169865	936281	
	known	Charles Dickens	The Life And Adventures Of Nicholas Nickleby	1839	333304	1951260	
	unknown	Herman Melville	Moby Dick	1851	219992	1270330	43,7
EN13	known	Leon Toltoi	War and Peace	1865	583077	3359550	
	known	Leon Toltoi	Anna Karenina	1877	362223	2026606	
	unknown	Leon Toltoi	The Kreutzer Sonata and Other Stories	1890	68864	384147	7,3
EN14	known	Leon Toltoi	War and Peace	1865	583077	3359550	
	known	Leon Toltoi	Anna Karenina	1877	362223	2026606	
	unknown	Leon Toltoi	Childhood	1852	42596	235951	4,5
EN15	known	Leon Toltoi	War and Peace	1865	583077	3359550	
	known	Leon Toltoi	Anna Karenina	1877	362223	2026606	
	unknown	Fyodor Dostoyevsky	The Brothers Karamazov	1879	362578	2044191	38,4
EN16	known	Leon Toltoi	War and Peace	1865	583077	3359550	
	known	Leon Toltoi	Anna Karenina	1877	362223	2026606	
	unknown	Pedro de Angelis	Colección de viajes y expediciónes	1837	58144	320284	6,2
EN17	known	Edgar Alan Poe	The Works of Edgar Allan Poe Vol 2	1840	100675	583046	
	known	Edgar Alan Poe	The Works of Edgar Allan Poe Vol 3	1842	104042	598631	
	unknown	Edgar Alan Poe	The Works of Edgar Allan Poe Vol 4	1844	91406	537725	44,6
EN18	known	Edgar Alan Poe	The Works of Edgar Allan Poe Vol 2	1840	100675	583046	
	known	Edgar Alan Poe	The Works of Edgar Allan Poe Vol 3	1842	104042	598631	
	unknown	Edgar Alan Poe	Eureka	1848	49082	296881	24,0
EN19	known	Edgar Alan Poe	The Works of Edgar Allan Poe Vol 2	1840	100675	583046	
	known	Edgar Alan Poe	The Works of Edgar Allan Poe Vol 3	1842	104042	598631	
	unknown	H. P. Lovecraft	The Dunwich Horror	1928	21235	123406	10,4
EN20	known	Edgar Alan Poe	The Works of Edgar Allan Poe Vol 2	1840	100675	583046	
	known	Edgar Alan Poe	The Works of Edgar Allan Poe Vol 3	1842	104042	598631	
	unknown	Mark Twain	Adventures of Huckleberry Finn	1884	116949	616320	57,1
EN21	known	Virginia Woolf	Night and Day	1919	172685	1001342	
	known	Virginia Woolf	Jacob's Room	1922	59118	340092	
	unknown	Virginia Woolf	The Voyage Out	1915	142665	799814	61,5
EN22	known	Virginia Woolf	Night and Day	1919	172685	1001342	
	known	Virginia Woolf	Jacob's Room	1922	59118	340092	
	unknown	Virginia Woolf	Monday or Tuesday	1921	21961	125059	9,5
EN23	known	Virginia Woolf	Night and Day	1919	172685	1001342	

	known	Virginia Woolf	Jacob's Room	1922	59118	340092	
	unknown	Jane Austen	Pride and Prejudice	1813	127368	724725	54,9
EN24	known	Virginia Woolf	Night and Day	1919	172685	1001342	
	known	Virginia Woolf	Jacob's Room	1922	59118	340092	
	unknown	Leon Toltoi	Anna Karenina	1877	362223	2026606	156,3
EN25	known	Homero	The Iliad		252365	1213898	
	unknown	Homero	The Odyssey		134274	709420	53,2
EN26	known	Homero	The Iliad		252365	1213898	
	unknown	Homero	The Adventures of Ulysses the Wanderer	157430	843123	843123	62,4
EN27	known	Homero	The Iliad		252365	1213898	
	unknown	Platon	Ion		10038	55890	4,0
