

## **Análisis de estilos de redacción para la identificación de autoría usando métodos de agrupamientos.**

### **Analysis of writing styles for the identification of authorship using grouping methods.**

Ariel Céspedes Pérez<sup>1</sup>

<sup>1</sup>Universidad de Las Tunas, Cuba y arielcespedes87@gmail.com.

#### **RESUMEN**

Internet es un espacio dinámico. Sus consumidores se han convertido en productores de la información que ellos mismos consumen, apoyados en herramientas y plataformas instaladas en grandes servidores, que facilitan el uso y la publicación de contenidos. Si bien este proceso ha democratizado el acceso de muchos a la información, también ha provocado una excesiva socialización de la propiedad intelectual y científica, pues los materiales se publican muchas veces bajo licencias que permiten su descarga sin el consentimiento del creador. En este contexto muchas veces se hace necesario determinar el autor de un documento anónimo, o uno cuyo autor esté en duda. Para realizar el análisis de autoría, como se le conoce a esta tarea, es necesario inferir características del autor a través de los documentos escritos por él y luego conformar un modelo de su estilo que pueda ser comparable con el de otra persona. Sin embargo, resulta impráctico realizar el procesamiento de todos los posibles autores que existen a partir de sus publicaciones. Por ello es necesario determinar procedimientos que, sin utilizar un conjunto de archivos de referencia, realicen un análisis del estilo en el texto y revelen sus variaciones estilográficas. En este trabajo se expone un procedimiento para lograr este propósito aplicando métodos de agrupamiento al documento que se desea analizar. Los resultados de los experimentos con determinados métodos de este tipo y varios rasgos lingüísticos, muestran que usando el algoritmo sIB en textos caracterizados por tri-gramas de caracteres y uni-gramas de palabras, se obtienen resultados aceptables.

Palabras clave: procesamiento de lenguaje natural; minería de textos; análisis de autoría; plagio; estilo de redacción.

#### **ABSTRACT**

The Internet is today a very dynamic and revolutionary space. Their consumers have become producers of the information they consume, supported by tools and platforms installed on large servers, which facilitate the use and publication of content. While this process has democratized the access of many to information, it has also caused excessive socialization of intellectual and scientific property, since materials are often published under licenses that allow their download without the consent of the creator. In this context it is often necessary to determine the author of an anonymous document, or one whose author is in doubt. To perform the authorship analysis, as this task is known, it is necessary to infer characteristics of the author through the documents written by him and then to conform a model of his style that can be comparable with that of another person. However, it is impractical to perform the processing of all possible authors that exist from their publications. For this reason, it is necessary to determine procedures that, without using a set of reference files, carry out an analysis of the style in the text and reveal its fountain variations. In this work, a procedure to achieve this purpose is presented, applying grouping methods to the document that is to be analyzed. The results of the experiments with certain methods of this type and several linguistic features show that using the sIB algorithm in texts characterized by three-grams of characters and uni-grams of words, acceptable results are obtained.

Keywords: natural language processing; text mining; authorship analysis; plagiarism; style of writing.

## 1. INTRODUCCIÓN

Internet es un espacio dinámico. Sus consumidores se han convertido en productores de la información que ellos mismos consumen. Todo esto a través herramientas y plataformas instaladas en grandes servidores que facilitan el uso y la publicación de contenidos. A juzgar por el crecimiento en la cantidad de servidores en funcionamiento, cada año existe un incremento considerable en el número de documentos disponibles en la red.

Si bien este proceso ha democratizado el acceso de muchos a la información, también ha provocado una excesiva socialización de la propiedad intelectual y científica, pues los materiales se publican muchas veces bajo generosas licencias que permiten su descarga y reutilización, en muchos casos sin el consentimiento del autor, abriendo así las puertas al plagio.

El plagio es una falta grave de ética que ya cuenta hoy con implicaciones legales. Según la Real Academia de la Lengua Española se define como la acción de copiar en lo sustancial obras ajenas, dándolas como propias (Española, 2001).

Uno de los plagios más difíciles de detectar de forma manual es el de texto. Este ocurre cuando en un documento se utilizan secciones de otro, sin proporcionar la fuente (Barrón Cedeño, 2011). Un caso muy frecuente es el uso del corta y pega (del inglés cut and paste) de un fragmento sin modificaciones, siendo este tipo el más fácil de descubrir. La tarea de detección se dificulta cuando, a diferencia de la copia exacta del fragmento original, se oculta un texto de tal manera que es difícil encontrar similitud entre ellos. Los trucos más usados son: utilizar sinónimos, cambiar el orden de las oraciones, adicionar palabras y sustituir frases largas por cortas (D. Funez & Errecalde, 2011). Por ello, es importante desarrollar mecanismos automatizados que permitan realizar esta tarea en un tiempo y confiabilidad razonables.

Para poder afirmar que un texto es resultado de plagio, se debe proporcionar la sección plagiada y la sección del documento fuente. El detector debe realizar comparaciones del contenido sospechoso con todos los posibles documentos. Es decir, a partir de una colección de referencia además del texto a analizar, se provee como resultado las secciones donde se produjo el plagio y las utilizadas para ello, encontradas en el archivo de referencia.

