UNIVERSIDAD SAN FRANCISCO DE QUITO USFQ

Colegio de Ciencias e Ingenierías

Aplicación de estilometría para la atribución de autorías en e-mails y documentos informáticos

Propuesta metodológica y trabajo experimental

Jorge Roberto Maldonado Galiano

Ingeniería en Sistemas

Trabajo de titulación presentado como requisito
para la obtención del título de
Ingeniero en Sistemas

Quito, 27 de octubre de 2015

UNIVERSIDAD SAN FRANCISCO DE QUITO USFQ

COLEGIO DE CIENCIAS E INGENIERÍAS

HOJA DE CALIFICACIÓN
DE TRABAJO DE TITULACIÓN

Aplicación de estilometría para la atribución de autorías en e-mails y documentos informáticos

Roberto Maldonado

Calificación:

Nombre del profesor, Título académico — Mauricio Iturralde, Ph.D.

Firma del profesor _____

Quito, 27 de octubre de 2015

Derechos de Autor

Por medio del presente documento certifico que he leído todas las Políticas y Manuales de la Universidad San Francisco de Quito USFQ, incluyendo la Política de Propiedad Intelectual USFQ, y estoy de acuerdo con su contenido, por lo que los derechos de propiedad intelectual del presente trabajo quedan sujetos a lo dispuesto en esas Políticas.

Asimismo, autorizo a la USFQ para que realice la digitalización y publicación de este trabajo en el repositorio virtual, de conformidad a lo dispuesto en el Art. 144 de la Ley Orgánica de Educación Superior.

Firma del estudiante: _____

Nombres y apellidos: Jorge Roberto Maldonado Galiano

Código: 00100258

Cédula de Identidad: 1714164272

Lugar y fecha: Quito, octubre de 2015

RESUMEN

La Estilometría es el análisis por el cual se puede determinar la autoría de un texto, que incluye el estudio de rasgos propios que utiliza un escritor al redactar documentos.
En este trabajo de investigación e integración se presenta un programa con la capacidad de extraer diversos rasgos característicos de escritura, los mismos que son comparados contra otro tipo de redacción con la finalidad de obtener un porcentaje de similitud entre estilos diferentes de composición manejados por uno o varios autores específicos. Esto se ha logrado mediante incorporación de varios parámetros que son considerados relevantes en el momento de realizar un análisis, los que incluyen a observaciones estadísticas sobre componentes léxicos, sintácticos, semánticos y estructurales aplicados al español.

Palabras clave: estilometría, redacción, autoría, estadística, porcentaje de similitud, léxica, semántica, sintaxis, estructura.

ABSTRACT

Stylometry is the analysis by which authorship of a written text can be determined, analyzing special features that are unconsciously placed by a writer in his publications.
In this integrative and research paper an application with the ability to extract various features of writing is presented. These features are compared against another profile, composed by features, in order to obtain a similarity percentage between two styles of composition that may be from the same or different authors.
This has been achieved by incorporating several selected features that are considered relevant in order to perform an stylometric analysis, which include statistical observations about the pattern presented in the document, without setting aside the fact that it is applied to the Spanish language.

Key words: Stylometry, writing, author, statistics, percent similarity, lexical, semantic, syntax, structure.

TABLA DE CONTENIDO

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

INTRODUCCIÓN

Estilometría. A pesar de que esta palabra no se encuentra definida de forma oficial en el diccionario de la Real Academia Española de la lengua, el lector podrá identificar de forma rápida que se trata de una caracterización matemática sobre el estilo, debido a la etimología de la palabra. En términos generales el término "estilo" se refiere al conjunto de rasgos peculiares que dan características a una persona u obra y le confiere una personalidad propia y reconocible.

Por lo que, en esta investigación se define el término "estilometría" que está estrechamente vinculado al manejo del lenguaje, con las características que presenta un estilo de redacción, el mismo que se atribuye como una expresión lingüística única de una persona.

Diferentes investigadores han sostenido la propuesta de que un estilo propio y único de expresión del ser humano proviene desde tiempos antiguos en los cuales famosos científicos, entre ellos matemáticos, físicos y filósofos han establecido distintos parámetros según los cuales se podría inferir la identificación de una persona. Por ejemplo se puede considerar el criterio de Thomas Corwin Mendehall, físico y meteorólogo de origen norteamericano quién propuso la existencia de una curva característica de composición que es determinada por la frecuencia en el uso de palabras que poseen diferentes longitudes [9][12]. Manifestación que efectivamente resultó ser verdadera puesto que investigaciones que se desarrollaron posteriores a está época aplicaban a la longitud de una palabra como una métrica importante para el análisis en los estilos de redacción.

La estilometría siempre es aplicada con un propósito concreto, que de forma universal persigue identificar un patrón o curva característica de escritura relacionada a la forma

de expresión de un escritor. En otras palabras, la estilometría busca generar un perfil de escritura por medio de métodos computacionales, el mismo que debe provenir de la recolección de datos, que incluyen el recuerdo de preferencias e información recolectada a partir de interacciones previas [3]. Cabe aclarar que en este contexto, el verbo recordar es empleado como un sinónimo de almacenamiento de los datos extraídos, para poder utilizarlo en el futuro ya sea para reinterpretar información o emplearla para realizar comparaciones. En la estilometría se enfrenta un problema del tipo de clasificación supervisada, puesto que a partir de un conjunto de datos provistos que están pre-clasificados, intentamos categorizar nuevos datos dentro de los conjuntos preexistentes, caso semejante al filtrado de información [8].

OBJETIVOS

Objetivo General

Implementar una herramienta informática, en el lenguaje de programación C/C++, la misma que permita identificar el autor de un e-mail basado en sus patrones de escritura.

Objetivos Específicos

- Implementación de algoritmos preexistentes dedicados a la identificación de autorías de textos digitales.
- Realización de cuadros comparativos entre los algoritmos preexistentes de mayor precisión.
- Proposición de un nuevo algoritmo que funcione de forma efectiva para textos en español, tomando en cuenta las estructuras esenciales del idioma.

JUSTIFICACIÓN

El presente trabajo se manifiesta como una investigación académica con la cual se busca otorgar un aporte positivo al desarrollo de herramientas informáticas para la detección de la criminalidad digital. La herramienta y sus métodos de funcionamiento en conjunto con las métricas del análisis textual, permitirán la identificación de la autoría de un texto con lo cuál se puede desenmascarar a un cyber-criminal que intentó ocultarse en el anonimato o en la suplantación de identidad, vulnerabilidad que está presente en el internet.

ANTECEDENTES

La identidad en el internet siempre ha sido uno de los mayores retos que enfrenta la informática, razón por la cual se han ido desarrollando varias técnicas para validar y llevar el control de los usuarios que utilizan aplicaciones con diversas funcionalidades. Para resolver estos retos, se debe reconocer la tendencia de que personas que desarrollan preferencias en cuanto a sus costumbres, en este caso en patrones de redacción.

Con este antecedente, el presente trabajo busca identificar dichas tendencias de redacción para emplearlas en el análisis de textos y finalmente relacionarlas a una persona que sería identificada como el autor de dicho texto.

TRABAJO RELACIONADO

Con la estilometría, o estudio cuantitativo del estilo de escritura de un autor, en este trabajo se busca trabajar sobre la extracción de un patrón de escritura único a partir de textos preliminares con los que se cuenta de forma anticipada antes de realizar un análisis comparativo y estadístico de un nuevo texto.

Una vez que se identifica el patrón único, es sencillo realizar el reconocimiento y atribución de la autoría para un nuevo texto, puesto que simplemente se debe realizar una comparación con los parámetros que se seleccionan según el lenguaje en el que se aplica el análisis. Formalmente hablando, la identificación de autorías es el trabajo que se realiza para determinar el autor de una porción de un documento o publicación [19]. En el trabajo desarrollado, como se verá en la sección de resultados, la atribución de autoría queda abierta para ser interpretada por el usuario de la herramienta quien podrá ver el resultado final en términos de porcentajes de similitud.

Además, en esta investigación el análisis se basará en las reglas básicas del lenguaje español, así como en la infracción de las mismas. Además se contará con otras propiedades que se verán a lo largo del trabajo cuando se explique el análisis ortográfico, léxico, estructural y sintáctico de los textos que se proveen de antemano. Por el momento es importante conocer que los parámetros de análisis caen en distintas categorías y que cada uno es evaluado con cuentas globales y únicas de todos los textos alimentados al sistema.

Cabe recalcar que la estilometría puede también identificar características sobre la educación de la persona puesto que revela el uso de palabras, signos de puntuación, contenido y finalmente define un estilo propio. Una distinción importante que no se debe dejar pasar es el concepto entre la identificación de una autoría y una

caracterización de una autoría para un texto establecido, según se indica en la publicación [1] titulada *Applying Autorship Analysis to Extremist-Group Forum Messages*:

- Identificación de autoría: Se refiere a la atribución de una autoría (a una pieza no identificada) que se otorga basándose en las similitudes estilísticas entre obras identificadas previamente como pertenecientes al autor y la pieza no identificada. Por lo que es un problema de clasificación.
- Caracterización de autoría: Intenta formular un perfil de autor para hacer inferencias sobre el género, el nivel de educación y la formación cultural basado en el estilo presente en la escritura.

La identificación de autorías para textos busca contribuir a la gran pregunta que surge ante la gran incógnita que representa el descubrimiento del autor de un documento, así como sus antecedentes demográficos y también su posible relación con documentos publicados anteriormente [12].

Es necesario tomar en cuenta que en la actualidad algunos métodos de comunicación han evolucionado, facilitado el acceso a la tecnología; Ofreciendo una simplicidad absoluta para el anonimato, lo que abre la posibilidad al envío de mensajes escritos de origen anónimo, que muchas veces llevan consigo intenciones abusivas hacia el destinatario, tales como: Cyber bullying, spam, hoaxes, entre otros [1][13].

La aplicación de las técnicas utilizadas para la estilometría o análisis que sean semejantes no están vinculadas específicamente a un área; sino que varias áreas son invocadas [7], que incluyen: criminalística, comercio digital, educación, entre otros. En concordancia también se emplea la estilometría "…en derecho civil para identificación de propiedad intelectual en caso de controversias; en derecho penal para la identificación de los autores de notas, cartas de pago y cartas de acoso; y para la minería

de datos en el contenido de correos electrónicos en el campo de la seguridad informática para identificar un autor." [12]. Finalmente, es sencillo demostrar que la aplicación de la estilometría es muy importante en tiempos actuales, debido a las facilidades que otorga la tecnología para esconder los verdaderos autores de documentos que pueden ir desde potenciales burlas hasta amenazas terroristas verdaderas y que pueden identificarse mediante las herramientas diseñadas para este fin.

La estilometría es un tema que ha ganado relevancia en la época actual debido a los factores expuestos anteriormente, razón por la cual varios investigadores han dedicado su tiempo para extraer los mejores indicadores. Dichos indicadores caen dentro de varias categorías que incluyen a la aplicación de métodos léxicos, sintácticos, semánticos, así como al estudio de caracteres y de contenido relacionado a temas específicos. Cada uno forma parte de un análisis dirigido, con sus componentes como se indica a continuación.

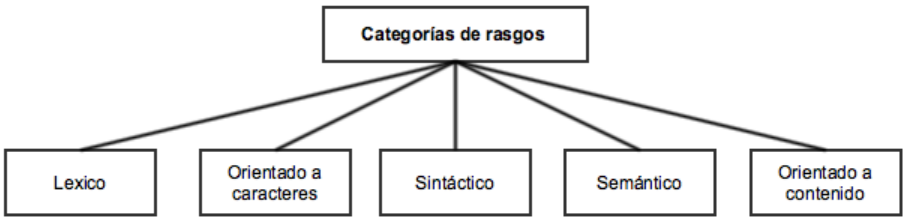

Figura 1. Categorización de métodos estilo-métricos

Los métodos léxicos se basan en el conteo de palabras y su distribución a lo largo del texto. Mientras que los métodos sintácticos se vinculan directamente a la frecuencia que tiene una palabra específica dentro del texto. Adicionalmente, los métodos semánticos se concentran en el vocabulario empleado y selección de palabras en la redacción [10]. Las categorías orientadas a caracteres definen una variedad de datos cuantificables a nivel carácter, incluyendo el conteo de caracteres alfabéticos, conteo de dígitos, conteos

de signos de puntuación y frecuencia de las letras. Esta categoría comprende a la palabra como una sencilla secuencia de caracteres [23] [25].