EN28	known	Homero	The Iliad		252365	1213898	
	unknown	Jane Austen	Pride and Prejudice	1813	127368	724725	50,5
EN29	known	Arthur Conan Doyle	The Adventures of Sherlock Holmes	1892	110682	594951	
	known	Arthur Conan Doyle	Memoirs of Sherlock Holmes	1894	93146	510601	
	unknown	Arthur Conan Doyle	The Return of Sherlock Holmes	1905	118273	653587	58,0
EN30	known	Arthur Conan Doyle	The Adventures of Sherlock Holmes	1892	110682	594951	
	known	Arthur Conan Doyle	Memoirs of Sherlock Holmes	1894	93146	510601	
	unknown	Arthur Conan Doyle	The Hound of the Baskervilles	1902	66002	352886	32,4
EN31	known	Arthur Conan Doyle	The Adventures of Sherlock Holmes	1892	110682	594951	
	known	Arthur Conan Doyle	Memoirs of Sherlock Holmes	1894	93146	510601	
	unknown	Agatha Christie	The Secret Adversary	1922	82038	473691	40,2
EN32	known	Arthur Conan Doyle	The Adventures of Sherlock Holmes	1892	110682	594951	
	known	Arthur Conan Doyle	Memoirs of Sherlock Holmes	1894	93146	510601	
	unknown	Platon	The Republic		224338	1239109	110,1
EN33	known	Platon	The Republic		224338	1239109	
	known	Platon	Symposium		36144	200852	
	unknown	Platon	Apology		19520	106904	7,5
EN34	known	Platon	The Republic		224338	1239109	
	known	Platon	Symposium		36144	200852	
	unknown	Platon	Protagoras		31821	176081	12,2
EN35	known	Platon	The Republic		224338	1239109	
	known	Platon	Symposium		36144	200852	
	unknown	Homero	The Iliad		252365	1213898	96,9

EN36	known	Platon	The Republic		224338	1239109	
	known	Platon	Symposium		36144	200852	
	unknown	Jane Austen	Pride and Prejudice	1813	127368	724725	48,9
EN37	known	F.Nietzsche	Beyond Good and Evil	1886	67488	408809	
	known	F.Nietzsche	Thus Spake Zarathustra	1896	121028	680764	
	unknown	F.Nietzsche	The Antichrist	1895	37900	226301	20,1
EN38	known	F.Nietzsche	Beyond Good and Evil	1886	67488	408809	
	known	F.Nietzsche	Thus Spake Zarathustra	1896	121028	680764	
	unknown	F.Nietzsche	Ecce Homo	1908	57691	304122	30,6
EN39	known	F.Nietzsche	Beyond Good and Evil	1886	67488	408809	
	known	F.Nietzsche	Thus Spake Zarathustra	1896	121028	680764	
	unknown	Jean Pierre Camus	The Spirit of St. Francis de Sales	1639	149163	826958	79,1
EN40	known	F.Nietzsche	Beyond Good and Evil	1886	67488	408809	
	known	F.Nietzsche	Thus Spake Zarathustra	1896	121028	680764	
	unknown	Charles Dickens	Oliver Twist	1838	169865	936281	90,1
EN41	known	Shakespear	The Complete Works of William Shakespeare	1600	1388740	5858792	
	known	Shakespear	Romeo and Juliet	1597	45632	179010	
	unknown	Shakespear	Hamlet	1599	36821	204885	2,6
EN42	known	Shakespear	The Complete Works of William Shakespeare	1600	1388740	5858792	
	known	Shakespear	Romeo and Juliet	1597	45632	179010	
	unknown	Shakespear	The Tragedie of Macbeth		23142	119845	1,6
EN43	known	Shakespear	The Complete Works of William Shakespeare	1600	1388740	5858792	