La detección de plagio es uno de los enfoques de una cuestión más general que es el análisis de autoría, aunque se interprete en ocasiones de forma equivalente. Este último consiste en determinar el autor de un documento anónimo o uno cuyo autor esté en duda (Stamatatos, 2009). Para ello es necesario inferir características del autor a través de los textos escritos por él. Estas, nos permitirán conformar un modelo de su estilo y medir qué tan similar puede ser un documento cualquiera a los creados por él (Stamatatos et al., 2014).

Para realizar la representación del estilo de redacción de una persona se debe contar con obras de referencia de la misma. El volumen de autores con referencias limita el alcance de las herramientas del análisis. Tal es el caso de un texto redactado supuestamente por un único autor, en el cual se desean identificar las secciones que fueron escritas por otros. En este caso, comparar el estilo de redacción del texto objeto de análisis a partir de los estilos adquiridos a priori con las obras de referencia, no resuelve el problema. Esto se debe a que es impráctico realizar el procesamiento de todos los posibles autores a partir de sus publicaciones, además existen muchos que no poseen publicaciones en formato electrónico. Por lo tanto, es necesario determinar procedimientos que, sin utilizar un conjunto de archivos de referencia, analicen el texto y revelen sus variaciones estilográficas.

## 2. METODOLOGÍA

El análisis de autoría se divide en enfoques o subtarear: detección de autoría, verificación de autoría, agrupamiento de autores, variaciones en el estilo, detección de plagio e identificación del perfil del autor. Las recientes investigaciones sobre el tema, están dirigidas fundamentalmente a la representación de los estilos de redacción de los autores de forma eficiente y al desarrollo de algoritmos enfocados en la detección de plagio, protección de derechos de autoría y análisis de flujos de información, a partir de la recuperación y extracción de información textual. (Potthast, 2012; Potthast et al., 2014).

La representación de estos estilos está caracterizada por rasgos lingüísticos. Estos son el núcleo del análisis de autoría, independientemente de la subtask en la que se enfoque. El propósito radica en intentar identificar un estilo propio de redacción para cada autor con respecto al resto.

Son varias las propuestas de rasgos publicadas por diferentes investigadores en sus trabajos, para capturar, identificar o representar el estilo de redacción del autor de un documento. Además, no solo es necesario identificar los rasgos, sino calcular qué tan importantes son y cómo podrían caracterizar a su autor. Generalmente se usa una distribución o identificación por capas lingüísticas.

### **Estrategia de descomposición de texto**

Una vez identificada la subtask en la que se enfoca la problemática presentada, el objetivo para cualquiera de ellas es representar el estilo de escritura, para su posterior análisis. Para ello es necesario que el documento sea dividido de acuerdo con una estrategia de descomposición, teniendo en cuenta que la que se determine influye directamente en el resultado final.

Las estrategias más simples pueden ser más sencillas de implementar y más rápidas en ejecución, pero no siempre obtienen un buen desempeño (D. G. Funez & Errecalde, 2012).

La estrategia más simple y rápida es la división del texto en bloques de tamaño fijo, tales como cantidad de palabras u oraciones. La división del texto en límites estructurales, tales como: oraciones, párrafos o capítulos suele ser una mejor opción ya que las secciones plagiadas suelen ser estructuras de este tipo. Otra alternativa es la división del texto por tópicos usando algún algoritmo de segmentación de textos, siendo esta la de mayor impacto en la actualidad, pero no aconsejable en el caso del agrupamiento de autores ya que al basarse en la representación de los temas es muy probable que a la hora de realizar el agrupamiento lo haga de acuerdo al tema específico y no respecto al estilo de redacción de los autores.

El siguiente paso en el análisis de autoría es la representación de los rasgos que caractericen el estilo de redacción de los autores. Existen diferentes modelos de representación. El más utilizado en las ciencias de la computación es el modelo de espacio vectorial, debido a su simplicidad y su clara base conceptual, que corresponde a la intuición humana en el procesamiento de información y de datos (Sidorov, 2013).

### **Rasgos lingüísticos**

Una vez seleccionado un modelo para la representación de las secciones del texto a analizar (en este caso el modelo de espacio vectorial), se deben seleccionar que rasgos (rasgos lingüísticos en el caso de esta investigación) deben ser tenidos en cuenta en la caracterización de dichas secciones.

Los rasgos lingüísticos son el núcleo de la tarea de análisis de autoría, independientemente de la subtask en la que se enfoque. El propósito radica en intentar identificar un estilo propio de redacción para cada autor con respecto al resto.

Generalmente estos rasgos son agrupados en capas, conocidas como capas lingüísticas. La mayoría de estos investigadores identifican cinco capas lingüísticas de rasgos: la capa de fonemas, la capa de caracteres, la capa léxica, la capa sintáctica y la capa semántica (Stamatatos, 2009).