Y la última categoría son los rasgos orientados al contenido que son una colección de palabras clave y frases de un determinado tema. Se ha demostrado que los rasgos de esta categoría son importantes características discriminantes en los mensajes electrónicos [5][25].

Debido a que los resultados de las investigaciones de [25] han demostrado que la combinación entre información semántica, léxica y sintáctica mejoraron la exactitud de clasificación, entonces en este trabajo se integran aquellas categorías en conjunto a las indicadas en la figura anterior.

En las secciones siguientes de este trabajo de investigación se demostrará la proposición, integración y modo de empleo de una herramienta diseñada para la comparación de documentos informáticos, con la finalidad de extraer un porcentaje de similitud entre sus patrones y tendencias de escritura.

SISTEMA PROPUESTO

De acuerdo con los objetivos señalados en esta investigación de carácter integrativo se propone un sistema en el cual se consideran bastantes categorías para la aplicación de la estilometría en documentos informáticos, por medio de una herramienta que permita extraer la información relacionada. Por lo que se propone un programa de consola, que es escrito en un sistema UNIX, con el lenguaje de programación C/C++. Para lograr el objetivo se utiliza el compilador: *g++*, (emitido por Free Software Foundation), que fue configurado en el ambiente de desarrollo *NetBeans 7.4.* En este caso la selección sucedió debido a la infraestructura que brinda para la organización de diferentes tipos de archivos y la preferencia del desarrollador.

Como complemento de la herramienta, también se empleó la técnica de Shell Scripting en Unix con *!bin-bash*, para colocar cadenas de instrucciones previamente escrititas y manejar con sencillez los datos y generar gráficos sobre las estadísticas obtenidas.

EL sistema que se propone incluye algunas de las técnicas usadas dentro de la rama del aprendizaje de máquinas, que se emplean constantemente en la actualidad para entrenar un modelo de comportamiento. El esquema para la aplicación que se ha propuesto, se indica en la siguiente figura:

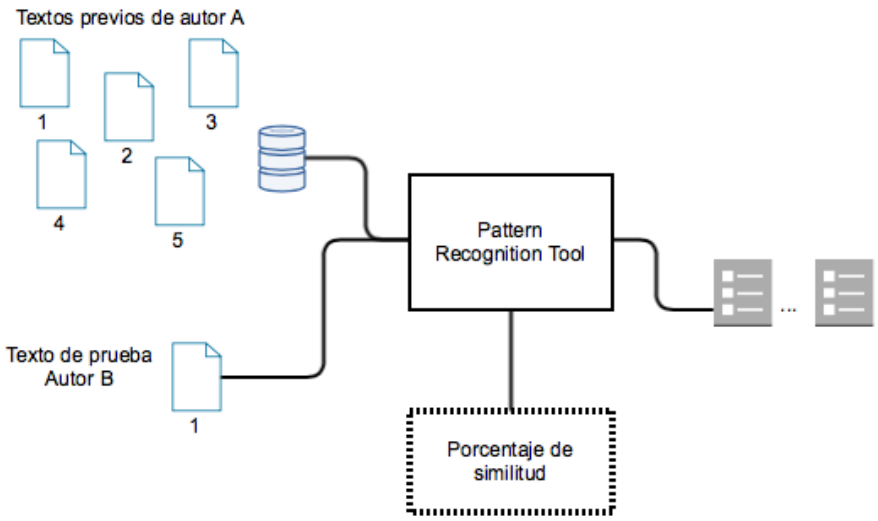

Figura 2. Arquitectura de la aplicación

Como se puede verificar, la aplicación requiere de cinco documentos distintos de un mismo autor (Autor A) y de un documento de un autor diferente(Autor B), los que son analizados por la herramienta para extraer datos relevantes, compararlos y generar un resultado porcentual en términos de la similitud hallada. El resultado indicado se presenta en conjunto con distintos archivos que guardan la información extraída para ser utilizada en los elementos gráficos que consisten en histogramas con barras de error generados por la desviación estándar de los datos.

Figura 3. Funcionalidad complementaria para la herramienta

Para cumplir con el funcionamiento completo de “Pattern Recognition Tool” se emplearon tres clases (objetos de datos empleados en la informática) [6] que cumplen diferentes roles en la actividad del programa, lo que incluye el ingreso, procesamiento de datos y generación de resultados tanto en archivos de planos de texto con extensión

.txt así como de gráficos *.eps*. Por medio de los gráficos, se busca facilitar la comprensión sobre el comportamiento de los datos al usuario de la aplicación. Dichos componentes visuales fueron desarrolladas con la utilidad portable GnuPlot 5.0 (An intercative plotting program), que tiene sus bases de funcionamiento por medio de la línea de comandos en varios sistemas operativos que incluyen Linux, MS Windows, OSX, y otras plataformas. La utilidad fue seleccionada debido a que se consideró su origen de distribución libre.

Desarrollo de la herramienta

Las funciones de cada una de las clases son las indicadas a continuación:

- FileHandler

  Esta es la clase encargada de administrar los archivos y alimentarlos al programa cada vez que realiza una inspección de un conjunto de textos. Además de esto, es la encargada de interactuar con el usuario para conocer con certeza si desea realizar un estudio completo de parámetros o si únicamente desea tomar en cuenta ciertos aspectos al criterio del que emplea la herramienta.

- TextParser

  Esta es la clase que realiza la inspección de cada uno de los textos almacenados para este fin. La clase intenta dividir el análisis desde la estructura mas grande hasta la más pequeña que en este caso están representadas por: todo el documento, cada uno de los párrafos que componen el documento, así mismo cada oración que es una estructura más básica del lenguaje. De igual manera cada palabra es analizada con respecto a su composición dada por los caracteres que se emplean para formarla.

Además de esto, la clase está encargada de buscar coincidencias con expresiones comunes que se utilizan en el lenguaje formal y coloquial del lenguaje que se habla en una zona específica. En el presente estudio se emplea el lenguaje español, con sus variantes para Latinoamérica.

- Calculator

  Finalmente, los cálculos estadísticos son realizados por esta clase que implementa varios métodos que obtienen el total de similitud entre el contenido de documentos escritos por dos diferentes autores, basándose en una agrupación de parámetros particulares. Dependiendo de la categoría se realiza el cálculo ya sea utilizando una media aritmética común o por medio de una media aritmética ponderada (Explicación mas detallada en sección "utilización de la media ponderada").

Los resultados son guardados dentro de la misma carpeta donde se encuentran los archivos de texto, en este caso el programa incluye *n* archivos de estadísticas, por medio de los cuales se puede colocar en forma gráfica los resultados.

Los archivos del programa deben tener una carpeta llamada *files*, en donde se colocan los cinco textos de un autor conocido bajo el nombre de: *default2.txt* hasta *default6.txt*.

El archivo perteneciente al autor desconocido (Autor B), es decir, aquel contra el que se comparan los parámetros de análisis, debe ser colocado dentro de un documento titulado *default1.txt.*

Puesto que la herramienta fue compuesta con la utilidad de Gnuplot 5.0, entonces se ha añadido la carpeta denominada *shells*, la misma que contiene un conjunto de comandos que se ejecutan para generar los resultados visuales de forma automática.

Selección de parámetros

Los parámetros que fueron seleccionados pueden ser clasificados en varios grupos, según lo que busca determinar sobre el autor:

- Análisis de la léxica empleada.

  En este análisis se busca determinar de que forma el autor de un texto utiliza un determinado conjunto de palabras dentro de su composición escrita.

- Análisis de sintaxis que tiende a ser utilizada.

  En este caso se enfoca un conjunto de posibles combinaciones que el autor emplea al momento de redactar un texto digital, por lo que se consideran aspectos como oraciones en una párrafo, palabras en una oración y letras en una palabra. Así mismo se aplica este análisis sobre la distribución de los signos de puntuación y su frecuencia.

- Análisis de distribución de letras.

  Para esta parte del análisis, se extrae el total de cada una de las letras del abecedario que han sido empleadas, puesto que se ha demostrado que un autor posee preferencia por las palabras con las que se siente cómodo empleando, debido a experiencias previas. Este análisis sirve para determinar también la frecuencia con la que se emplean letras mayúsculas en relación a letras minúsculas en el momento en que se comparan textos. No se debe olvidar que también se incluye a las vocales minúsculas o mayúsculas que llevan acentuación, representada por medio de la tilde.

- Utilización de números descritos en palabras contra signos numéricos.

  Nuevamente aquí se determina otra tendencia relacionada con el comportamiento que expresa un autor, en el instante que debe utilizar un número o dejar una referencia en el contenido de un mensaje electrónico.

- Análisis sobre presencia y ausencia de palabras de saludo y despedida.

En esta parte se puede determinar la formalidad con la que se expresa el autor y los sentimientos que posee hacia la persona o institución a la que se dirige. Se debe tomar en cuenta que el autor generalmente posee expresiones que se utilizan en el medio en el que se desarrollan sus actividades de la vida cotidianas, puesto que el mismo le dictará reglas distintas sobre la forma en que una expresión es aprobada o desaprobada por el medio.

Utilización de los métodos de estimación

Una de las principales diferencias que compone la investigación desarrollada y en la que se hace hincapié a lo largo del trabajo es que no se emplean los métodos de aprendizaje de máquinas para generar un modelo predictivo o de clasificación basado en conjuntos de parámetros selectos. La exclusión se realiza puesto que clasificadores tal como los clasificadores bayesianos se apoyan en la admisión de un supuesto, a menudo defectuoso, que cada una de las características son igual de importantes e independientes [14]. Este es el caso que se presenta en la estilometría porque existen datos que en realidad son muy dependientes de otros, por los que no se puede asumir de ninguna manera este supuesto. En la investigación se presenta un acercamiento a través de la ponderación de la media aritmética que permite que una característica tenga un peso definido, dependiendo de la importancia que tiene dicho parámetro. En este caso se otorga un peso considerable únicamente a los parámetros que han presentado la existencia de similitud porque como fue discutido antes, el emparejamiento de la inexistencia de un rasgo común, no es un indicador que otorgue garantía alguna de similitud como ocurre en el caso contrario. Es por esta razón que se aplica la

ponderación como se indica en la siguiente sección, además de otros métodos que analizan las distancias relativas que se mantienen con respecto a un vector.

Con el fin de ilustrar lo que sucede en el momento que se reemplaza el método convencional de la media aritmética por la media aritmética ponderada se supone el siguiente caso:

Tabla 1. Parámetros del caso demostrativo

ID Parámetro	Valores Obtenidos para Autor A	Valores Obtenidos para Autor B
1	30	29
2	0	0
3	10	5
4	4	3
5	2	1

Tomando en cuenta cinco características únicas de los textos de un autor A y considerando la semejanza por medio de una regla de proporción, conceptualizada con la diferencia del valor absoluto entre los valores obtenidos por los autores, como se visualiza en la tabla 2, el resultado sin ponderar que es equivalente al 74.33%.

Tabla 2. Demostración de funcionamiento de media aritmética

ID Parámetro	**Valor Determinado. (Autor A)**	**Regla de Proporción [%/n]**
1	30	19.34%
2	0	20%
3	10	10%
4	4	15%
5	2	10%
TOTAL	46	$\sum RP = 74.34\%$

Como se visualiza en la tabla 2, todos los parámetros son tomados con el mismo peso matemático que es dictado por la ecuación:

$$\frac{1}{n} * 100\%$$

siendo el valor de $n = 5$

lo que resulta en un porcentaje equitativo del 20%.

Así mismo se demuestra la suposición en la cual la inexistencia de un parámetro compartido entre ambos textos es tomada en cuenta como una similitud completa, por lo que en la regla de proporción calculada, se estima directamente el 100% de similitud hallada. Posteriormente la similitud se traduce al 20% del total global, atribución que debe ser corregida con el fin de optimizar los resultados del análisis estilométrico aplicado a los textos.