	known	Shakespear	Romeo and Juliet	1597	45632	179010	
	unknown	Christoper Marlowe	The Tragical History of Dr. Faustus	1592	33343	147233	2,3
EN44	known	Shakespear	The Complete Works of William Shakespeare	1600	1388740	5858792	
	known	Shakespear	Romeo and Juliet	1597	45632	179010	
	unknown	Leon Toltoi	Anna Karenina	1877	362223	2026606	25,3
EN45	known	Fyodor Dostoyevsky	The Brothers Karamazov	1879	362578	2044191	
	known	Fyodor Dostoyevsky	The Idiot	1868	250369	1425309	
	unknown	Fyodor Dostoyevsky	The Grand Inquisitor	1879	13070	75448	2,1
EN46	known	Fyodor Dostoyevsky	The Brothers Karamazov	1879	362578	2044191	
	known	Fyodor Dostoyevsky	The Idiot	1868	250369	1425309	
	unknown	Fyodor Dostoyevsky	Poor Folk	1844	57988	317754	9,5
EN47	known	Fyodor Dostoyevsky	The Brothers Karamazov	1879	362578	2044191	

	known	Fyodor Dostoyevsky	The Idiot	1868	250369	1425309	
	unknown	Leon Toltoi	Anna Karenina	1877	362223	2026606	59,1
EN48	known	Fyodor Dostoyevsky	The Brothers Karamazov	1879	362578	2044191	
	known	Fyodor Dostoyevsky	The Idiot	1868	250369	1425309	
	unknown	Jane Austen	Pride and Prejudice	1813	127368	724725	20,8
EN49	known	Charlotte Brontë	Jane Eyre	1847	198454	1070330	
	known	Charlotte Brontë	Villette	1853	200872	1137607	
	unknown	Charlotte Brontë	Shirley	1811	228913	1274038	57,3
EN50	known	Charlotte Brontë	Jane Eyre	1847	198454	1070330	
	known	Charlotte Brontë	Villette	1853	200872	1137607	
	unknown	Charlotte Brontë	The Professor	1857	93665	536673	23,5
EN51	known	Charlotte Brontë	Jane Eyre	1847	198454	1070330	
	known	Charlotte Brontë	Villette	1853	200872	1137607	
	unknown	Jane Austen	Pride and Prejudice	1813	127368	724725	31,9
EN52	known	Charlotte Brontë	Jane Eyre	1847	198454	1070330	
	known	Charlotte Brontë	Villette	1853	200872	1137607	
	unknown	Shakespear	The Complete Works of William Shakespeare	1600	1388740	5858792	347,8
EN53	known	Lewis Carroll	Alice’s Adventures in Wonderland	1865	31709	173595	
	known	Lewis Carroll	Through the Looking-Glass	1871	36005	193607	
	unknown	Lewis Carroll	Tangled Tale	1885	35205	183983	52,0
EN54	known	Lewis Carroll	Alice’s Adventures in Wonderland	1865	31709	173595	
	known	Lewis Carroll	Through the Looking-Glass	1871	36005	193607	
	unknown	Lewis Carroll	Sylvie and Bruno	1889	72096	402973	106,5
EN55	known	Lewis Carroll	Alice’s Adventures in Wonderland	1865	31709	173595	
	known	Lewis Carroll	Through the Looking-Glass	1871	36005	193607	
	unknown	Charles Dickens	Oliver Twist	1838	169865	936281	250,9
EN56	known	Lewis Carroll	Alice’s Adventures in Wonderland	1865	31709	173595	
	known	Lewis Carroll	Through the Looking-Glass	1871	36005	193607	
	unknown	Leon Toltoi	Anna Karenina	1877	362223	2026606	534,9
EN57	known	Alexandre Dumas	The Count of Monte Cristo	1844	480561	2783699	
	known	Alexandre Dumas	The Three Musketeers	1844	241731	1385830	
	unknown	Alexandre Dumas	The Man in the Iron Mask	1857	182632	1066135	25,3