Estas capas son muy abstractas, lo que explica la diversidad existente en el uso de rasgos, incluso una vez escogidos puede variar la implementación debido al pre-procesado realizado en los documentos o la representación de frecuencias utilizada. En el estudio realizado se identificaron 47 rasgos citados en la bibliografía.

Rasgos de la capa de caracteres: signos de puntuación (Argamon & Juola, 2011; Brooke & Hirst, 2012; Castillo, Cervantes, Vilariño, Pinto, & León, 2014; Fréry, Largeron, & Juganaru-Mathieu, 2014; Halvani & Steinebach, 2014; Halvani, Steinebach, & Zimmermann, 2013; Mayor et al., 2014; Tanguy, Sajous, BasilioCalderone, & Hathout., 2012; Vilariño, Pinto, Gómez, León, & Castillo, 2013)

sufijos (Argamon & Juola, 2011; Castillo et al., 2014; Halvani & Steinebach, 2014; Halvani et al., 2013; Mayor et al., 2014; Ruseti & Rebedea, 2012; Vilariño et al., 2013); prefijos (Argamon & Juola, 2011; Castillo et al., 2014; Castillo et al., 2012; Halvani & Steinebach, 2014; Halvani et al., 2013; Mayor et al., 2014)

letras por palabras (Argamon & Juola, 2011; Brooke & Hirst, 2012; Feng & Hirst, 2013; Ghaeini, 2013; Giraud & Artières, 2012; Halvani & Steinebach, 2014; Ruseti & Rebedea, 2012); n-grama de caracteres

(Fréry et al., 2014; Giraud & Artières, 2012; Halvani & Steinebach, 2014; Ruseti & Rebedea, 2012; Tanguy et al., 2012); letras por oración (Argamon & Juola, 2011; Feng & Hirst, 2013; Fréry et al., 2014; Halvani & Steinebach, 2014; Halvani et al., 2013); letras por párrafo (Argamon & Juola, 2011; Feng & Hirst, 2013; Fréry et al., 2014; Halvani & Steinebach, 2014; Halvani et al., 2013); n-grama de letras (Halvani & Steinebach, 2014; Halvani et al., 2013); contracciones y abreviaturas (Argamon & Juola, 2011); comas (Argamon & Juola, 2011; Ledesma, Fuentes, Jasso, Toledo, & Meza, 2013); puntos (Argamon & Juola, 2011; Ledesma et al., 2013); paréntesis (Argamon & Juola, 2011; Ledesma et al., 2013); sílabas (Brooke & Hirst, 2012); números (Ledesma et al., 2013); mayúsculas (Ledesma et al., 2013); caracteres (Halvani & Steinebach, 2014)

Rasgos de la capa léxica: palabras de función (del inglés stop words) (Castillo et al., 2014; Castillo et al., 2012; Feng & Hirst, 2013; Halvani et al., 2013; Mayor et al., 2014; Vilariño et al., 2013); palabras por oración (tamaño de oración) (Argamon & Juola, 2011; Feng & Hirst, 2013; Fréry et al., 2014; Halvani & Steinebach, 2014; Ledesma et al., 2013; Mayor et al., 2014); n-gramas de palabras (Argamon & Juola, 2011; Castillo et al., 2014; Fréry et al., 2014; Giraud & Artières, 2012; Halvani et al., 2013; Ledesma et al., 2013); n-gramas de palabras del discurso (Castillo et al., 2012; Giraud & Artières, 2012; Ruseti & Rebedea, 2012; Tanguy et al., 2012; Vilariño et al., 2013); palabras por párrafo (tamaño de párrafo) (Argamon & Juola, 2011; Giraud & Artières, 2012; Halvani & Steinebach, 2014; Ledesma et al., 2013); pronombres (Argamon & Juola, 2011; Brooke & Hirst, 2012); palabras del discurso (POS, por sus siglas en inglés) (Argamon & Juola, 2011; Brooke & Hirst, 2012; Giraud & Artières, 2012); verbos modales (Argamon & Juola, 2011); errores ortográficos (Argamon & Juola, 2011); emoticones (Argamon & Juola, 2011); polisemia (Argamon & Juola, 2011); tiempo verbal (Brooke & Hirst, 2012); riqueza de vocabulario (Fréry et al., 2014; Giraud & Artières, 2012); enlace de palabras (Giraud & Artières, 2012); frecuencia de palabras del discurso (Feng & Hirst, 2013); cantidad de oraciones (Ghaeini, 2013); cantidad de párrafos (Ghaeini, 2013); promedio de oraciones (Ghaeini, 2013); promedio de párrafos (Ghaeini, 2013); palabras únicas por oración (Ghaeini, 2013); n-grama de palabras de función (Halvani et al., 2013); oraciones que comienzan con palabras de función (Halvani et al., 2013); n-grama de palabras de función (Mayor et al., 2014); cohesión léxica (Tanguy et al., 2012); complejidad morfológica (Tanguy et al., 2012); entropía de palabras del discurso (Feng & Hirst, 2013).