Tabla 3. Demostración de funcionamiento de media ponderada

ID Parámetro	Peso Atribuido	Regla de Proporción	Media Ponderada para la regla de Proporción
1	0.6521	96.67%	63.04%
2	0	0%	0%
3	0.2174	50%	10.86%
4	0.0870	75%	6.52%
5	0.0435	50%	2.17%
TOTAL	1	----	$\sum RPP = 82.6\%$

El peso atribuido esta dado por:

$$\frac{p_i}{\sum_{i=0}^{n_p} p_i}$$

definiendo a p_i como el valor numérico obtenido para un parámetro.

La regla de proporción es la misma que en el caso anterior, pero esta vez un parámetro es ignorado si no existe la similitud común entre los autores y el porcentaje ignorado es redistribuido entre los parámetros restantes.

Una vez que se obtiene este valor, éste es multiplicado por el porcentaje de similitud hallado en la regla de proporción, lo que brinda el porcentaje de similitud basado tanto en la semejanza o acercamiento al parámetro promediado como en la importancia del elemento en cuestión.

A diferencia del caso anterior, en el trabajo de investigación se utiliza una forma diferente para obtener el porcentaje de similitud final, en el cual el significado que tiene

cada uno de los parámetros es independiente y no equitativo. Esta suposición busca maximizar la similitud entre los contenidos de los documentos analizados.

Arquitectura de la aplicación y funcionamiento interno

En esta sección se atiende la arquitectura de la aplicación presentada y utilizada por el autor para realizar las pruebas estilométricas empleadas en la siguiente sección.

De forma general se demuestran cada uno de los procesos que se realizan de forma global, que incluyen a las fases de alimentación de archivos, al análisis global de rasgos característicos predefinidos en la herramienta, a la escritura de los resultados en archivos de texto plano, al análisis comparativo realizado por medio de una media aritmética ponderada y finalmente, a la extracción de los resultados a partir de los análisis realizados en las etapas anteriores.

Alimentación de archivos

Al inicio del programa interviene el manejador de archivos, que se encarga de preguntar al usuario cuantos son los textos a analizar. A partir de este número que es ingresado por el usuario se generan referencias a vectores de datos que contienen vectores del tipo decimal. En este punto también se asigna a cada uno de los archivos un número de identificación para mantener a lo largo del programa un claro entendimiento de lo que sucede con sus datos una vez que son extraídos.

Teniendo conocimiento del número de textos, el hilo principal del programa ingresa en un lazo de control repetitivo del tipo *for*. Más adelante se explica lo que sucede con el mismo.

Análisis general de parámetros

Una vez que han sido realizadas las referencias de los vectores que se utilizarán memoria, se procede a transferir el contenido de los archivos de texto plano. Aquí es utilizada una nueva clase, denominada *"Reader"* puesto que cumple las funciones de un scanner que verifica varios parámetros de escritura.

Para realizar el proceso la clase encargada de analizar los archivos genera vectores con caracteres específicos y expresiones o formas de composición que han sido previamente comprobadas con respecto a su relevancia en el uso del lenguaje expresado de forma escrita.

Posteriormente se llenan con datos los vectores que fueron inicialmente referenciados, con las estadísticas obtenidas al examinar los contenidos del archivo de texto.

Escritura de resultados

Dentro del lazo de control *for* la clase lectora de los archivos genera nombres específicos para cada uno de los parámetros, los mismos que son escritos en archivo de salida compuesto por el nombre del archivo que se analizó con una parte extra concatenada, la que indica el tipo de análisis que cumple.

El archivo escrito se compone de un encabezado que indica a que se refiere cada uno de los identificadores tomados para describir un rasgo particular y el contenido que simplemente es el identificador seguido del número de veces que se encontró una coincidencia importante entre el archivo y los parámetros cargados de forma previa.

Los nombres que son anexados al archivo original, generan diferentes archivos, los mismos que en el directorio se verán de la siguiente forma:

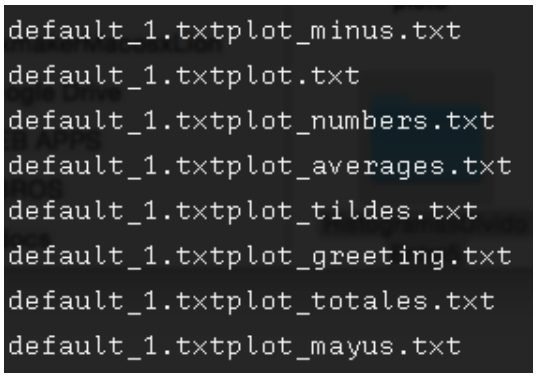

Figura 4. Archivos de almacenamiento para los datos obtenidos

Por medio de estos archivos se procede a realizar las gráficas combinando técnicas de comunicación entre la consola y la utilidad de Gnuplot. Para ello se emplean métodos de awk, lenguaje desarrollado para el procesamiento de datos basado en textos y grep (utilidad de consola de comandos por la existe en sistemas basados en UNIX).

Análisis comparativo por medio de la media ponderada

En este punto, en el que ya han sido extraídos los índices de repetición en los vectores de cada uno de los archivos de los autores, se procede al análisis porcentual de similitudes.

Los contenidos de cada uno de los vectores de los archivos de texto deben ser comparados, por lo que se transfieren a una clase especializada en la comparación de características similares.

Un aspecto esencial de la investigación es que la ausencia compartida de un rasgo entre dos autores no es un indicativo de similitud, puesto que la presencia de rasgos similares no tiene la misma relevancia que la ausencia del mismo. Es por esta razón que se tuvo que implementar, en la clase que realiza el análisis, un cálculo por medio de la media ponderada (Explicación detallada en sección sobre el desarrollo de la herramienta).

Extracción de porcentajes de similitud

Ahora que los vectores de cada uno de los autores han sido comparados, se obtiene la sumatoria total de cada uno de los vectores, los mismos que incrementan su similitud mientras sea menor la diferencia numérica entre las veces que se repite un rasgo, entonces mayor es el índice de coincidencia. Dichas coincidencias que fueron consideradas previamente con respecto a la media ponderada son ahora porcentajes de similitud individuales. Como se puede ver en la sección de resultados, se presentan los porcentajes de similitud con respecto a cada una de las categorías indicadas. Para extraer el resultado final, el programa ha sido indicado que debe sumar todas las categorías y obtener el resultado realizando un promedio global.

ESTIMACIÓN DE SIMILITUD

Tomando como punto de partida los valores que son obtenidos para cada uno de los parámetros analizados en el proceso estilo-métrico, en primer lugar se determina almacenarlos en una estructura de datos tipo vector. El vector es definido con el enfoque que es otorgado por la matemática y la programación, estructuras conceptualmente equivalentes.

Por lo que se define el vector:

$$\overrightarrow{T_i} = \begin{pmatrix} t_1 \\ t_2 \\ \cdots \\ \cdots \\ t_n \end{pmatrix}$$

que contiene los parámetros de un $texto_i$, para el autor A.

Con los valores de $t_i \in \mathbb{R}^n$.

Esto compone la matriz T:

$$T = \begin{pmatrix} \overrightarrow{T_1} & \overrightarrow{T_2} & \overrightarrow{T_3} & \overrightarrow{T_4} & \overrightarrow{T_5} \end{pmatrix}$$

A partir de estos vectores se extrae el promedio para cada fila o registro del vector, en este caso con $n = 5$ textos para obtener el vector promedio del autor A.

$$\vec{A} = \begin{pmatrix} a_1 \\ a_2 \\ \cdots \\ \cdots \\ a_n \end{pmatrix}$$

Es decir, para cada entrada a_i se tiene:

$$a_i = \frac{1}{n}\sum_{i=1}^{n} t_i$$

A diferencia del vector $\vec{A}$, el vector de valores vinculados al texto redactado por el autor B, simplemente contiene los parámetros de dicho texto.

$$\vec{B} = \begin{pmatrix} b_1 \\ b_2 \\ \dots \\ \dots \\ b_n \end{pmatrix}$$

Con la existencia de los vectores $\vec{A}$ y $\vec{B}$, ahora se establece la relación de distancia entre ambos vectores que es dada por la ecuación:

$$d_i = \frac{1}{n} \sum_{i=1}^{n} L_i$$

en la cual el valor de L_i esta dado por:

$$L_i = \begin{cases} \dfrac{|a_i - b_i|}{a_i} & , si \ \ |a_i - b_i| < a_i \\ 0 & , si \ \ |a_i - b_i| \geq a_i \end{cases}$$

resultando en el vector de distancias $\vec{D}$

$$\vec{D} = \begin{pmatrix} d_1 \\ d_2 \\ \dots \\ \dots \\ d_n \end{pmatrix}$$

El proceso de estimación de similitud entre los vectores, se basará en el vector resultante de distancias, con diferentes consideraciones según el concepto que se aplique. En la investigación se presentan cuatro métodos principales que incluyen a distancias definidas por equivalencias porcentuales no ponderadas o ponderadas y también las distancias definidas de forma estricta por la matemática como son las distancias de Minkowski y Chebyshev. Cada método se explica en su totalidad en las subsecciones siguientes.

Distancia con regla de proporción simple

El primer acercamiento al concepto de similitud entre los dos vectores se basa exclusivamente en la aplicación de la regla de proporción, por lo que estrictamente no se refiere a la distancia como indica su definición matemática. Para obtener un porcentaje de similitud se realizan los cálculos de la siguiente forma:

A partir de la obtención de distancias se recuerda el vector $\vec{D}$:

$$\vec{D} = \begin{pmatrix} d_1 \\ d_2 \\ \dots \\ \dots \\ d_n \end{pmatrix}$$

Se estima la distancia inversa representado la cercanía al vector original

$$\vec{C} = \begin{pmatrix} 1 - d_1 \\ 1 - d_2 \\ \dots \\ \dots \\ 1 - d_n \end{pmatrix}$$

Como se visualiza en la tabla #, todos los parámetros son tomados con el mismo peso matemático que es dictado por la ecuación:

$$\vec{E} = \begin{pmatrix} \left(\frac{1}{n}\right) * 100 \\ \left(\frac{1}{n}\right) * 100 \\ \dots \\ \dots \\ \left(\frac{1}{n}\right) * 100 \end{pmatrix}$$

Realizando el producto entre cada una de las entradas de los vectores $\vec{C}$ y $\vec{E}$, se halla el vector de similitudes $\vec{S}$, se obtiene $\vec{S} = \vec{C} * \vec{E}$

$$\vec{S} = \begin{pmatrix} c_1 * e_1 \\ c_2 * e_2 \\ \dots \\ \dots \\ c_n * e_n \end{pmatrix}$$

Finalmente la sumatoria de todas las entradas del vector $\vec{S}$ representará el porcentaje de similitud existente entre los vectores $\vec{A}$ y $\vec{B}$ en cuestión.

$$Similitud\ porcentual = \sum_{i=1}^{n} s_i$$

Distancia con regla de proporción ponderada

Otro acercamiento que sigue la conceptualización de distancia para hallar similitudes entre los dos vectores es la media ponderada aplicada sobre la regla de proporción incluyendo el vector de distancias que existen entre ambos.

Recordando el vector $\vec{C}$ que representa la cercanía al vector original

$$\vec{C} = \begin{pmatrix} 1 - d_1 \\ 1 - d_2 \\ \dots \\ \dots \\ 1 - d_n \end{pmatrix}$$

A cada punto se le asigna un peso dado por la importancia que tiene el parámetro, lo que define el siguiente conjunto vectorial:

$$\vec{W} = \begin{pmatrix} w_1 \\ w_2 \\ \dots \\ \dots \\ w_n \end{pmatrix}$$

en el mismo que w_i es:

$$w_i = \frac{p_i}{\sum_{i=0}^{n_p} p_i}$$

con p representando al valor acumulado promedio específicamente de un solo parámetro dentro de las entradas del vector analizado.

Con estos valores en hallados, esta vez se busca el valor de la media ponderada que esta vez mantendrá una distribución no equitativa entre los parámetros analizados en el sistema, estableciendo:

$$\bar{X} = \frac{\sum_{i=1}^{n} c_i \times w_i}{\sum_{i=1}^{n} w_i}$$

Dicha fórmula de forma expandida se simplifica a:

$$\bar{X} = \frac{c_1 w_1 + c_2 w_2 + c w_3 + \cdots + c_n w_n}{w_1 + w_2 + w_3 + \cdots + w_n}$$

Nótese que siempre se toma la suposición matemática que $c_i, w_i \geq 0$.

