UNIVERSITAT POLITÈCNICA DE VALÈNCIA

SISTEMAS DE ALMACENAMIENTO Y RECUPERACIÓN DE INFORMACIÓN

Detección de Plagio en texto

Grado en Ingeniería Informática

Escuela Técnica Superior de Ingeniería Informática

Curso 2020-21

Autores:

- Lishuang Sun (María)
- Antonio José Romero Barberá
- Vicent González Gramage
- Fabián Scherle Carboneres

ÍNDICE

1- Introducción.

El término de plagio queda definido según La Real Academia Española en 2008 como el acto de "*copiar en lo sustancial obras ajenas, dándolas como propias*". Pero claro, el concepto propio de una obra puede ser ambiguo, ya que una obra puede tratarse desde una pintura, una pieza musical, una fotografía hasta un fragmento de texto o un código de programación. La definición más completa que se considera actualmente está establecida por la IEEE en 2008:

plagiar es reusar las ideas, procesos, resultados o palabras de alguien más sin mencionar explícitamente a la fuente y a su autor.

Dada la enorme cantidad de documentos existentes, y disponibles, la detección manual de plagio resulta imposible, por lo que es necesario desarrollar herramientas computacionales que asistan al ser humano en esta tarea: los detectores automáticos de plagio. Esta tarea de desarrollar herramientas de detección de plagio es una tarea tanto del **procesamiento del lenguaje natural** (PLN) como de la **recuperación de información** (RI).

2- Desarrollo

2.1- Procesos

Existen diferentes tipos de plagios, no es igual plagiar un código de programación de forma que copiamos y pegamos el código que por ejemplo la idea de cómo y porqué se ha desarrollado este código. Con esta idea en mente vamos a distinguir distintos tipos de plagio, mencionados y comentados en el documento *"Detección automática de plagio: de la copia exacta a la paráfrasis *"* (Barrón Cedeño et al., n.d.):

1.de ideas; Este tipo de plagio consiste en que el autor de un documento, ya sea plagiado o original, adopta las ideas, pensamientos o teorías de otra persona sin darle crédito.

2.palabra por palabra; Este plagio consiste en copiar una frase, el autor del documento puede realizar una copia exacta o efectuar modificaciones de esa copia. Este tipo de plagio es un poco más complicado puesto que depende de la sintaxis, se verá posteriormente en el plagio por paráfrasis.

3.de fuentes; Este plagio consiste en que el autor del documento incluye las referencias bibliográficas que otro autor haya incluido en su propio documento, pero, el autor del documento plagiado no señala que las referencias que ha obtenido son extraídas de otro documento.

4.de autoría; Este plagio consiste en que el autor del documento presume ser el autor de un documento que en realidad ha sido escrito por otra persona.

Ahora que ya sabemos que tipos de plagios existen, vamos a explicar como funciona la detección de plagio hoy en día con internet.

Hoy en día, con internet, hay miles de millones de documentos y para poder detectar plagios de forma efectiva se utiliza un proceso dividido en tres fases: Dados un documento potencialmente sospechoso y una colección de documentos D (D suele ser en la actualidad internet por ejemplo, o una colección enorme de documentos). Primero, se realiza una recuperación heurística, esta operación nos devuelve aquellos documentos en D que son más similares al documento sospechoso. Segundo, se realiza una comparación exhaustiva, donde el documento sospechoso se compara con cada uno de los documentos de la

subcolección que hemos recuperado en el primer paso dando como resultado conjuntos de pares de fragmentos de texto. Y por último, tercero, se realiza un postprocesamiento donde se diferencian los casos donde sean o no verdaderos casos de plagio, comprobando las referencias incluidas y otros puntos similares.

2.2- Tipología de plagios

Ahora que sabemos los diferentes tipos de plagios y también cómo funciona la detección hoy en día por internet, podemos considerar los modelos de detección automática relevantes que propuso Maurer, descritos como diferentes tipologías. Estas tipologías se plantean y explican en el documento *"Detección automática de plagio: de la copia exacta a la paráfrasis *"* (Barrón Cedeño et al., n.d.) por *Alberto Barrón-Cedeño, Marta Vila y Paolo Rosso*.