Rasgos de la capa sintáctica: dependencias sintácticas (Tanguy et al., 2012); primera/tercera persona (Tanguy et al., 2012); estructura y complejidad sintáctica (Ruseti & Rebedea, 2012); acotaciones (Tanguy et al., 2012); frases verbales (Tanguy et al., 2012);

Se reconocen notablemente, en correspondencia con la frecuencia y resultado de su uso en las investigaciones publicadas, los siguientes rasgos: signos de puntuación, sufijos, prefijos, letras por palabras, n-grama de caracteres, letras por oración, letras por párrafo, palabras de función (del inglés stop words), palabras por oración, n-gramas de palabras, n-gramas de palabras del discurso, palabras por párrafo.

En correspondencia con lo antes planteado se propone la representación de esos rasgos utilizando n-gramas tradicionales. Se utilizó la representación vectorial para cada n-grama de rasgo, por lo que cada muestra (segmento de texto de 1 párrafo) es representada por 36 vectores correspondiendo con cada uno de los rasgos, representado en forma de n-grama.

Los rasgos fueron asociados en tres grupos: basados en caracteres, basados en palabras y basados en lema e información morfológica.

Basados en caracteres: n-grama de caracteres. (n=1,2,3); n-grama de prefijos de tamaño 2. (n=1,2,3); n-grama de sufijos de tamaño 2. (n=1,2,3); uni-grama de signos de puntuación.

Basados en palabras: n-grama de palabras. (n=1,2,3); uni-grama de palabras de función.

Basados en lema e información morfológica: n-grama de lemas. (n=1,2,3); n-grama de categoría sin información morfológica. (n=1,2,3); n-grama de categoría con información morfológica. (n=1,2,3).

Para realizar la extracción de los rasgos se implementaron una serie de métodos usando las potencialidades que brinda FreeLing, una librería sobre la cual se pueden desarrollar potentes aplicaciones de Procesamiento de Lenguaje Natural.

Una vez obtenidos los segmentos del texto a analizar, caracterizados por los rasgos escogidos, se aplica algún método de agrupamiento para identificar diferencias entre estos segmentos. Aquellos segmentos que formen parte de grupos diferentes suponen un estilo de redacción diferente.

### **Procedimiento para identificar autores usando métodos de agrupamiento**

Dado un conjunto de objetos definidos en términos de un grupo de rasgos (tal como se propone en el modelo de espacio vectorial), los métodos de agrupamiento intentan construir particiones o cubrimientos de este conjunto, donde la semejanza intra-grupo sea máxima y la semejanza inter-grupos sea mínima (Pons Porrata, 2004). Pueden ser vistos además como procedimiento para unir una serie de vectores de acuerdo con un criterio de cercanía. Esta cercanía se define en términos de una determinada función de distancia o similitud. Generalmente, los vectores de un mismo grupo (o clúster) comparten propiedades comunes (Kaufman & Rousseeuw, 1990). El conocimiento de los grupos puede permitir una descripción sintética de un conjunto de datos multidimensional complejo. Esta descripción sintética se consigue sustituyendo la descripción de todos los elementos de un grupo por la de un representante característico del mismo.

La selección de un algoritmo de agrupamiento depende del tipo de dato disponible y del propósito o aplicación particular. Si el análisis de agrupamiento es usado como una herramienta descriptiva o exploratoria, es común aplicar varios algoritmos sobre los mismos datos para ver que conocimiento puede manifestarse (Fung, 2001).

Existen múltiples estrategias jerárquicas que pueden ser aplicadas para determinar las diferencias entre los segmentos de acuerdo con los rasgos propuestos. Se tomaron como referencia 5 métodos de agrupamiento implementados en la herramienta Weka, los que permitieron determinar la efectividad del procedimiento para identificar variaciones en el estilo de redacción: Single Link (SL), Average Link (AL), Complete Link (CL), COBWEB (CW), siB.

El procedimiento para realizar la identificación de autores usando métodos de agrupamiento se fundamenta en el hecho de que cada autor tiene un estilo de escritura propio, el cual se mantiene a lo largo de todo el texto (Chen, Yeh, & Ke, 2010). Este estilo de escritura necesita estar representado por un modelo, en este caso modelo de espacio vectorial, que estará compuesto de rasgos lingüísticos que definen el estilo de escritura individual de cada creador. La detección de los autores se realiza identificando las variaciones estilográficas significativas en el documento caracterizado. Esto se hace a partir de los grupos obtenidos aplicando algoritmos de agrupamiento, ya que no se poseen caracterizaciones de los estilos de escritura de los autores y no se pueden realizar comparaciones con muestras de las que se conoce su clase (autor). En el procedimiento utilizado en la investigación pueden ser determinadas tres tareas fundamentales: construcción del modelo de estilo, aplicación de una estrategia de agrupamiento y determinación de los posibles segmentos de cada autor.

### **Descripción de los experimentos**

En los experimentos se considera la tarea de evaluar la representación de diferentes rasgos lingüísticos para la identificación de partes estilísticamente homogéneas, a partir de los algoritmos de agrupamiento descritos. Esencialmente, se tiene un texto del cual se conoce fue escrito por más de un autor. Además, no es posible comparar los estilos de redacción de dicho texto, adquiridos a través de las variables lingüísticas definidas, con los estilos de autores obtenidos a partir de un corpus de datos.