De la naturaleza de esta fórmula surge un sistema mediante el cual se pudo controlar la relevancia de la existencia o no existencia de una misma característica dentro de textos que provienen de diferentes autores. Además queda expresado de forma matemática que cuando el peso *w* cumple la condición en que

$$w_i = 0$$

entonces el producto dado entre los elementos es:

$$x_i w_i = 0$$

con una influencia de peso 0 por lo que se puede decir que la relevancia del parámetro es anulada tanto en el numerador, así como en el denominador.

Aquí radica la importancia del método empleado que permite descartar aquellos elementos que no demostraron tener similitud entre sí.

Finalmente se obtiene la similitud de la siguiente forma:

$$Similitud\ porcentual = \sum_{i=1}^{n} \overline{x_i}$$

Con estas consideraciones la métrica emplea el concepto de distancia con respecto a los puntos que definen un vector, con la finalidad de permitir un acercamiento más significante cuando los valores de un mismo parámetro son semejantes o cercanos a aquel establecido como referencia. En este caso particular, la herramienta considera como referencia al vector promedio extraído a partir de diferentes textos redactados por un mismo autor A.

Distancia Minkowski

Otra métrica que se puede emplear para calcular la similitud entre dos vectores, en este caso los vectores resultantes del estudio de un promedio para el Autor A y el resultado de totales de un Autor B, es buscar la similitud entre ambos aplicando el concepto de distancia. Con el fin de generalizar los cálculos y estimaciones de similitud por medio de las distancias, se aplica la distancia de Minkowski, métrica representativa de la generalización de las distancias Euclidianas o Manhattan. Adicionalmente, esta distancia tiene la característica de ser una métrica normalizada en el espacio vectorial.

La distancia de Minkowski de orden p para dos vectores posee la siguiente definición matemática:

$$\left(\sum_{i=1}^{n} |a_i - b_i|^p\right)^{1/p}$$

De igual forma, por definición de la desigualdad de Minkowski, el valor de p es métrico siempre que sea mayor o igual a uno.

$$\forall\, p \geq 1 \Longrightarrow Valor\ Métrico$$

Por lo tanto, los valores para $p < 1$ no son aplicables para ser utilizados con la definición presentada sobre distancia.

Distancia Chebyshev

De forma análoga a la distancia de Minkowski, se analiza el caso particular en donde el valor de la distancia está definido por:

$$\lim_{p\to\infty}\left(\sum_{i=1}^{n}|a_i - b_i|^p\right)^{\frac{1}{p}}$$

Con esta consideración, la métrica resultante es conocida como la distancia de Chebyshev, que al ser establecida como el límite cuando el valor de $p \to \infty$ los valores menores al valor máximo son considerados como insignificantes por lo que la expresión anterior se transforma en la siguiente ecuación de forma reducida:

$$D_{Chebyshev}(A, B) = max(d_i)$$

recordando que:

$$d_i = |a_i - b_i|$$

Por lo que la métrica representa únicamente a la distancia mayor en magnitud que existe entre un par de coordenadas semejantes, excluyendo distancias menores que se encuentran en coordenadas diferentes a la tomada en cuenta.

RESULTADOS

Los resultados que se presentan en esta sección han sido obtenidos después de haber realizado una serie de pruebas con documentos digitales de varias categorías, que incluyeron emails, resúmenes académicos y redacción libre de 35 distintos autores. Dado que los autores no conocían la finalidad del experimento, se asume que no hicieron ninguna consideración especial en el momento de la redacción de los contenidos de los documentos electrónicos. Con esta consideración se evidencia que los estilos lingüísticos de escritura no han sido modificados.

Pruebas

Las pruebas que se muestran a continuación se basaron en textos reales que fueron escritos por diferentes autores, los mismos que establecían comunicación dentro del ámbito académico relacionado a trabajos, notas, exámenes, deberes, puntos de vista, resúmenes, entre otros. Para conseguir un acercamiento con mayor exactitud en la obtención de los porcentajes de similitud, se ingresaron cinco texto de un mismo autor, que fueron utilizados para extraer el patrón de escritura y posteriormente compararlo con un documento informático ajeno, es decir redactado por otra persona. Para comprender los que sucede con las estadísticas los resultados se presentan en dos grupos de análisis, en los que los gráficos iniciales muestran coincidencias sobre un mismo parámetro, con su barra de error y posteriormente los resultados de la prueba de similitud con dos puntajes: *High Score Value* y *Low Score Value*. Siendo *High Score Val*ue la métrica que toma en cuenta la división entre la sumatoria de todos los resultados para el número de elementos; y, *Low Score Value* que únicamente toma en cuenta el resultado

de un grupo de parámetros cuando ha la semejanza es diferente a cero, aplicando la misma división mencionada con anterioridad.

$$\begin{cases} High\ Value\ Score = \dfrac{\sum_{i=0}^{n} \bar{x}_i}{m} \\ Low\ Value\ Score = \dfrac{\sum_{i=0}^{n} \bar{x}_i}{n} \end{cases}$$

donde: $m = \#\ de\ promedios\ \neq 0$ y $n = \#\ de\ promedios$

Nótese que puesto a que $m \leq n$, entonces se cumple la condición:

$$Low\ Value\ Score\ \leq High\ Value\ Score$$

Tomando en cuenta ambos puntajes se busca maximizar la información entregada al usuario en relación a la similitud de los textos, con el fin de realizar una conclusión sobre la autoría del texto, que sea la más aproximada a la realidad.

Análisis de textos

Para la prueba se diseñó al sistema con la capacidad de extraer datos de la estilometría a partir de cinco texto previos que se conocían que fueron redactados por un mismo autor (Autor A) de forma natural. Es decir, desconocían que sus textos iban a ser analizados por el programa para posteriormente poder identificarlos mediante pruebas comparando textos propios o ajenos al autor. En las siguientes imágenes se identifica al autor de varios textos como “Known Author” (Autor A), y el autor con un solo texto sobre el cual se extrae la similitud, como “Unknown Author” (Autor B).

Debido a que se conocía que los múltiples textos pertenecían a un solo autor se pudo calcular también el error de la medición, dado por la siguiente expresión matematica:

Sea la muestra *M* representada por el conjunto:

$$M\ = \{m_1, m_2, m_3, \dots, m_n\}$$

Entonces la desviación estándar está dado por:

$$desvEst = \sqrt{\frac{\sum_{i=1}^{n}(\bar{x}_i - m_i)^2}{n}}$$

siendo *n* el número total de muestras.

Se debe saber que ahora la media es calculada sobre las medias ponderadas individuales, obtenidas con anterioridad:

$$\bar{X} = \frac{\sum_{i=1}^{n} x_i}{n}$$

También conocido como la desviación estándar que en este caso ha sido empleada para facilitar la lectura y comprensión de los datos por medio de una inspección visual sencilla.

Análisis de textos pertenecientes a autores distintos

El primer análisis que se realizó, consiste en una inspección simple sobre los caracteres seleccionados por los autores en su lenguaje expresado de forma escrita, con la finalidad de identificar sus preferencias que incluyen los signos de puntuación, signos de agrupación, caracteres matemáticos y caracteres no comunes.

A continuación se pueden apreciar los resultados, en las figuras 5 a la 15:

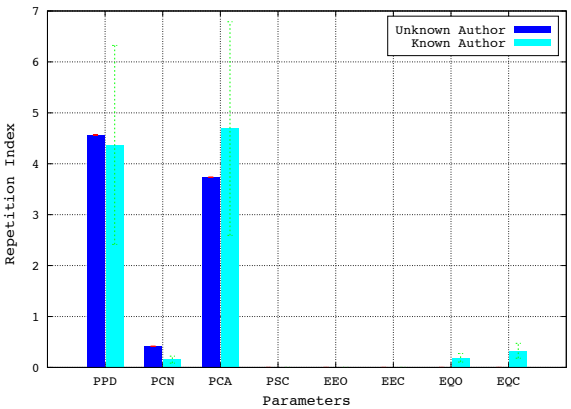

Figura 5. Signos comunes y número de incidencias, para textos de autores diferentes.

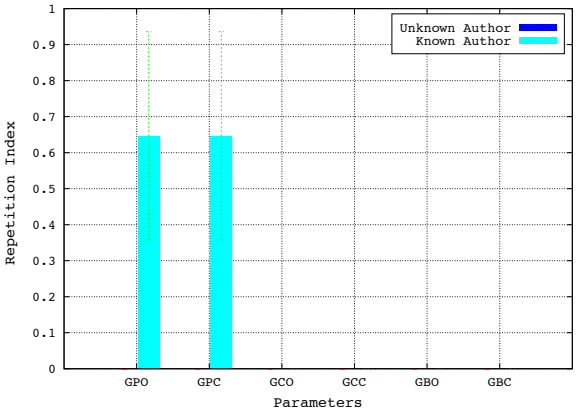

Figura 6. Signos de agrupación y número de incidencias, para textos de autores diferentes.

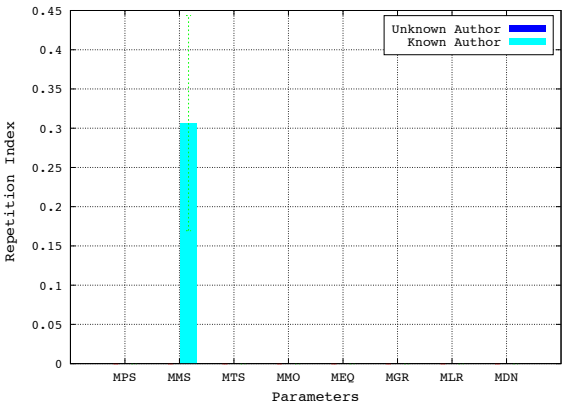

Figura 7. Signos matemáticos y número de incidencias, para textos de autores diferentes.

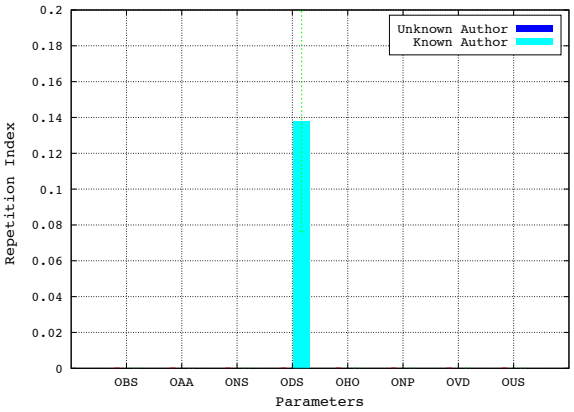

Figura 8. Signos no comunes y número de incidencias, para textos de autores diferentes.

A continuación la aplicación del sistema diseñado toma en cuenta los atributos relacionados a caracteres lingüísticos, esto es la utilización de tildes tanto en letras mayúsculas como minúsculas y la tendencia que tiene el autor para expresar los números en palabras o como dígitos.

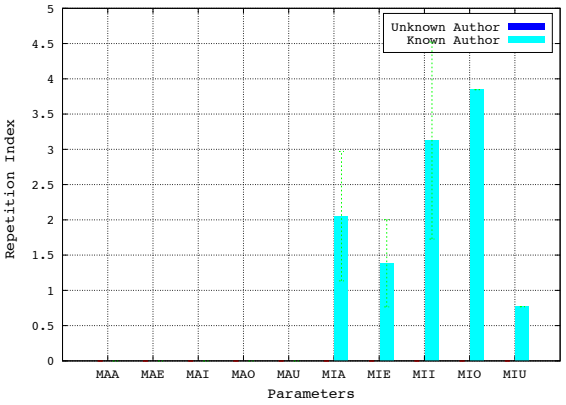

Figura 9. Acentuación en minúsculas y minúsculas con su número de incidencias, para textos de autores diferentes.