EN58	known	Alexandre Dumas	The Count of Monte Cristo	1844	480561	2783699	
	known	Alexandre Dumas	The Three Musketeers	1844	241731	1385830	

	unknown	Alexandre Dumas	Camille (La Dame aux Camilias)	1848	71957	382972	10,0
EN59	known	Alexandre Dumas	The Count of Monte Cristo	1844	480561	2783699	
	known	Alexandre Dumas	The Three Musketeers	1844	241731	1385830	
	unknown	Victor Hugo	Les Misérables	1862	589758	3324334	81,7
EN60	known	Alexandre Dumas	The Count of Monte Cristo	1844	480561	2783699	
	known	Alexandre Dumas	The Three Musketeers	1844	241731	1385830	
	unknown	Shakespear	Romeo and Juliet	1597	45632	179010	6,3
EN61	known	Gustave Flaubert	Madame Bovary	1856	122031	694672	
	known	Gustave Flaubert	A Simple Soul	1877	15661	90018	
	unknown	Gustave Flaubert	Salammbô	1862	110974	645642	80,6
EN62	known	Gustave Flaubert	Madame Bovary	1856	122031	694672	
	known	Gustave Flaubert	A Simple Soul	1877	15661	90018	
	unknown	Gustave Flaubert	The Temptation	1874	56269	331620	40,9
EN63	known	Gustave Flaubert	Madame Bovary	1856	122031	694672	
	known	Gustave Flaubert	A Simple Soul	1877	15661	90018	
	unknown	Charlotte Brontë	Jane Eyre	1847	198454	1070330	144,1
EN64	known	Gustave Flaubert	Madame Bovary	1856	122031	694672	
	known	Gustave Flaubert	A Simple Soul	1877	15661	90018	
	unknown	Jane Austen	Pride and Prejudice	1813	127368	724725	92,5
EN65	known	Oscar Wilde	The Happy Prince and Other Tales	1888	21249	111140	
	known	Oscar Wilde	The Picture of Dorian Gray	1890	85387	465976	
	unknown	Oscar Wilde	The Canterville Ghost	1887	15282	86711	14,3
EN66	known	Oscar Wilde	The Happy Prince and Other Tales	1888	21249	111140	
	known	Oscar Wilde	The Picture of Dorian Gray	1890	85387	465976	
	unknown	Oscar Wilde	For Love of the King	1926	7786	44922	7,3
EN67	known	Oscar Wilde	The Happy Prince and Other Tales	1888	21249	111140	
	known	Oscar Wilde	The Picture of Dorian Gray	1890	85387	465976	
	unknown	Lewis Carroll	Alice's Adventures in Wonderland	1865	31709	173595	29,7
EN68	known	Oscar Wilde	The Happy Prince and Other Tales	1888	21249	111140	
	known	Oscar Wilde	The Picture of Dorian Gray	1890	85387	465976	
	unknown	Jane Austen	Pride and Prejudice	1813	127368	724725	119,4
EN69	known	Eleanor Hallowell Abbott	The Stingy Receiver	1917	31598	181580	
	known	Eleanor Hallowell Abbott	Old-Dad	1923	45339	264631	
	unknown	Eleanor Hallowell Abbott	Rainy Week	1921	44887	263207	58,3

EN70	known	Eleanor Hallowell Abbott	The Stingy Receiver	1917	31598	181580	
	known	Eleanor Hallowell Abbott	Old-Dad	1923	45339	264631	
	unknown	Eleanor Hallowell Abbott	The Sick-a-Bed Lady	1911	80529	457784	104,7
EN71	known	Eleanor Hallowell Abbott	The Stingy Receiver	1917	31598	181580	
	known	Eleanor Hallowell Abbott	Old-Dad	1923	45339	264631	
	unknown	Jane Austen	Sense and Sensibility	1811	127536	706152	165,8