Para la realización de los experimentos se usa un corpus, construido por el autor de la investigación a partir de artículos periodísticos de 47 autores, de 8 medios digitales de prensa: Periódico Granma, Periódico Trabajadores, Periódico Juventud Rebelde, Periódico 26, Periódico Adelante, Agencia de Información Nacional y Sitio Web SoyCuba. Cada documento del corpus es un texto multiautor generado a través del procedimiento descrito en la Tabla 1, con el cual se obtienen textos mezclados uniformemente.

**Tabla 1 Procedimiento para generar textos multiautores.**



Entrada:	$k$ documentos, $D_1, D_2 \dots D_k$ , cada uno escrito por un autor diferente
	<p>Escoger <math>D_{R(i)}</math>, donde <math>R(i)</math> es una función que devuelve un número aleatorio entre 1 y <math>k</math>, con la condición de que <math>R(i) \neq R(i - 1)</math></p> <p>Agregar a <math>D_{multiautor}</math> las primeras <math>x</math> oraciones no utilizadas del texto <math>D_{R(i)}</math>, donde <math>x</math> es escogida aleatoriamente entre <math>y</math> y <math>n</math>. Se debe distinguir que <math>n</math> es la máxima cantidad de oraciones que se desea tener de cada autor en cada iteración.</p> <p>Los pasos 1 y 2 se repiten mientras existan oraciones no utilizadas en cualquiera los documentos <math>k</math> documentos de entrada, <math>D_1, D_2 \dots D_k</math></p>
Salida:	$D_{multiautor}$

### Métricas para la validación

Es difícil definir, cuándo, el resultado del agrupamiento es aceptable. Se han creado un conjunto de técnicas e índices para realizar su validación. En general existen dos tipos de validación: externa e interna. La principal diferencia es si se usa o no información externa para la validación, es decir, información que no es producto de la técnica de agrupación utilizada (León Guzmán, 2005).

Las técnicas de validación interna miden el agrupamiento únicamente basadas en información de los datos. Evalúan que tan buena es la estructura del agrupamiento sin necesidad de información ajena al propio algoritmo y su resultado (León Guzmán, 2005).

Como la validación externa mide la calidad del agrupamiento conociendo información externa de antemano, es principalmente usada para escoger un algoritmo de agrupamiento óptimo sobre un grupo de datos (data set) específicos.

Cuando se tiene información externa tal como la clase de cada dato, es común y ampliamente utilizado el análisis mediante la matriz de confusión. Una vez realizado el agrupamiento mediante algún algoritmo para ese propósito, este puede sugerir un agrupamiento de los datos, diferente al que indicaran las clases conocidas de antemano.

La matriz de confusión es completada a partir de cuatro términos: verdaderos positivos (VP), falsos positivos (FP), falsos negativos (FN) y verdaderos negativos (VN). El término VP hace referencia a aquellos puntos que fueron ubicados por el algoritmo en el mismo grupo que indicaba la clase con la que se contaba de antemano. El término FP hace referencia a aquellos puntos que fueron ubicados por el algoritmo en un grupo y que en realidad pertenecían a otro grupo. El término FN hace referencia a aquellos elementos de un grupo que fueron ubicados en un grupo diferente al que indicaba su etiqueta. El término VN hace referencia a aquellos elementos que fueron ubicados correctamente fuera de un grupo, es decir, aquellos elementos ajenos al grupo en cuestión y que efectivamente no correspondían a este (León Guzmán, 2005). Existen otras métricas externas ampliamente utilizadas y provenientes del campo de la Recuperación de la Información: precisión (precision) y exhaustividad (recall). La precisión es la proporción entre el número de documentos relevantes recuperados y el número de documentos recuperados, como se muestra en la ecuación 1 (Dillon, 1983). De esta forma, cuanto más se acerque el valor de la precisión al valor nulo, mayor será el número de documentos recuperados que no consideren relevantes. Si por el contrario, el valor de la precisión es igual a uno, se entenderá que todos los documentos recuperados son relevantes. Acorde con la definición y en correspondencia con los términos definidos en la matriz de confusión esta puede expresarse como se presenta en la ecuación 2. En el caso de la exhaustividad expresar la proporción de documentos relevantes recuperados, comparado con el total de los documentos que son relevantes existentes en la base de datos, con total independencia de que éstos, se recuperen o no. La ecuación en este caso se expresa como se muestran en la ecuación 3 (Kent, Berry, Luehrs, & Perry, 1955). Si el resultado de la ecuación 3 arroja como valor 1, se tendrá la exhaustividad máxima posible, y esto indica que se ha encontrado todo documento relevante que residía en la base de datos, siendo la recuperación de documentos perfecta. Por el contrario, en el caso que el valor de la exhaustividad sea igual a cero, se tiene que los documentos obtenidos no poseen relevancia alguna. Acorde con la definición y en correspondencia con los términos definidos en la matriz de confusión esta puede expresarse como se presenta en la ecuación 4.