Para hacer una explicación más didáctica tomaremos un fragmento de un supuesto documento original: "El androide fue destruido con éxito antes del amanecer".

2.2.1- Copia exacta

En este tipo de plagio se copia el mismo fragmento del documento original sin realizar ningún cambio. Usando el ejemplo definido previamente como un fragmento de un documento original, se tiene el plagio: "El androide fue destruido con éxito antes del amanecer". Ambos fragmentos son el mismo.

Una forma para detectar este tipo de plagios tomando como base un gran número de documentos, se basa en dividir cada documento de la colección en fragmentos, donde a cada fragmento se le aplicará una determinada función hash capaz de generar un valor numérico que puede ser considerado único dependiendo de la función utilizada. Los fragmentos podrían ser bien oraciones o conjuntos de palabras (n-gramas).

Una vez obtenidos los valores para todos los fragmentos, se guardan en una base de datos.

Para ilustrar un ejemplo, usamos el valor hash que se obtiene al aplicar el algoritmo Karp-Rabin (*Algoritmo De Rabin-Karp Usando Hash Polinomial Y Aritmética Modular*, n.d.) mostrado en las siguientes imágenes:

$$H = c_1 \times b^{m-1} + c_2 \times b^{m-2} + c_3 \times b^{m-3} \ldots + c_m \times b^0$$

c = caracteres en la cadena m = longitud de la cadena b = constante

Ecuación para obtener valor hash [2]

```
1   def KarpAndRabin(cadena,base) :
2       res = 0
3       for i in range(len(cadena)) :
4           res += ord(cadena[i]) * pow(base, len(cadena)-i-1)
5       print(str(res))
```

Código en python de la ecuación anterior

Tomando en cuenta el fragmento sospechoso "El androide fue destruido con éxito antes del amanecer" y el fragmento en la base de datos "El androide fue destruido con éxito antes del amanecer" con un valor hash asociado de 53197840388956338096808076102928588273124624782407802264268725281413320609405 3 usando el código python con base 27, obtenemos el valor hash del fragmento sospechoso. En este caso coinciden al comparar ambos fragmentos, por lo que se detecta una **copia exacta**.

Cabe destacar que cambiar un carácter del fragmento hará que el valor de la función hash cambie. Por ejemplo: "El androide fue destruido con éxito antes **de** amanecer", generará 19702903847761606702521509667751328990046157326817704542321750083446917556075, este valor es diferente al que se encuentra en la base de datos para el fragmento del documento original.

2.2.2- Copia modificada

En el caso que se realicen ligeras modificaciones en un fragmento no es posible modelos de *fingerprinting*, como el explicado anteriormente. Por

lo que se busca estimar la similitud entre dos fragmentos a través de diferentes medidas, como la similitud coseno o el coeficiente de Jaccard.

Considerando el coeficiente de Jaccard:

$$\mathcal{J}(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

Fórmula del coeficiente de Jaccard

Siendo A un conjunto de elementos del fragmento del documento original: "El androide fue destruido con éxito antes del amanecer" y B un conjunto de elementos de un fragmento sospechoso: "El androide fue **exterminado** con éxito **previo al** amanecer", se representan los elementos de A y B de las siguientes formas:

- Elementos que constituyen palabras:

  A = {El, androide, fue, destruido, con, éxito, antes, del, amanecer}

  B = {El, androide, fue, exterminado, con, éxito, previo, al, amanecer}

  A ∩ B = {El, androide, fue, con, éxito, amanecer}

  A U B = {El, androide, fue, destruido, exterminado, con, éxito, antes, previo, del, al, amanecer}

  -> similitud = 6/12 = 0,5

- Elementos formados por n-gramas, considerando 2-gramas:

  A = { (El, androide), (androide, fue), (fue, destruido), (destruido, con), (con, éxito), (éxito, antes), (antes, del), (del, amanecer) }

  B = { (El, androide), (androide, fue), (fue, exterminado), (exterminado, con), (con, éxito), (éxito, previo), (previo, al), (al, amanecer) }