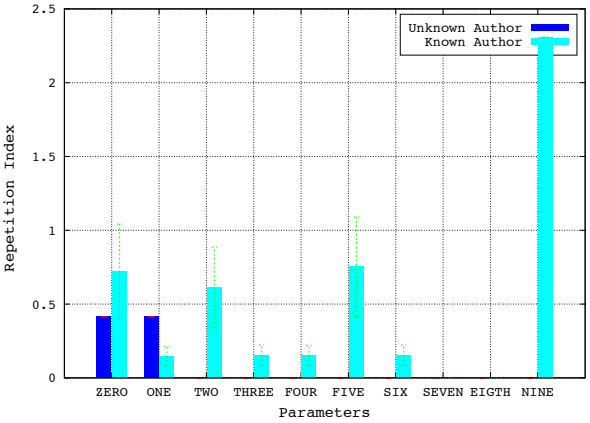

Figura 10. Preferencia escrita o numérica con número de incidencias, para textos de autores diferentes.

Además de los análisis presentados en las figuras anteriores, también se realizó una revisión sobre la distribución de las letras mayúsculas y minúsculas en todo el documento. Los gráficos correspondientes fueron omitidos debido a que su relevancia ha probado ser muy baja, pero aún debe ser considerada como un parámetro válido.

Una vez que se realizaron estos análisis, se procede al estudio léxico y sintáctico, como indican las figuras 11 a 14.

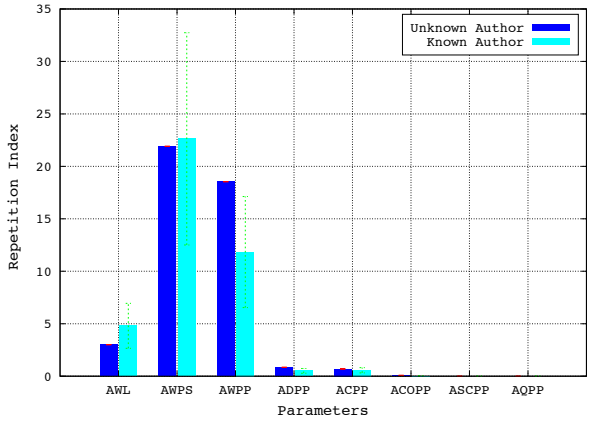

Figura 11. Promedios léxicos y número de incidencias, para textos del mismo autor, parte 1.

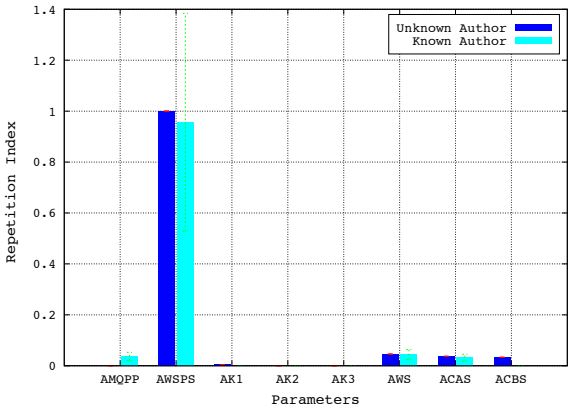

Figura 12. Promedios léxicos y número de incidencias, para textos del mismo autor, parte 2.

.

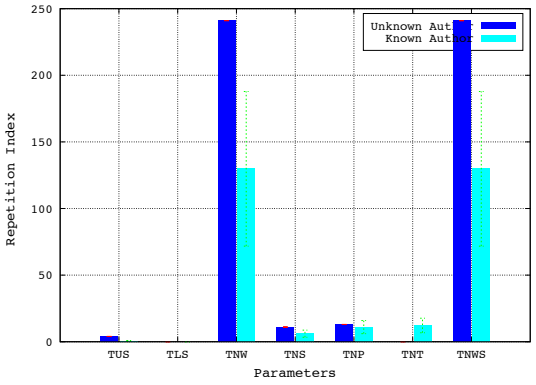

Figura 13. Totales léxicos y número de incidencias, para textos de autores diferentes, parte 1.

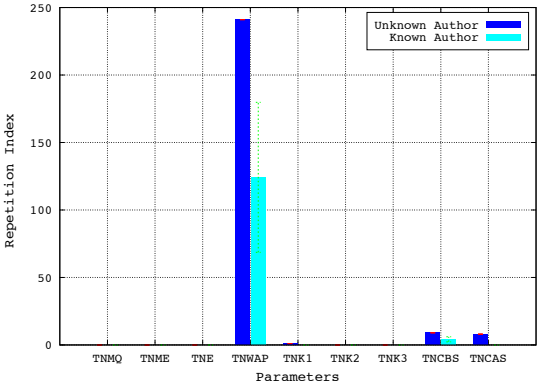

Figura 14. Totales léxicos y número de incidencias, para textos de autores diferentes, parte 2

Estos resultados han calculado cada uno de los parámetros tanto en totales como en promedios obtenidos con respecto a la cantidad total de palabras en un documento. Finalmente se realiza una última verificación para conocer si el texto de los autores posee similitudes en palabras clave de saludo o despedida, los mismos que son utilizadas con mucha frecuencia en documentos informáticos.

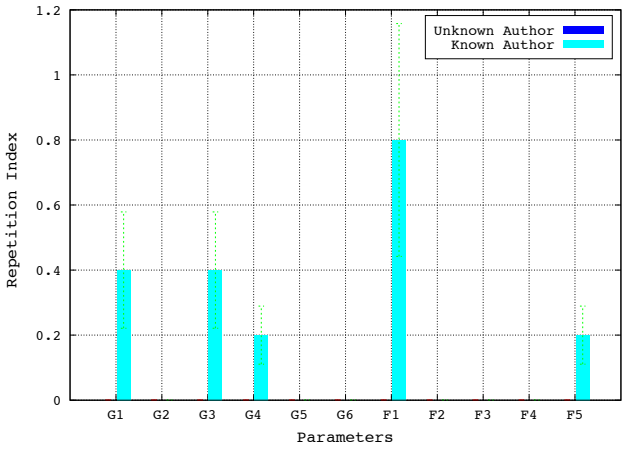

Figura 15. Presencia de saludos y con su número de incidencias, para textos del mismo autor

Después de obtener la media aritmética de los atributos analizados (ver Anexo A), el resultado del estudio proporciona los porcentajes de similitud como se detalla a continuación:

```
######      ######      ######      ######

High value score: 51.9734% over 100%

Low  value score: 38.98% over 100%

######      ######      ######      ######
```

Figura 16. Porcentaje de similitud estimado para textos de autores diferentes.

A partir de los datos finales entregados por la herramienta, se puede estimar que los textos no presentan una similitud significativa porque la similitud obtenida en puntuación alta es cercana al 51.95% y la puntuación baja es de 38.98%. Dichos resultados sugieren que los autores son distintos puesto que se presenta la carencia de similitud tanto gráfica como porcentualmente después de haber realizado una comparación intensiva sobre rasgos considerados.

De igual forma que se ha analizado el resultado de similitud de un solo autor, se realiza el análisis global de resultados. Estos son presentados en la figura 17 y 18 por medio de histogramas.

Como se observa en la figura, es sencillo verificar que la mayor cantidad de datos del análisis presentaron similitudes entre 50% y 55% para el puntaje alto.

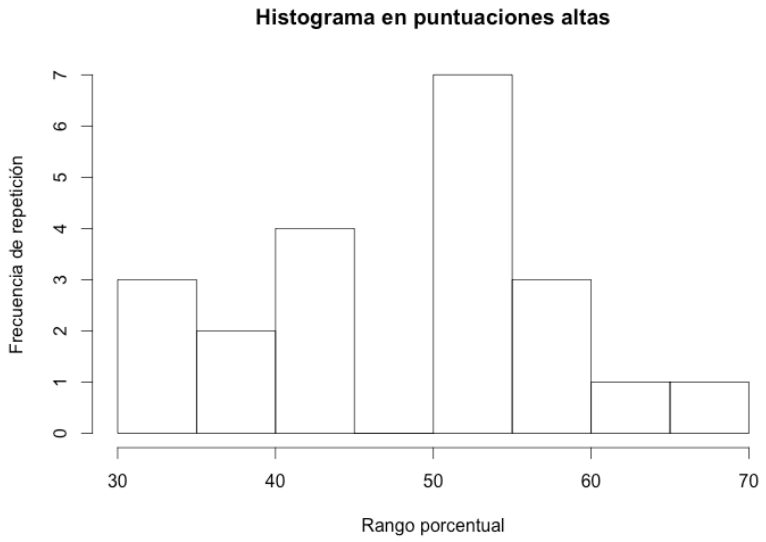

Figura 17. Histograma Puntajes Altos, autores diferentes

A continuación, también se presenta el histograma de las puntuaciones bajas en el análisis, que esta vez indican una variación y tendencia ente en 30% y 35% cuando los textos pertenecen a diferentes autores.

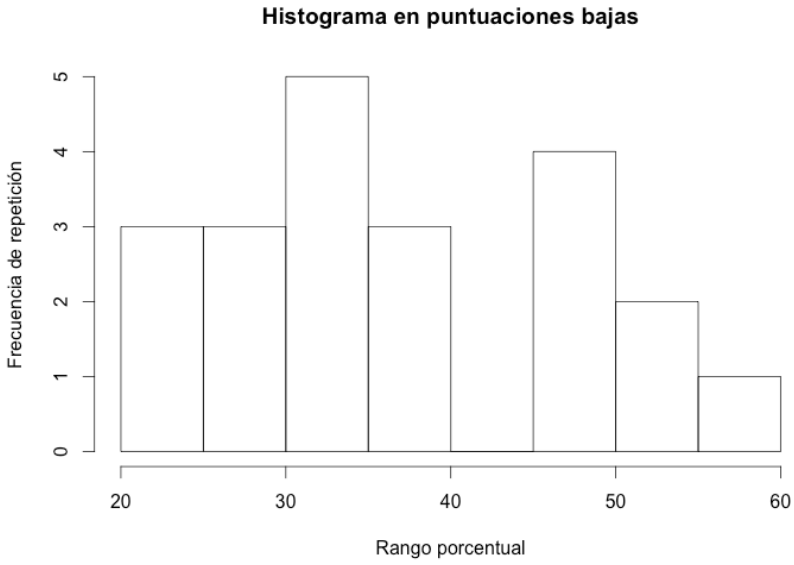

Figura 18. Histograma Puntajes Bajos, autores diferentes

A más de los histogramas, el cálculo de las estadísticas sobre los resultados son presentados en la tabla 4:

Tabla 4. Datos estadísticos sobre las distribuciones de resultados

Dato Estadístico	High Score [%]	Low Score [%]
1er Cuartil	40.79	28.68
3er Cuartil	54.83	46.58
Mediana	51.60	33.69
Media Aritmética	48.16	36.85

Análisis de textos pertenecientes a un mismo autor

En igual forma que los resultados presentados recientemente, se considera la misma agrupación para los caracteres empleados, los que son indicados en las figuras 19 a 29 de la subsección.

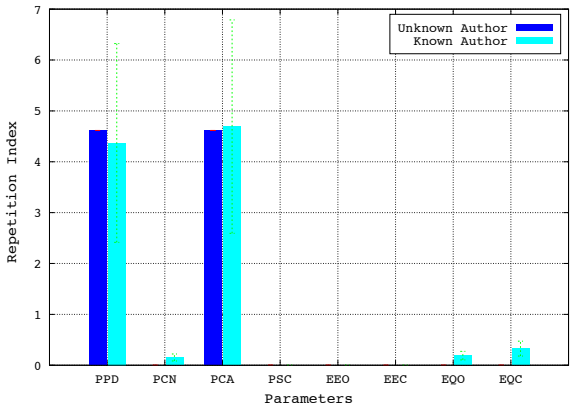

Figura 19. Signos comunes y número de incidencias, para textos del mismo autor.

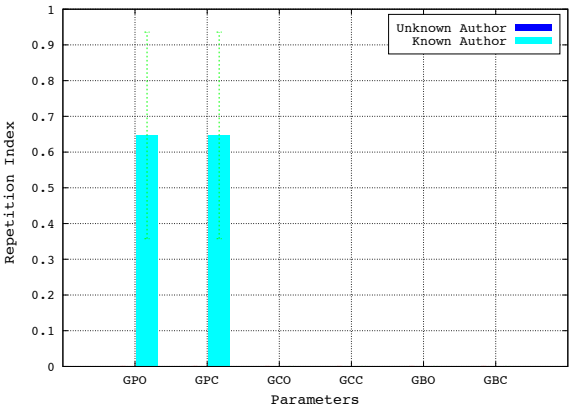

Figura 20. Signos de agrupación y número de incidencias, para textos del mismo autor.