$$Precision = \frac{|\{documentos\ relevantes\} \cap \{documentos\ recuperados\}|}{|\{documentos\ recuperados\}|} \quad (1)$$

$$Precision = \frac{VP}{VP + FP} \quad (2)$$

$$Exhaustividad = \frac{|\{documentos\ relevantes\} \cap \{documentos\ recuperados\}|}{|\{documentos\ relevantes\}|} \quad (3)$$

$$Exhaustividad = \frac{VP}{VP + FN} \quad (4)$$

Con los conceptos de precisión y exhaustividad es posible definir otro tipo de métrica llamada "Medida F". Esta se da en función de las dos métricas ya vistas y puede ser interpretada como la media armónica de ambas. En particular la medida F es manejada por un parámetro  $\alpha$ , como se muestra en la ecuación 5.

$\alpha = 1$  media armónica.

$\alpha \in (0: 1)$  preferencia por la precisión.

$\alpha > 1$  preferencia por la exhaustividad

$$F_{\alpha} = \frac{1 + \alpha}{\frac{1}{precision} + \frac{\alpha}{cobertura}} \quad (5)$$

### 3. RESULTADOS Y DISCUSIÓN

Se generaron más de 15 mil textos multiautores, asociados en 4 grupos (de 2 hasta 5 autores). Esta división tiene como objetivo comprobar el comportamiento de los rasgos y algoritmos en correspondencia con los autores involucrados en la mezcla.

En el texto generado quedan identificados los párrafos por autores, obteniéndose etiquetas referidas a un autor (párrafos puros de un solo autor) y otras referidas a varios (párrafos mezclados con oraciones de varios autores). Para realizar los experimentos se escogieron de forma aleatoria los textos, para lograr una mejor confianza en los resultados.

En cada representación del texto basada en caracteres, basada en palabras y basada en lemas e información morfológica, fue aplicado cada uno de los algoritmos seleccionados. Los métodos fueron empleados ignorando las clases de las tuplas que conforman cada representación. Una vez obtenidos los grupos a partir de cada procedimiento, se realizó la evaluación del mismo para cada rasgo. Para hacer esta evaluación se calcularon los valores de precisión, exhaustividad y medida F, en este caso F1 como combinación armónica de precisión y exhaustividad.

#### Resultados de la experimentación con rasgos basados en caracteres.

En la Tabla 2 se resumen los resultados obtenidos para los uni-gramas de caracteres, en correspondencia con la cantidad de autores que se encuentran mezclados en el texto. Los valores mostrados en la tabla son los promedios de los valores de las medidas (precisión, exhaustividad, medida F, columnas B, C y D respectivamente) para los 4000 textos procesados, 1000 por cada mezcla de autores. Además, se muestra el F1 promedio de cada procedimiento sobre todos los textos procesados (columna E).

La representación basada en uni-gramas de caracteres no muestra resultados de F1 promedio por encima de 0.48. Como se muestra en la Tabla 2 B, la máxima precisión mostrada es de 0.98 con el procedimiento CW cuando la mezcla es de dos autores, sin embargo, la exhaustividad en este caso es de 0.12 (Tabla 2 C). En el caso de la exhaustividad, se destaca el 0.7 obtenido por el sLB (Tabla 2 C), en la mezcla de dos autores. El método sLB es de manera general el de mejor desempeño, con un F1 promedio de 0.48 (Tabla 2 E). Es notable el resultado de este método en cuanto a precisión, exhaustividad y valor F1 (0.7, 0.7 y 0.68 respectivamente) obtenidos en la mezcla de dos autores (Tabla 2 B, C, D).

El desempeño de los métodos sobre las representaciones basadas en bi-grama de caracteres es similar al obtenido sobre representaciones basadas en uni-grama de caracteres (Tabla 2). La principal diferencia se encuentra en los resultados aplicando el procedimiento sLB con un ligero aumento en la medida F1

promedio. Este valor tuvo un incremento de 0.48 a 0.55. En el caso de la medida F1 para este procedimiento aplicado a mezclas de textos de dos autores (Tabla 2 D), también hubo un ligero aumento. Este fue de 0.68 a 0.77.

En el caso de los tri-gramas de caracteres se observa, un F1 promedio máximo de 0.65. La mayor precisión, exhaustividad y F1 se alcanza con el método sIB (0.86, 0.87 y 0.85 respectivamente) en los textos mezclados de 2 autores.