  A ∩ B = { (El, androide), (androide, fue), (con, éxito) }

  A U B = { (El, androide), (androide, fue), (fue, destruido), (fue, exterminado), (destruido, con), (exterminado, con), (con, éxito), (éxito, antes), (éxito, previo), (antes, del), (previo, al), (del, amanecer), (al, amanecer) }

  -> similitud = 3/13 = 0,23

2.2.3- Plagio traducido

Este tipo de plagio consiste en la reutilización de una frase de un documento que proviene de una lengua como el español, y traducirla a otro idioma como el inglés.

Una posible solución para detectar este tipo de plagio podría consistir en traducir las frases por medio de alguna herramienta como un traductor online (translate.google.com) y posteriormente aplicar cualquiera de las técnicas explicadas anteriormente.

Sin embargo, hay que tomar en cuenta que los traductores suelen cometer errores. Por ello existen alternativas, de los cuales destacamos los tres modelos siguientes (Barrón Cedeño et al., n.d.):

- CL-ESA (Cross-Language Explicit Semantic Analysis): el fragmento de texto sospechoso y el fragmento del documento original son comparados con un conjunto de artículos que se relacionen por el mismo tema, estos artículos pertenecen a wikipedia.
- CL-ASA (Cross-Language Alignment-based Similarity Analysis): de todas las palabras del documento sospechoso se consideran las posibles traducciones en un diccionario probabilístico estimado, con el objetivo de reducir el error de traducción. En este modelo se consideran las longitudes de los fragmentos de texto, puesto que algunos fragmentos al ser traducidos serán más largos.
- CL-CNG (Cross-Language Character n-grams): Consta de tres pasos, para facilitar la explicación se utilizará un ejemplo con un fragmento sospechoso en inglés "The android was successfully destroyed before dawn" y el fragmento original utilizado en apartados anteriores:
  - Se descartan espacios y signos de puntuación y se pasa a minúsculas.

    "elandroidefuedestruidoconéxitoantesdelamanecer"

    "theandroidwassuccessfullydestroyedbeforedawn"

  - Se obtiene un conjunto de 3-gramas por caracteres.

[('e','l','a'),('l','a','n'),('a','n','d'),('n','d','r'),('d','r','o'),('r','o','i'),('o','i','d'),('i','d','e'),('d','e','f'),('e','f','u'),('f','u','e'),('u','e','d'),('e','d','e'),('d','e','s'),('e','s','t'),('s','t','r'),('t','r','u'),('r','u','i'),('u','i','d'),('i','d','o'),('d','o','c'),('o','c','o'),('c','o','n'),('o','n','é'),('n','é','x'),('é','x','i'),('x','i','t'),('i','t','o'),('t','o','a'),('o','a','n'),('a','n','t'),('n','t','e'),('t','e','s'),('e','s','d'),('s','d','e'),('d','e','l'),('l','a','m'),('a','m','a'),('m','a','n'),('a','n','e'),('n','e','c'),('e','c','e'),('c','e','r')]

[('t','h','e'),('h','e','a'),('e','a','n'),('a','n','d'),('n','d','r'),('d','r','o'),('r','o','i'),('o','i','d'),('i','d','w'),('d','w','a'),('w','a','s'),('a','s','s'),('s','s','u'),('s','u','c'),('u','c','c'),('c','c','e'),('c','e','s'),('e','s','s'),('s','s','f'),('s','f','u'),('f','u','l'),('u','l','l'),('l','l','y'),('l','y','d'),('y','d','e'),('d','e','s'),('e','s','t'),('s','t','r'),('t','r','o'),('r','o','y'),('o','y','e'),('y','e','d'),('e','d','b'),('d','b','e'),('b','e','f'),('e','f','o'),('f','o','r'),('o','r','e'),('r','e','d'),('e','d','a'),('d','a','w'),('a','w','n')]

- El conjunto obtenido se compara usando una medida de similitud, como las mencionadas anteriormente. En este caso consideramos el coeficiente de Jaccard sobre los conjuntos de trigramas obtenidos.