EN72	known	Eleanor Hallowell Abbott	The Stingy Receiver	1917	31598	181580	
	known	Eleanor Hallowell Abbott	Old-Dad	1923	45339	264631	
	unknown	Alexandre Dumas	The Count of Monte Cristo	1844	480561	2783699	624,6
EN73	known	L. Frank (Lyman Frank) Baum	The Land of Oz	1900	50779	277145	
	known	L. Frank (Lyman Frank) Baum	The Magic of Oz	1919	52866	265813	
	unknown	L. Frank (Lyman Frank) Baum	The Scarecrow of Oz	1915	55245	283781	53,3
EN74	known	L. Frank (Lyman Frank) Baum	The Land of Oz	1900	50779	277145	
	known	L. Frank (Lyman Frank) Baum	The Magic of Oz	1919	52866	265813	
	unknown	L. Frank (Lyman Frank) Baum	Tik-Tok of Oz	1914	56305	298267	54,3
EN75	known	L. Frank (Lyman Frank) Baum	The Land of Oz	1900	50779	277145	
	known	L. Frank (Lyman Frank) Baum	The Magic of Oz	1919	52866	265813	
	unknown	Mark Twain	Adventures of Huckleberry Finn	1884	116949	616320	112,8
EN76	known	L. Frank (Lyman Frank) Baum	The Land of Oz	1900	50779	277145	
	known	L. Frank (Lyman Frank) Baum	The Magic of Oz	1919	52866	265813	
	unknown	Jane Austen	Pride and Prejudice	1813	127368	724725	122,9
EN77	known	Aldous Huxley	Crome Yellow	1921	61556	357754	
	known	Aldous Huxley	Limbo	1920	54613	309961	
	unknown	Aldous Huxley	Leda	1920	27137	113256	23,4
EN78	known	Aldous Huxley	Crome Yellow	1921	61556	357754	
	known	Aldous Huxley	Limbo	1920	54613	309961	
	unknown	Aldous Huxley	The Burning Wheel	1920	11961	52276	10,3
EN79	known	Aldous Huxley	Crome Yellow	1921	61556	357754	
	known	Aldous Huxley	Limbo	1920	54613	309961	
	unknown	Ray Bradbury	A Little Journey	1927	6180	34971	5,3
EN80	known	Aldous Huxley	Crome Yellow	1921	61556	357754	
	known	Aldous Huxley	Limbo	1920	54613	309961	
	unknown	Platon	The Republic		224338	1239109	193,1
EN81	known	H.L.Sayler	In the Clouds for Uncle Sam	1910	45764	254904	

	known	H.L.Sayler	The Airship Boys' Ocean Flyer	1909	73115	406463	
	unknown	H.L.Sayler	The Aeroplane Express	1910	50754	281270	42,7
EN82	known	H.L.Sayler	In the Clouds for Uncle Sam	1910	45764	254904	
	known	H.L.Sayler	The Airship Boys' Ocean Flyer	1909	73115	406463	
	unknown	H.L.Sayler	The Stolen Aeroplane	1911	46955	261851	39,5
EN83	known	H.L.Sayler	In the Clouds for Uncle Sam	1910	45764	254904	
	known	H.L.Sayler	The Airship Boys' Ocean Flyer	1909	73115	406463	
	unknown	Alexandre Dumas	The Count of Monte Cristo	1844	480561	2783699	404,2
EN84	known	H.L.Sayler	In the Clouds for Uncle Sam	1910	45764	254904	
	known	H.L.Sayler	The Airship Boys' Ocean Flyer	1909	73115	406463	
	unknown	Platon	The Republic		224338	1239109	188,7
EN85	known	Felix Dahn	The Scarlet Banner	1861	98432	541289	
	known	Felix Dahn	A Struggle for Rome v1	1876	111084	623201	
	unknown	Felix Dahn	A Struggle for Rome v2	1876	117371	652843	56,0
EN86	known	Felix Dahn	The Scarlet Banner	1861	98432	541289	
	known	Felix Dahn	A Struggle for Rome v1	1876	111084	623201	
