Detección de Plagio

Sergi Albiach Caro, Manel Angresola Navarro,
Stéphane Díaz-Alejo León & Antonio Martínez Leal

Índice

1. INTRODUCCIÓN

1. Introducción

El plagio, según la IEEE consiste en “*reusar las ideas, procesos, resultados o palabras de alguien más sin mencionar explícitamente a la fuente y su autor*”.

Esta es una técnica que se ha vuelto más fácil con la llegada de internet, llegando a acuñar el término “*ciberplagio*” o la existencia del síndrome de “*copia y pega*”.

1. Introducción

La acción de detectar un plagio corresponde tanto al ámbito del **procesamiento del lenguaje natura**l como al de la **recuperación de la información**.

Es importante destacar que en numerosas ocasiones el plagio es resultado de la **ignorancia** o **criptomnesia,** por lo que un proceso informático es insuficiente para tomar medidas punitivas.

1. Introducción

Se pueden clasificar los plagios según las partes del texto que se han intentado plagiar:

- Ideas
- Palabra a palabra
- Fuentes
- Autoría

Siendo los más difíciles de detectar, el plagio de **ideas** y el de **fuentes**.

2. TIPOS DE PLAGIO

2. Tipos de plagio

Existen primordialmente tres tipos de plagio:

- Copia exacta.
- Copia modificada.
- Plagio traducido.

2.1. COPIA EXACTA

2.1. Copia exacta

En el caso de plagio de copia exacta, no se altera de ninguna manera el texto original, por lo que los métodos que ofrecen mejores resultados son los modelos de huella digital.

2.2. COPIA MODIFICADA

2.2. Copia modificada

La complejidad del problema es incrementada.

Se precisa de representaciones con mayor flexibilidad: bolsa de palabras, n-gramas.

Se busca la **similitud** entre los fragmentos, medidas más utilizadas: similitud de coseno, coeficiente de Jaccard.

2.3. PLAGIO TRADUCIDO

2.3 Plagio traducido

Resultado de convertir un fragmento de un documento de una lengua a otra. La complejidad de su detección ha resultado en la aparición de tres modelos principales:

- **CL-ESA**: artículos de Wikipedia.
- **CL-ASA**: diccionario probabilístico.
- **CL-CNG**: relaciones interlingüísticas.

3. PARÁFRASIS

3. Paráfrasis

Según la Real Academia de la lengua Española (RAE), la paráfrisis se define como:

> "*Frase que expresa el mismo contenido que otra pero con diferente estructura sintáctica.*"

Esta paráfrasis se puede realizar mediante métodos como:

- Sustitución.
- Eliminación.
- Transformación
- Segmentación.
- Cambio de orden.

4. PRINCIPALES MÉTODOS DE DETECCIÓN

4. Principales métodos de detección

Existen dos planteamientos que destacan en el sector:

- Basados en el contenido del texto
- Basados en la estructura del texto

No obstante, a pesar de sus diferencias comparten las siguientes 4 fases:

- Representación
- Búsqueda de candidatos
- Acotación de pasajes
- Postproceso

4. Principales métodos de detección

Representación

A partir de un texto completo se aplica uno o varios procesos: Tokenización, Normalización, Eliminación de stop-words, Stemming y/o Lematización.

Formatos más comunes: bolsa de palabras, n-grams, por bloques.

Búsqueda de candidatos

Existen dos métodos básicos:

- Metodología IR, basada en queries y ranking de documentos.
- “Huellas” de documentos formados por hashes de chunks de longitud fija (shingles).

4. Principales métodos de detección

Acotación de pasajes

Analizar en más detalle los pasajes plagiados y determinar sus fronteras.

Postproceso

Filtrar los pasajes de la anterior fase, eliminando o fusionando casos de pasajes cortos y superposición o detecciones ambiguas.

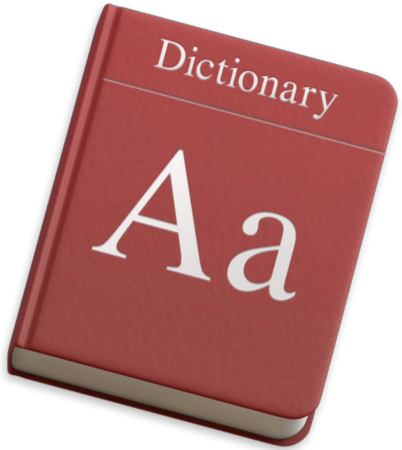
4.1. BASADOS EN EL CONTENIDO

4.1. Basados en el contenido

Se descartan las stopwords porque no aportan significado por ellas mismas, pero se pueden utilizar para detectar otras palabras relevantes en el texto.

Sin embargo, para los casos de detección de plagio, está demostrado que eliminar las stopwords afecta negativamente al funcionamiento.