**Tabla 2 Desempeño del rasgo uni-grama de caracteres en la identificación de autores.**

Algoritmos (A)	Precisión (B)				Exhaustividad (C)				F1 (D)				Promedio (según F1) (E)
	2	3	4	5	2	3	4	5	2	3	4	5	
SL	0.66	0.63	0.54	0.53	0.54	0.4	0.3	0.25	0.46	0.33	0.27	0.24	0.3239321
AL	0.67	0.64	0.52	0.52	0.54	0.41	0.31	0.27	0.48	0.34	0.28	0.26	0.3405318
CL	0.66	0.57	0.48	0.46	0.58	0.45	0.38	0.34	0.53	0.4	0.35	0.32	0.4026507
CW	0.98	0.98	0.91	0.89	0.12	0.13	0.12	0.12	0.21	0.23	0.21	0.21	0.2168044
sIB	0.7	0.56	0.39	0.29	0.7	0.57	0.45	0.37	0.68	0.55	0.39	0.3	0.4796913
Promedio	0.73	0.68	0.57	0.54	0.5	0.39	0.31	0.27	0.47	0.37	0.3	0.27	0.3527221

Con los uni-gramas, bi-gramas y tri-gramas de prefijos se obtiene un resultado similar. Se alcanza una buena precisión de 0.98 aplicando CW, sin embargo, se observa una baja exhaustividad. El mejor desempeño se obtiene con el procedimiento sIB con un valor promedio F1 de 0.56, y como F1 en mezclas de dos autores de 0.77.

Los resultados arrojados por los uni-gramas de sufijos no se diferencian de los obtenidos con las representaciones de uni-gramas de caracteres. Los bi-gramas y tri-gramas de sufijos muestran resultados similares. La mejor precisión es de 0.98 y se alcanza aplicando CW en textos de dos autores. El mejor desempeño se logra con el procedimiento sIB con valor F1 promedio de 0.51.

En el caso de las representaciones basadas en uni-gramas de signos se obtiene un F1 promedio por debajo de 0.41.

Con la representación a partir de uni-gramas de palabras, se presenta un desempeño notable de acuerdo con los resultados analizados. Con el método sIB se alcanza un F1 promedio de 0.63. Los resultados particulares de este método en cuanto a precisión, exhaustividad y F1 en textos de dos autores son de 0.86, 0.87 y 0.85 respectivamente.

Los bi-gramas y tri-gramas de palabras registran un desempeño menos considerable que el de uni-gramas de palabras, con un F1 promedio máximo de 0.59. La máxima precisión se alcanza es de 0.98 aplicando CW con textos de dos autores, sin embargo, en este caso la exhaustividad es de solo 0.12. El método con mejor desempeño es el sIB con valores de precisión, exhaustividad y F1 de 0.77, 0.79 y 0.77 respectivamente.

Los uni-gramas de palabras de función presentan un bajo desempeño, el F1 promedio no supera el 0.49. Se mantienen los mejores resultados aplicando sIB. En el caso de los rasgos basados en lemas e información morfológica los mejores resultados se obtienen realizando la representación a partir de los uni-gramas, bi-gramas y tri-gramas de lemas, y los tri-gramas de categorías con información morfológica. Su desempeño aplicando sIB es como promedio de 0.56 y de 0.77 en el caso del análisis de documentos mezclados de dos autores.

Como tendencia usual para todos los rasgos, la precisión y la exhaustividad (y consecuentemente el F1) disminuyen a medida que se incrementa la cantidad de autores.

No se observan diferencias significativas entre los resultados obtenidos al aplicar los métodos SL, AL y CL sobre los rasgos seleccionados. El promedio F1 se encuentra entre 0.36 y 0.28.



El mejor desempeño para todos los rasgos se obtiene aplicando el procedimiento sIB, con un F1 promedio de 0.65 como máximo, usando tri-gramas de caracteres y 0.4 como mínimo, usando bi-gramas de categorías sin información morfológica.

#### 4. CONCLUSIONES

1. A partir del estudio del análisis de autoría y sus enfoques o subtarefas, se determinan los elementos que deben ser considerados en este tipo de análisis: descomposición del texto; selección de los rasgos lingüísticos; representación de los textos, de acuerdo con un modelo computacional, y la identificación de la similitud entre los textos. Se destaca el uso del modelo de espacio vectorial como modelo de representación.
2. La caracterización de los métodos de agrupamiento de acuerdo con diversas categorías permite identificar a los procedimientos jerárquicos y particionales, como las clasificaciones más difundidas en el área. De acuerdo con las características de la investigación, y basados en las propias de cada procedimiento se escogieron los métodos jerárquicos: Single Link, Average link, Complete Link, COBWEB y sIB, para realizar los experimentos.
3. En el estudio realizado sobre las formas de representación computacional de los estilos de redacción se identificaron 23 variables relevantes: 10 basadas en caracteres, 4 basadas en palabras y 9 basadas en lemas e información morfológica.
4. Los experimentos mostraron un mejor desempeño para los rasgos tri-grama de caracteres y uni-grama de palabras, aplicando el procedimiento sIB, con un F1 promedio de 0.65 y 0.63 respectivamente, y un F1 máximo de 0.85. Los rasgos restantes no muestran un buen desempeño.
5. Se propone realizar la identificación de autores mediante algoritmos de agrupamiento aplicando el algoritmo sIB a las representaciones de los estilos de redacción de los segmentos, a partir de los rasgos tri-grama de caracteres y uni-grama de palabras. Como tendencia usual, la precisión y exhaustividad (y consecuentemente el F1) disminuyen a medida que se incrementa la cantidad de autores, para todos los rasgos.