	unknown	Felix Dahn	Felicitas	1892	43179	239791	20,6
EN87	known	Felix Dahn	The Scarlet Banner	1861	98432	541289	
	known	Felix Dahn	A Struggle for Rome v1	1876	111084	623201	
	unknown	L. Frank (Lyman Frank) Baum	The Land of Oz	1900	50779	277145	24,2
EN88	known	Felix Dahn	The Scarlet Banner	1861	98432	541289	
	known	Felix Dahn	A Struggle for Rome v1	1876	111084	623201	
	unknown	Jane Austen	Pride and Prejudice	1813	127368	724725	60,8
EN89	known	Dante Alighieri	The Divine Comedy	1308	115138	641442	
	known	Dante Alighieri	The Divine Comedy of Dante Alighieri	1308	150098	715517	
	unknown	Dante Alighieri	The Vision of Hell	1320	41139	228502	15,5
EN90	known	Dante Alighieri	The Divine Comedy	1308	115138	641442	
	known	Dante Alighieri	The Divine Comedy of Dante Alighieri	1308	150098	715517	
	unknown	Dante Alighieri	The New Life (La Vita Nuova)	1320	33845	163301	12,8
EN91	known	Dante Alighieri	The Divine Comedy	1308	115138	641442	
	known	Dante Alighieri	The Divine Comedy of Dante Alighieri	1308	150098	715517	
	unknown	Pedro de Angelis	Colección de viajes y expediciónes	1837	58144	320284	21,9
EN92	known	Dante Alighieri	The Divine Comedy	1308	115138	641442	
	known	Dante Alighieri	The Divine Comedy of Dante Alighieri	1308	150098	715517	

	unknown	Jane Austen	Pride and Prejudice	1813	127368	724725	48,0	
EN93	known	Richard Harding Davis	Cinderella	1891	45236	247090		
	known	Richard Harding Davis	A Charmed Life	1891	8214	46940		
	unknown	Richard Harding Davis	Captain Macklin	1902	80210	437369	150,1	
EN94	known	Richard Harding Davis	Cinderella	1891	45236	247090		
	known	Richard Harding Davis	A Charmed Life	1891	8214	46940		
	unknown	Richard Harding Davis	In the Fog	1897	26067	143730	48,8	
EN95	known	Richard Harding Davis	Cinderella	1891	45236	247090		
	known	Richard Harding Davis	A Charmed Life	1891	8214	46940		
	unknown	L. Frank (Lyman Frank) Baum	The Land of Oz	1900	50779	277145	95,0	
EN96	known	Richard Harding Davis	Cinderella	1891	45236	247090		
	known	Richard Harding Davis	A Charmed Life	1891	8214	46940		
	unknown	Jane Austen	Pride and Prejudice	1813	127368	724725	238,3	
EN97	known	Philip K.Dick	Beyond the Door	1954	6037	33871		
	known	Philip K.Dick	Beyond Lies the Wub	1952	6407	35767		
	unknown	Philip K.Dick	The Defenders	1953	13015	70710	104,6	
EN98	known	Philip K.Dick	Beyond the Door	1954	6037	33871		
	known	Philip K.Dick	Beyond Lies the Wub	1952	6407	35767		
	unknown	Philip K.Dick	The Crystal Crypt	1954	11355	63200	91,2	
EN99	known	Philip K.Dick	Beyond the Door	1954	6037	33871		
	known	Philip K.Dick	Beyond Lies the Wub	1952	6407	35767		
	unknown	L. Frank (Lyman Frank) Baum	The Land of Oz	1900	50779	277145	408,1	
EN100	known	Philip K.Dick	Beyond the Door	1954	6037	33871		
	known	Philip K.Dick	Beyond Lies the Wub	1952	6407	35767		
	unknown	Herman Melville	Moby Dick	1851	50779	277145	408,1	