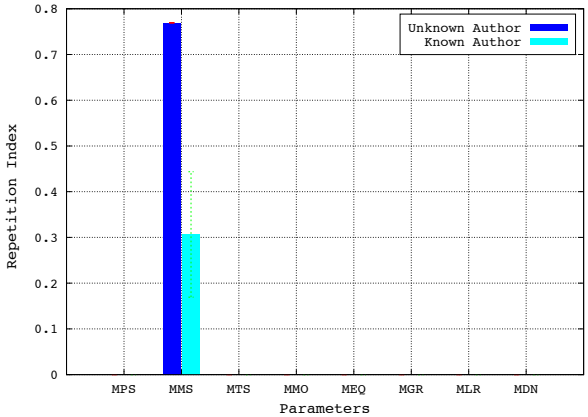

Figura 21. Signos matemáticos y número de incidencias, para textos del mismo autor.

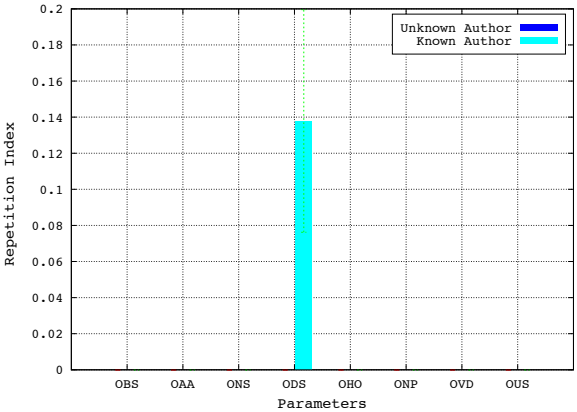

Figura 22. Signos no comunes y número de incidencias, para textos del mismo autor.

Paso seguido, se consideran las categorías relacionados a caracteres lingüísticos y preferencia de expresión numérica.

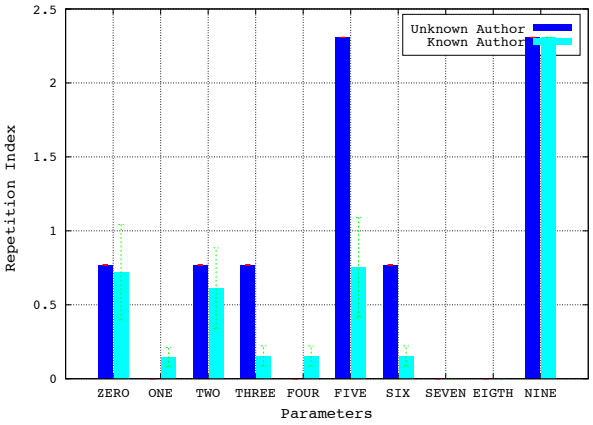

Figura 23. Acentuación en minúsculas y minúsculas con su número de incidencias, para textos del mismo autor.

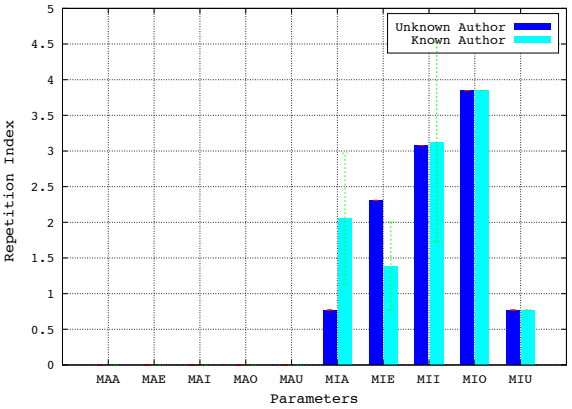

Figura 24. Preferencia escrita o numérica con número de incidencias, para textos del mismo autor.

Nuevamente se procede a extraer los datos relacionados a los métodos sintácticos y léxicos, como se aprecia en las figuras 25 a 28.

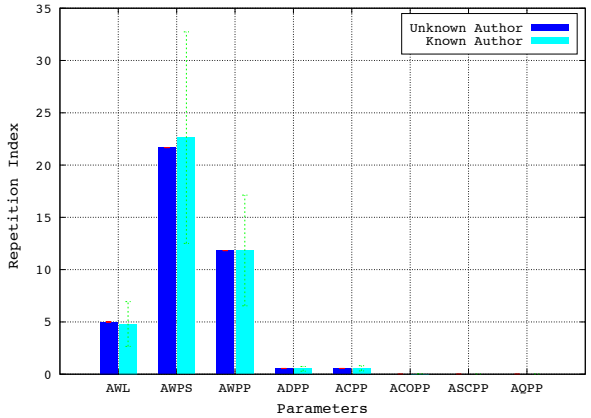

Figura 25. Promedios léxicos y número de incidencias, para textos del mismo autor, parte 1

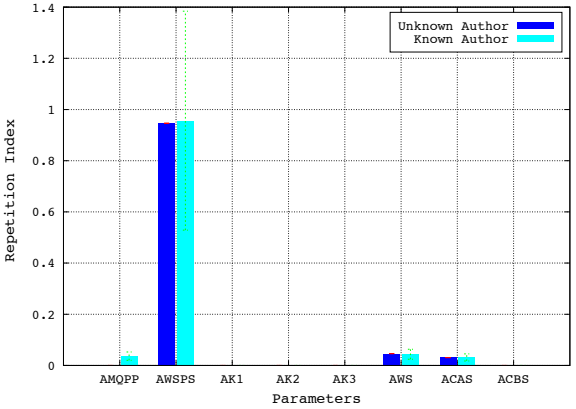

Figura 26. Promedios léxicos y número de incidencias, para textos del mismo autor, parte 2

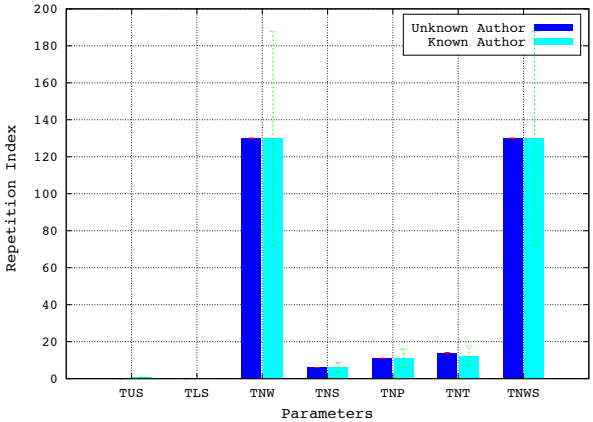

Figura 27. Totales léxicos y número de incidencias, para textos del mismo autor, parte 1

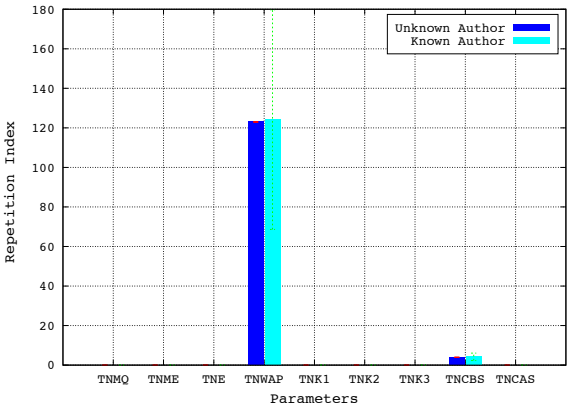

Figura 28. Totales léxicos y número de incidencias, para textos del mismo autor, parte 2

Finalmente se consideran a los atributos semánticos que son influenciados directamente por la selección y preferencia de palabras y estructuras comúnmente aplicadas en el lenguaje escrito utilizado en Latinoamérica.

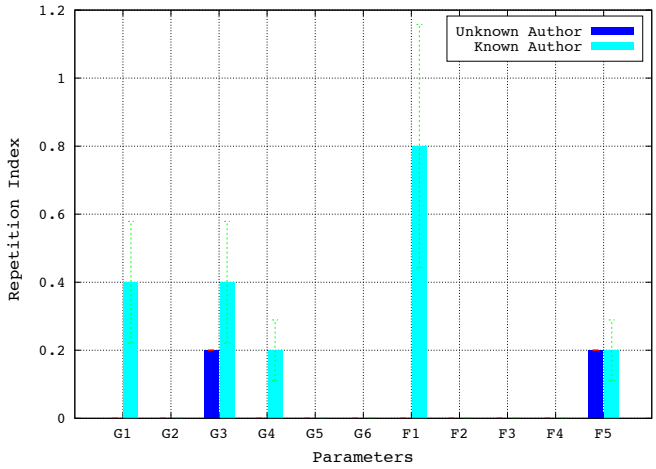

Figura 29. Presencia de saludos y con su número de incidencias, para textos del mismo autor

Todos los parámetros se incluyeron para lograr el resultado final, que indica los porcentajes de similitud presentados por la consola de la herramienta:

```
######     ######      ######      ######
High value score: 86.6739% over 100%
Low  value score: 86.6739% over 100%
######     ######      ######      ######
```

Figura 30. Porcentaje de similitud estimado para textos del mismo autor.

Esta vez los resultados indican un puntaje alto y bajo de 86.67% lo que indica una gran cercanía entre los documentos analizados. Esta es la razón por la cual se puede considerar que ambos textos tienen una gran probabilidad de provenir de un mismo escritor.

Nuevamente se presentan los resultados globales, en este caso de la comparación de documentos pertenecientes al mismo autor. Se determinan los siguientes valores estadísticos:

Tabla 5. Datos estadísticos sobre las distribuciones de resultados

Dato Estadístico	High Score [%]	Low Score [%]
1er Cuartil	75	74
3er Cuartil	77	77
Mediana	76	76
Media Aritmética	77.8	77.6

Por lo que se estima que, para el autor del que se conocen los patrones y características únicas, existe el valor de rigurosidad mínima que indica 77.7% en promedio.

Por lo que cuando se emplea la media aritmética ponderada y el resultado del análisis de similitud sobrepasa este valor, entonces se puede realizar la atribución del texto al autor previamente conocido.

Aplicando los distintos métodos matemáticos para obtener el porcentaje de similitud, compuesto por todos los valores obtenidos de los atributos analizados en el texto, se ha obtenido la tabla 6.

Tabla 6. Resumen de porcentajes de similitud promedio, obtenidos para cada método matemático.

	Dato Estadístico		
Método Aplicado	**Mismo Autor [%]**	**Autores Distintos [%]**	**Intervalo de decisión [%]**
Media Aritmética	92.7	83.4	9.3
Media Aritmética Ponderada	77.8	36.9	40.9
Distancia Minkowski	89.6	65.8	23.8
Distancia Chebyshev	88.2	83.4	4.8

A partir de la tabla anterior, se puede observar la diferencia que existe cuando se busca la similitud entre textos de autorías diferentes y textos de la misma autoría. La diferencia hallada entre ellos ha sido llamada intervalo de decisión, y es un valor con el que se puede determinar la efectividad que poseen los métodos, puesto que mientras mayor es dicho valor entonces el método posee un mayor rango de valores que diferencian los resultados cuando se trata del análisis de documentos que provienen del mismo u otro autor.

Los mismos resultados se pueden apreciar en forma de cuadro de barras, como presenta la figura a continuación.

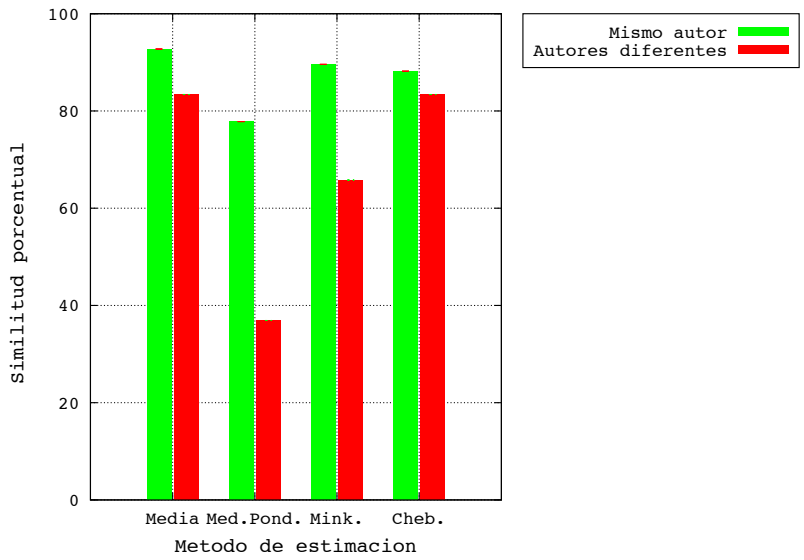

Figura 31. Resumen en cuadro de barras para porcentajes de similitud obtenidos en las pruebas.