Modelo ejemplo: **PAN-10-1**



4.1. Basados en el contenido

PAN-10-1

Representación del texto	Búsqueda de candidatos	Acotación de pasajes	Postproceso
Bloques de palabras de k-gramas ordenadas, con uso de hashes	Se eligen los que más bloques en común tengan con el documento original	Se aplican heurísticas	Se eliminan las detecciones con pocas palabras en común

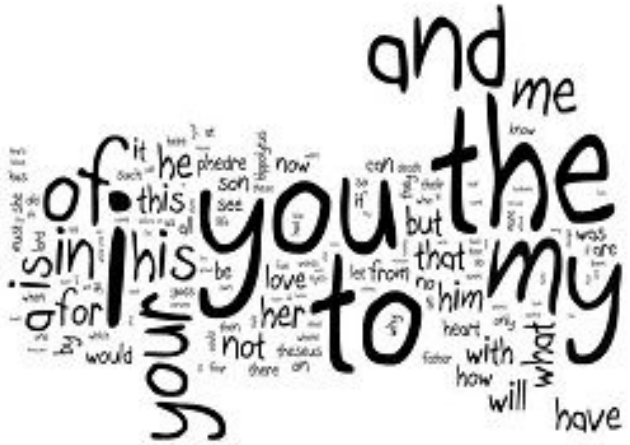
4.2. BASADOS EN LA ESTRUCTURA

4.2. Basados en la estructura

Los plagios más comunes suelen depender de los sinónimos.

En consecuencia, la estructura principal de las oraciones no se ve alterada, hecho que puede ser explotado a través de las stopwords.

Modelo ejemplo: **SWNG**



4.2. Basados en la estructura

Stopwords N-grams (SWNG)

Representación del texto	Búsqueda de candidatos	Acotación de pasajes	Postproceso
N-gramas de stopwords	N-gramas comunes en documento original y sospechoso	Representación gráfica de los n-gramas comunes, eliminando n-gramas muy frecuentes	Comprobación de los pasajes detectados y puntuarlos según la similitud de los pasajes

5. CONCLUSIÓN

5. Conclusión

Conclusions

Plagiarism detection in large document collections should be both efficient and effective. The former requires that the measures used to represent documents are easily available and capture local similarities so as to enable the identification of a short plagiarized passage within a long document. Moreover, the document representation measures should be flexible in modifications intentionally made by plagiarists to hide the similarity with the original passages. In contrast to the vast majority of the existing approaches that are (entirely or in part) based on content terms, in this paper we presented a method that uses only a small list of stopwords to represent documents. It has been demonstrated that the stopword n-gram method is reliable when it is used to identify similarity in the document level as well the exact passage boundaries in the plagiarized and the source documents.

Experiments using publicly available corpora for plagiarism detection show that the performance of the present method is very competitive when compared with methods based on content information. Interestingly, the proposed

4. Conclusiones

En este capítulo hemos presentado el panorama actual de los modelos automáticos para la detección de plagio, una tarea en la que se combinan métodos de recuperación de información y procesamiento del lenguaje natural. Desde la perspectiva de este último, y teniendo en cuenta el planteamiento que acabamos de exponer, podemos considerar que el plagio es la aplicación de mecanismos parafrásticos orientados a un determinado fin: copiar lo que han escrito otros autores pero procurando que el lector no lo note. Así, la paráfrasis está en la base de distintos tipo de plagio.

ESPAÑOL INGLÉS FRANCÉS

La detección de plagio en grandes colecciones de documentos debería
ser a la vez eficiente y efectivo. El primero requiere que el
las medidas utilizadas para representar documentos están
fácilmente disponibles
y capturar similitudes locales para permitir la identificación
de un pasaje corto plagiado dentro de un documento largo.
Además, las medidas de representación de documentos deben ser

5. Conclusión

Se precisa de herramientas automáticas que actúen de manera eficaz y eficiente, por lo que surgen diferentes planteamientos.

Existen diferentes metodologías que, pese a basarse en principios distintos, ofrecen resultados realmente positivos.

No obstante, con el fin de obtener los mejores resultados se pueden combinar los diferentes tipos de modelos y así contrarrestar los puntos débiles de cada uno.

Bibliografía

- **Stamatatos, Efstathios**. (2011). Plagiarism Detection Using Stopword n-grams. Journal of the American Society for Information Science and Technology. 62. 2512 - 2527. 10.1002/asi.21630

- **Barrón-Cedeño, Alberto & Vila, Marta & Rosso, Paolo**. (2010). Detección automática de plagio: de la copia exacta a la paráfrasis

- **Mohamed El Bachir Menai** (2012) Detection of Plagiarism in Arabic Documents. Department of Computer Science, College of Computer and Information Sciences, King Saud University.

¡Muchas gracias por vuestra atención!