#### 5. REFERENCIAS BIBLIOGRÁFICAS

- Argamon, S., & Juola, P. (2011). Overview of the International Authorship Identification Competition at PAN-2011.
- Barrón Cedeño, L. A. (2011). *Detección automática de plagio en texto*.
- Brooke, J., & Hirst, G. (2012). Paragraph Clustering for Intrinsic Plagiarism Detection using a Stylistic Vector-Space Model with Extrinsic Features.
- Castillo, E., Cervantes, O., Vilariño, D., Pinto, D., & León, S. (2014). Unsupervised method for the authorship identification.
- Castillo, E., Vilariño, D., Pinto, D., Olmos, I., González, J. A., & Carrillo, M. (2012). Graph-based and Lexical-Syntactic Approaches for the Authorship Attribution Task.
- Chen, C.-Y., Yeh, J.-Y., & Ke, H.-R. (2010). Plagiarism detection using ROUGE and WordNet. *arXiv preprint arXiv:1003.4065*.
- Dillon, M. (1983). Introduction to modern information retrieval: G. Salton and M. McGill. McGraw-Hill, New York (1983). xv+ 448 pp., \$32.95 ISBN 0-07-054484-0: Pergamon.
- Española, R. A. (Ed.) (2001) (22 ed.).
- Feng, V. W., & Hirst, G. (2013). Authorship Verification with Entity Coherence and Other Rich Linguistic Features.
- Fréry, J., Langeron, C., & Juganaru-Mathieu, M. (2014). UJM at CLEF in Author Verification based on optimized classification trees.
- Funez, D., & Errecalde, M. L. (2011). *Detección de plagio intrínseco usando la segmentación de texto*. Paper presented at the XVII Congreso Argentino de Ciencias de la Computación.
- Funez, D. G., & Errecalde, M. L. (2012). *Detección de plagio intrínseco basad en histogramas*. Paper presented at the XVIII Congreso Argentino de Ciencias de la Computación.

- Fung, G. (2001). A Comprehensive Overview of Basic Clustering Algorithms.
- Ghaeini, M. R. (2013). Intrinsic Author Identification Using Modified Weighted KNN.
- Giraud, F.-M., & Artières, T. (2012). Feature Bagging for Author Attribution.
- Halvani, O., & Steinebach, M. (2014). VEBAV - A Simple, Scalable and Fast Authorship Verification Scheme.
- Halvani, O., Steinebach, M., & Zimmermann, R. (2013). Authorship Verification via k-Nearest Neighbor Estimation.
- Kaufman, L., & Rousseeuw, P. J. (1990). Finding Groups in Data: An Introduction to Cluster Analysis. .
- Kent, A., Berry, M. M., Luehrs, F. U., & Perry, J. W. (1955). Machine literature searching VIII. Operational criteria for designing information retrieval systems. *American documentation*, 6(2), 93-101.
- Ledesma, P., Fuentes, G., Jasso, G., Toledo, A., & Meza, I. (2013). Distance learning for Author Verification.
- León Guzmán, E. (2005). Métricas para la validación de Clustering.
- Mayor, C., Gutierrez, J., Toledo, A., Martinez, R., Ledesma, P., Fuentes, G., & Meza, I. (2014). A Single Author Style Representation for the Author Verification Task.
- Pons Porrata, A. (2004). *Desarrollo de algoritmos para la estructuración dinámica de información y su aplicación a la detección de sucesos*. (Doctorado).
- Potthast, M. (2012). Technologies for reusing text from the Web.
- Potthast, M., Hagen, M., Beyer, A., Busse, M., Tippmann, M., Rosso, P., & Stein, B. (2014). *Overview of the 6th International Competition on Plagiarism Detection*. Paper presented at the CLEF (Online Working Notes/Labs/Workshop).
- Ruseti, S., & Rebedea, T. (2012). Authorship Identification Using a Reduced Set of Linguistic Features
- Sidorov, G. (2013). Construcción no lineal de n-gramas en la lingüística computacional.
- Stamatatos, E. (2009). A survey of modern authorship attribution methods. *Journal of the American Society for information Science and Technology*, 60(3), 538-556.
- Stamatatos, E., Daelemans, W., Verhoeven, B., Potthast, M., Stein, B., Juola, P., . . . Barrón-Cedeño, A. (2014). Overview of the Author Identification Task at PAN 2014. *analysis*, 13, 31.
- Tanguy, L., Sajous, F., BasilioCalderone, & Hathout, N. (2012). Authorship attribution: using rich linguistic features when training data is scarce.
- Vilariño, D., Pinto, D., Gómez, H., León, S., & Castillo, E. (2013). Lexical-Syntactic and Graph-Based Features for Authorship Verification.

## **SOBRE LOS AUTORES**

Profesor en la Universidad de Las Tunas desde 2013. Profesor Asistente y Máster en Informática Aplicada. Miembro de la Asociación Cubana de Reconocimiento de Patrones y la Sociedad Cubana de Matemática y Computación.