Es evidente que la en figura anterior se refleja lo discutido anteriormente sobre la diferencia que existe cuando se analizan textos de un mismo autor y textos de diferentes autores.

DISCUSIÓN

A partir de los resultados demostrados surge la observación de que la existencia común de un rasgo característico entre el contenido de dos documentos informáticos es un indicador válido de similitud, puesto que demuestra una preferencia a una forma de expresión. Mientras que la ausencia de este rasgo lingüístico, no es un indicador debido a que no representa dicha tendencia.

Esta observación que se demuestra por medio del uso de la herramienta es de gran utilidad en la estilometría porque indica que los métodos tradicionales de extracción de similitud, que emplean distancias, pueden generar una desviación de los resultados.

Es por esta razón que la herramienta enseña el uso de la media aritmética ponderada, que otorga un peso matemático distinto, según la relevancia de cada una de las características analizadas y que se pueden verificar en la sección de anexos (Anexo A).

Otra observación que surge a partir de la arquitectura, que no fue diseñada con el propósito de ser un software predictivo, se presenta cuando el usuario de la herramienta se pregunta sobre cuál sería el parámetro que se debe considerar en el momento final en el que se atribuye una redacción de un texto a la persona. La respuesta a esta incógnita es que se debe considerar el promedio de los porcentajes de similitud y el comportamiento que presentan los datos.

Para la atribución de un documento a un autor, se debe realizar varias corridas de la herramienta para verificar un umbral mínimo que tiende a ser cercano a 66% en puntuación baja y al 74% en puntuación alta, que da una precisión matemática de 72%.

Adicionalmente, es de mucha importancia aclarar al lector de esta investigación que el estilo de redacción no es inmutable ni mucho menos constante, porque el ser humano adapta su composición a experiencias previas o las etapas de la vida, con el fin de

mejorarla en cohesión, concisión y coherencia. Por lo que siempre que se realiza un análisis estilo-métrico se recomienda poner especial importancia en el intervalo de tiempo al cual pertenecen los contenidos de los documentos que son tomados en cuenta para dicho análisis.

Esto es indicado de forma puntual en el trabajo de Fazli [4], donde se analizan tres parámetros los que incluyen la frecuencia de las palabras, longitud de caracteres y palabras más frecuentes con los que se demuestra la existencia de cambios en el estilo de escritura con respecto al tiempo. Adicionalmente, el autor menciona que de ninguna manera son estos tres marcadores usados los únicos que pueden ser usados para medir una variación en el estilo con el tiempo [4] .

De todas formas, esta consideración no puede ser pasada por alto al estudiar la redacción de un documento.

TRABAJO FUTURO

En el trabajo a futuro de la presente investigación se propone adicionar más parámetros y métodos de análisis que vayan más allá del lenguaje Español y las reglas dictadas por la Real Academia de la Lengua, con las palabras y expresiones que son utilizadas de forma coloquial por la región en donde se desee aplicar el programa con la extracción de métricas enfocadas en el estilo de expresión lingüístico utilizado en la escritura para otorgar la autoría de un texto a una persona específica. También se debe tomar en cuenta que la estructura de los textos son dependientes del contexto en el que fueron redactados como sucede en los blogs de discusión, emails dirigidos dentro de una empresa, o trabajos con fines académicos, jurídicos, entre otros.

Con el objetivo de brindar escalar el funcionamiento de la herramienta a otro nivel, se buscará implementar la recolección de los patrones de escritura resultantes en una base de datos para que las comparaciones puedan ser realizadas entre centenas de autores y se atribuya un porcentaje de similitud con respecto a uno de los autores almacenados, que evidentemente debe tener la mayor probabilidad de ser acertada. Si bien se busca escalabilidad en la aplicación, se tiene que realizar una observación interesante que proviene de trabajos previos, en los que "se ha demostrado que el autor de un texto de prueba a menudo puede ser determinado mediante análisis manual o automatizado del estilo de escritura, pero únicamente cuando el autor es conocido dentro de un pequeño conjunto de posibilidades (hasta 300)" [17] Aquí se puede observar una limitación de la herramienta y de la estilometría, puesto que se limita su capacidad a tres centenas de patrones identificados que pueden coexistir en la memoria de una máquina para otorgar los resultados más exactos.

Adicionalmente, en la herramienta que fue desarrollada, se deseará incrementar la cantidad de textos para extraer rasgos característicos de autores, pero sin olvidar que siempre existe la presencia de asíntotas matemáticas en el aprendizaje en máquinas, por lo que es esencial encontrar el valor que debe ser usado como límite para maximizar la exactitud y precisión de la herramienta.

Con respecto a los métodos que se creerían convenientes para adicionar en la aplicación se destacan los siguientes:

- Análisis Ortográfico. Por este medio es posible determinar el nivel de formalidad y educación del autor que está siendo fuente para la extracción de datos estilográficos. Con la ortografía se podría evidenciar la utilización de palabras que no existen definidas en el diccionario del idioma y palabras que son propias de un dialecto al que se ha adaptado una persona.
- Análisis Gramatical. Este análisis viene a ser superior al sintáctico, puesto que incluye al mismo. Con ayuda de un análisis gramatical en relación a las leyes dictadas por la lengua propia, se podrían notar patrones repetitivos dentro de los textos redactados por un mismo autor.
- Análisis sobre el contenido y temas relacionados en el cuerpo del mensaje digital. Es un aspecto de importancia porque por medio del mismo se pueden detectar las áreas de conocimiento del autor, puesto que su vocabulario será más amplio en pocas áreas.
- Análisis basado en el reconocimiento de sustantivos, sintagmas nominales (frases nominales), y frases preposicionales, tal como son propuestas en los resultados presentados por el árbol de análisis gramático implementado en el Stanford Parser. [17]

Finalmente el lector de este trabajo y usuario de la herramienta debe considerar que hasta la fecha, “aproximadamente 1000 marcadores del estilo de redacción han sido identificados. Se debe buscar la identificación de todos y cada uno de los marcadores que componen un estilo; para mapear al estilo en la misma forma que los biólogos han mapeado los genes.” [22]. En esto, se demuestra que , la estilometría representa un gran reto para la informática porque requiere de una gran cantidad de variantes y cualidades para aplicar los estudios del estilo en un lenguaje específico, como fue en esta investigación el lenguaje Español, utilizado en la ciudad de Quito, capital del Ecuador.

Se debe tomar en cuenta que la estilometría es una parte necesaria en el mundo actual, debido a la facilidad que presenta el mundo del internet para el anonimato.

CONCLUSIONES

La investigación, motivo de este estudio, nace con el fin de aplicar la estilometría para determinar los indicativos que definen el estilo de escritura de un autor y, de esta forma, atribuir su autoría a textos que tienen presentes las cualidades que forman un patrón de redacción. La estilometría es la exploración de cualidades lingüísticas que se encuentran presentes en un documento escrito, ya sea de carácter investigativo, académico o coloquial.

Una vez que se realiza la caracterización del autor se procede a la identificación de la autoría con el objeto de aclarar la procedencia y la legalidad del mismo.

A diferencia de investigaciones previas, en este trabajo, se ha realizado la integración de varios métodos desarrollados por diferentes investigadores que han señalado la importancia que tienen los análisis de carácter léxico, semántico y sintáctico y otras relacionadas a la existencia, frecuencia y distribución de caracteres, expresiones, estructuras y temas discutidos en el contenido de un archivo digital.

Esta es una lista que a pesar de ser amplia, no incluye todos los parámetros que surgen para la identificación de una forma de escribir, porque los atributos siguen creciendo mientras se identifican nuevas características, que también dependen de cada idioma. Incluso se ha comprobado que el estilo cambia con variables externas como son el ambiente, época y contexto, así como cualidades de personalidad del autor como la edad, madurez y nivel cultural.

Adicionalmente, se demuestra que para realizar una estimación estilométrica, no es un requisito utilizar complejas rutinas de aprendizaje de máquinas tales como las redes neurales (NN), Supporting Vector Machines(SVN) o métodos de clasificación que incluyen al

vecino más cercano o árboles de decisión. Por lo que se comprueba que para aplicar la estilometría simplemente se requiere de una aproximación matemática, realizada con estimadores efectivos.

Es importante recalcar que cuando se presenta el método de estimación matemática es de mucha importancia la ponderación que tiene cada parámetro en el resultado global. La herramienta demuestra esto mediante el uso de la media aritmética ponderada dándole mayor eficiencia con respecto a los otros métodos presentados que incluyeron la media aritmética simple y conceptualizaciones generales de distancias.

También se toma en cuenta que mientras existen más rasgos de estilo presentes en un documento y mientras crece la cantidad de estimadores, la confiabilidad que se tiene en los indicadores se incrementa.

El desarrollo de la presente investigación será un aporte importante para diferentes áreas puesto que puede ser aplicada con versatilidad en varios campos porque cubre desde escritos ilegítimos simples hasta problemas relacionados con la seguridad nacional, que pueden causar serios problemas al no ser identificable su autor.

REFERENCIAS BIBLIOGRÁFICAS

[1] Abbasy, A. & Chen, H. (2005). *Applying Autorship Analysis to Extremist-Group Forum Messages.* Intelligent Systems, IEEE (Impact Factor: 1.92). 10/2005; 20(5):67 - 75. DOI: 10.1109/MIS.2005.81 Obtenido el 2 de mayo 2015 de: http://www.researchgate.net/publication/3454295_Chen_H._Applying_Authorship_Analysis_to_Extremist_Group_Web_Forum_Messages._IEEE_Intelligent_Systems_20(5)_67-75

[2] Ali, N. (2014*). Text stylometry for chat bot identification and intelligence estimation.* University of Louisville. ThinkIR: The University of Louisville's Institutional Repository.

[3] Bourcier, D. (2003). De *l'intelligence artificielle à la personne virtuelle: émergence d'une entité juridique?*. Droit et société (n°49), p. 847-871.

[4] Can, F. & Patton, J. (2004). *Change of Writing Style with Time*. Computers and the Humanities 38: 61–82. Obtenido el 12 de septiembre 2015 de: http://www.users.miamioh.edu/canf/papers/chum04.pdf

[5] Chen, X., Hao, P., Chandramouli, R. & Subbalakshmi, K. (2011). *Authorship Similarity Detection from Email Messages.* P. Perner (Ed.): MLDM 2011, LNAI 6871, pp. 375–386.

[6] Deitel, P., Deitel, H. (2012). *C++ How to Program*. Massachusetts: Pearson, Prentice Hall.

[7] El Manar El Bouanani, S. & Kassou, I. (2013). *Using lexicometry and vocabulary analysis techniques to detect a signature for web profile.* 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. The 2012 International Workshop on Semantic Social Network Analysis and Design (SSNAD)

[8] Foltz, P. (1992). *Personalized Information Delivery: An Analysis of Information Filtering Methods*. Comminucations of the ACM, 35(12), 51-60. Obtenido el 12 de septiembre 2015 de: http://www-psych.nmsu.edu/~pfoltz/cacm/cacm.html

[9] Gokhale, A., Borkar, K. & Prasad, R. (2013). *A Proposed System for Author Identification Using Statistical Method.* International Journal of Engineering Research & Technology (IJERT) ISSN: 2278-0181 Vol. 2 Issue 9.

[10] Hong Riu Tan, R. & Tsai, F. (2010). *Authorship Identification for Online Text.* 2010 International Conference on Cyberworlds.

[11] Joula, P. & Vescovi, D. (2010*). Analyzing stylometric approaches to author obfuscation*. Advances in digital forensics VII, pp. 115-125.

[12] Kumar P., Laksghmi & Raj G. (2014). *A Pragmatic Validation of Stylometric Techniques using BPA.* 978-1-4799-4236-7/14.

[13] Lakshmi & Kumar, P. (2010). *A Study on Author Identification through Stylometry.* International Journal of Computer Science & Communication Networks,Vol 2(6), 653-657.

[14] Lantz, B. (2013). *Machine Learning with R: Learn how to use R to apply powerful machine learning methods and gain an insight into real-world applications.* Livery Place: Packt Publishing Ltd.

[15] Luart, L., Tazhibayeva, S., Wagoner, A, & Taylor, J. (2013). *Style Features for Authors in Two Languages.* Obtenido el 1 de mayo 2015 de http://ieeexplore.ieee.org/xpl/login.jsp?tp=&arnumber=6690051&url=http%3A%2F%2Fieeexplore.ieee.org%2Fxpls%2Fabs_all.jsp%3Farnumber%3D6690051

[16] Mohtasseb, H. & Ahmed, A. (2009). *Two-Layer Classification and Distinguished Representations of Users and Documents for Grouping and Authorship Identification.* IEEE 978-1-4244-4738-1/09

[17] Narayanan, A., Paskov, H., Zhenqiang, N. & Bethencourt, J. (2012). *On the Feasibility of Internet-Scale Author Identification.* Obtenido el 1 de mayo 2015 de http://randomwalker.info/publications/author-identification-draft.pdf

[18] Neme, A., Lugo, B., & Cervera, A. (2010). *Detection of Different Authorship of Text Sequences through Self-organizing Maps and Mutual Information Function.* G. Sidorov et al. (Eds.): MICAI 2010, Part II, LNAI 6438, pp. 186–195.

[19] Orebaugh, A. (2006) *An Instant Messaging Intrusion Detection System Framework: Using Character frecuency analysis for authorship identification and validation.* Obtenido el 2 de mayo 2015 de http://ieeexplore.ieee.org/xpl/login.jsp?tp=&arnumber=4105332&url=http%3A%2F%2Fieeexplore.ieee.org%2Fxpls%2Fabs_all.jsp%3Farnumber%3D4105332

[20] Ragel, R., Herath P. & Senanayake, U. (2013). *Authorship Detection of SMS Messages Using Unigrams.* 2013 IEEE 8th International Conference on Industrial and Information Systems, ICIIS 2013, Sri Lanka

[21] Ramyaa, He, C. & Rasheed, K. (2012). *Using Machine Learning Techniques for Stylometry.* Obtenido el 1 de mayo 2015 de http://www2.tcs.ifi.lmu.de/~ramyaa/publications/stylometry.pdf

[22] Rudman, J. (1998). *The State of Authorship Attribution Studies: Some Problems and Solutions*. Computers and the Humanities 31: 351–365. Obtenido el 12 de septiembre 2015 de: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.455.9936&rep=rep1&type=pdf

[23] Stamatos, E. (s.f.). *A Survey of Modern Authorship Attribution Methods.* Obtenido el 1 de mayo 2015 de: http://www.icsd.aegean.gr/lecturers/stamatatos/papers/survey.pdf

[24] Valera, P., Justino, E., Oliveira, L. (2011). *Selecting Syntactic Attributes for Authorship Attribution.* Proceedings of International Joint Conference on Neural Networks, San Jose, California, USA

[25] Zheng, R., Li, J., Chen, H., & Huang, Z. (2006). *A framework for authorship identification of online messages: Writing style features and classification techniques*. Journal of the American Society of Information Science and Technology, 57(3), 378-393.

ANEXO A: MODO DE EMPLEO DE LA HERRAMIENTA

La utilización de la herramienta consiste en pasos básicos que incluyen compilar, cargar archivos de texto, ejecutar el programa principal y finalmente graficar las estadísticas obtenidas. A continuación se explica cada uno de los procesos de forma detallada.

A.1. Descargar la carpeta de la herramienta Pattern Recognition.

Con mentalidad de que la persona que quiera acceder a la aplicación tenga derecho a editar y presentar mejorías la herramienta esa licenciada mediante Creative Commons, bajo el tipo de licencia Attribution-NonCommercial 4.0 International.

El proyecto ha sido colocado en un repositorio digital en gitHub desde donde se lo puede clonar a otra máquina que desee hacer las veces de host para la herramienta. Para facilidad del usuario, el enlace al proyecto se indica en la siguiente URL:

https://github.com/robertomaldonado/PatternTool

Una vez que se ha descargado la carpeta, se debe verificar que dentro de la raíz de la carpeta, existan los archivos *TextParser*, *FileHandler* y *Analyzer* (Con ambas extensiones: *.cpp* y *.h*). De igual forma se debe determinar que existan en el directorio las carpetas llamadas *shells* y *files*. Esto es importante puesto que los resultados de los análisis se almacenarán en *files* y los resultados visuales en *shells*, una vez que todo el proceso se haya cumplido de forma satisfactoria.

A.2. Compilar el código fuente de la herramienta.

Con el propósito de generar la traducción del programa C++ a lenguaje de máquina se utiliza el compilador[6]. En este caso se emplea el compilador g++, de Free Software Foundation, el mismo que es basado en nix y usualmente operado por la línea de comandos. Al compilar no se debe olvidar de cargar todas las clases

vinculadas al programa principal, por lo que la instrucción debe tener algún parecido a la siguiente sentencia:

g++ -std=c++11 main.cpp TextParser.cpp Analyzer.cpp FileHandler.cpp –o executable

La sentencia es ejecutada por medio de una consola de comandos que debe permanecer abierta dentro del mismo directorio. La aparición de un archivo ejecutable, sucede después de compilar y realizar la acción de enlace con librerías o clases externas que se utilizan dentro del código.

A.3. Cargar archivos de textos para la comparación.

Antes de ejecutar el programa para verificar la similitud entre dos autores distintos, se deben cargar los archivos de forma correcta para que sean accesibles al programa escrito en *C++*. Esto implica colocar el texto obtenido del primer autor (Autor B) en el archivo *default1* con extensión *.txt.* De forma semejante, se debe colocar el contenido de cada uno de los textos del segundo autor (Autor A) en los archivos *default2.txt* hasta *default6.txt.*

A.4. Ejecución del programa principal.

Ahora, desde la consola que fue utilizada para ejecutar la compilación en el paso anterior, se realiza un llamado al archivo ejecutable con el nombre que fue otorgado. Siguiendo con las instrucciones anteriores, en este caso la sentencia será:

./executable

Es importante especificar que el archivo es de carácter ejecutable por lo que no se puede omitir la estructura inicial *./* que precede al nombre del mismo.

Cuando el programa entra en ejecución, el usuario deberá seguir las instrucciones del mismo. Los únicos parámetros solicitados son el modo de *debugging* y el número de textos totales que posee el operador de la aplicación. En caso de que el usuario

desee realizar pruebas independientes con agrupaciones específicas de parámetros, entonces el programa verificará cuales son los el usuario desea incluir.

En la consola se verán los resultados del análisis, los que son presentados en porcentajes de similitud. Mientras mayor es la similitud, esto quiere decir que la autoría puede atribuirse al mismo autor en un porcentaje de probabilidad. Los resultados quedan abiertos, para que la conclusión sea obtenida por una persona, puesto que el sistema desarrollado no es predictivo.

A.5. Verificar y otorgar los permisos necesarios a !binBash

En el mismo directorio debe habilitarse los permisos de ejecución para lanzar los procesos por medio de instrucciones pre-escritas en un shell script. Para ello se debe hacer la siguiente sentencia:

chmod 777 /shells

sentencia que concede los permisos a todos los archivos que se encuentran en el directorio especificado.

A.6. Ejecución de los comandos para producir componentes visuales

Una vez que los permisos han sido otorgados a los shells y los datos estadísticos han sido escritos se procede al proceso de procesamiento de gráficos. Para ello se debe ingresar en el directorio de *shells* con el comando "*cd shells"*. Como se puede divisar existen varios archivos, los cuales han sido escritos de forma anticipada, puesto que el orden de los parámetros es de extrema importancia. Al ser ejecutable, se debe llamar al archivo *plot.sh* de la siguiente forma:

./plot.sh.

Este comando nos dará el retroalimentación en la consola sobre los procesos que han sido completados. Finalmente estos comandos generan una carpeta denominada *histograms* que contienen los gráficos estadísticos.

ANEXO B: REFERENCIA A LOS PARÁMETROS ANALIZADOS

File: General Characters

Punctuation Group:

1. PPD: Punctuation Period
2. PCN: Punctuation Colon
3. PCA: Punctuation Comma
4. PSC: Punctuation Semicolon
5. EEO: Expression Exclamation Open
6. EEC: Expression Exclamation Close
7. EQO: Expression Question Open
8. EQC: Expression Question Close

Mathematical Group:

1. MPS: Mathematical Plus
2. MMS: Mathematical Minus
3. MTS: Mathematical Times
4. MMO: Mathematical Modulus
5. MEQ: Mathematical Equals
6. MGR: Mathematical Greater
7. MLR: Mathematical Lesser
8. MDN: Mathematical Division

Group Symbols:

1. GPO: Grouping Parenthesis Open
2. GPC: Grouping Parenthesis Close
3. GCO: Grouping Curly Bracket Open
4. GCC: Grouping XCurly Bracket Close
5. GBO: Grouping Bracket Open
6. GBC: Grouping Bracket Close

Other Symbols:

1. OBS: Other Back Slash
2. OAA: Other Arroba
3. ONS: Other Number Symbol
4. ODS: Other Dollar Sign
5. OHO: Other Carret
6. ONP: Other Ampersand
7. OVD: Other Pipe
8. OUS Other Underscore

File: Tildes

1. MAA: Accented Upper Case A
2. MAE: Accented Upper Case E
3. MAI: Accented Upper Case I
4. MAO: Accented Upper Case O
5. MAU: Accented Upper Case U

6. MIA: Accented Lower Case A
7. MIE: Accented Lower Case E
8. MII: Accented Lower Case I
9. MIO: Accented Lower Case O
10. MIU: Accented Lower Case U

File: Numerical

1. ZERO: 0
2. ONE: 1
3. TWO: 2
4. THREE: 3
5. FOUR: 4
6. FIVE: 5
7. SIX : 6
8. SEVEN: 7
9. EIGHT: 8
10. NINE: 9

File: Mayus Case Letters

- List of upper case letters belonging to the alphabet (26 characters)

File: Lower Case Letters

- List of lower case letters belonging to the alphabet (26 characters)

File: Greetings And Farewells

1. G1: Hola
2. G2: Buen Día
3. G3: Estimado
4. G4: Estimada
5. G5: Buenos
6. G6: Buenas
7. F1: Gracias
8. F2: Atentamente
9. F3: Cordialmente
10. F4: Suerte
11. F5: Saludos

File: Averages

1. AWL: Average word length
2. AWPS: Average words per sentence
3. AWPP: Average words per paragraph
4. ADPP: Average Dot Per Paragraph
5. ACP: Average Comma Per Paragraph
6. ACOPP: Average Colon Per Paragraph
7. ASCPP: Average Semicolon Per Paragraph
8. AQPP: Average question marks per paragraph
9. AMQPP: Average multiple question marks per paragraph
10. AWSPS: Average whitespaces per sentence
11. AK1: Average times of expression
12. AK2: Average times of expression
13. AK3: Average times of expression
14. AWS: Average white spaces
15. ACAS: Average comma after space
16. ACBS: Average comma before space

File: Totals

1. TUS: Total number of sentences beginning with upper case
2. TLS: Total number of sentences beginning with lower case
3. TNW: Total number of words
4. TNS: Total number of spaces
5. TNP: Total number of paragraphs
6. TNT: Total number of tildes
7. TNWS: Total number of white spaces
8. TNMQ: Total number of multiple question marks
9. TNME: Total number of multiple exclamation marks
10. TNE: Total number of ellipsis (…)
11. TNWAP: Total number of apostrophes
12. TNK1: Total number of “Basicamente”
13. TNK2: Total number of “Bueno”
14. TNK3: Total number of “Sin embargo”
15. TNCBS: Total number of commas before spaces
16. TNCAS: Total number of commas after spaces

Total of parameters analysed: 145 including various categories.