  -> similitud = 8 / 77 = 0.104

  Se aprecia como resultado un valor bajo, sin embargo, estos valores logran diferenciar fragmentos de textos que son plagios traducidos de los que no lo son.

En el último modelo hay que tener en cuenta que no se pueden considerar lenguas con diferentes alfabetos, por ejemplo: español y japonés.

2.2.4- Paráfrasis

La tipología de plagio correspondiente a la paráfrasis se ha convertido en un fenómeno de alta complejidad, implicando una alta variedad de conocimientos morfológicos, léxicos, sintácticos, semánticos y pragmáticos. Por lo que métodos utilizados en los tipos de plagios expuestos anteriormente no resultan útiles en este caso.

Pretendemos dar una visión amplia sobre la tipología de paráfrasis sin entrar en detalle, concretamente enunciamos cinco tipos según la operación realizada para generar una nueva frase a partir de otra original:

- Sustitución de una o varias piezas léxicas por otras:

  Mediante el uso de sinónimos, antónimos, piezas léxicas más genéricas y/o piezas léxicas que representan uno de los actores que realizan la acción de la pieza léxica a sustituir.

- Eliminación de una o varias piezas léxicas:

  Se elimina el contenido no proposicional, elementos adicionales del predicado y/o argumentos pertenecientes al verbo de la oración.

- Transformación en la estructura de la oración, por ejemplo pasar de voz activa a voz pasiva o viceversa.
- Segmentación de una oración en varias oraciones.
- Cambio de orden en la estructura de las piezas léxicas.

Hay que destacar que estas operaciones requieren que los fragmentos de texto modificados cambien ligeramente algunas palabras para que estos tengan sentido al ser leídos.

2.3- Detectores de plagio alternativos

2.3.1- Contenido del texto: términos

La indexación clásica de una colección de documentos consiste en la extracción de términos para formar su vocabulario.

Es una decisión polémica la de que los stopwords formen parte de ese vocabulario, ya que son útiles para mantener la semántica de los sintagmas pero requieren mucho espacio de almacenamiento. (Existen métodos de compresión de posting lists eficientes).

La virtud de este tipo de índices se encuentra a la hora de hacer búsquedas de documentos según los términos de las consultas. Sin embargo, en el ámbito de detección de plagio en documentos altamente modificados puede no ser el ideal; habría que encontrar e incluir en el índice todos los sinónimos de todos los términos obtenidos, entre otros aspectos.

2.3.2- Estructura del texto: *stopwords*

El contenido del texto se puede cambiar fácilmente por sustitución a sinónimos, sin embargo, se ha comprobado que la estructura de un texto se mantiene y es muy difícil de modificar. De modo que, con una lógica contraria a los métodos clásicos, se plantea un método de detección de plagio basado en la conservación exclusiva de *stopwords* de la colección de documentos para la creación del índice invertido, priorizando así las similitudes sintácticas en lugar de las semánticas.

Nos basamos en el artículo "*Plagiarism Detection Using Stopword n-grams"* (Stamatatos, n.d.) para explicar este método original, al que llamaremos *SWNG* (StopWords N-Grams) .

En este método, cada documento *d* se convierte en un conjunto de n-gramas de stopwords, que será el perfil del documento, *P(n,d)*.

La detección de plagio se compone de las siguientes fases:

1. Recuperación de documentos candidatos

Tenemos dos colecciones de documentos: sospechosos (*Dx*) y fuentes (*Ds*).

Dado un documento sospechoso d_x, se pretende recuperar su conjunto de documentos fuente. Todo documento fuente d_s que tenga algún n-grama en común con d_x será recuperado.

Para evitar o reducir los falsos positivos (d_s seleccionado aunque no sea similar a d_x), necesitaremos que $n >= n_1$, siendo n_1 un valor entero bien escogido; un valor demasiado bajo puede dar falsos positivos, uno demasiado alto puede descartar documentos fuentes que nos interesan si el fragmento plagiado contiene menos de n_1 stopwords en secuencia o ha sido muy modificado (*verbatim plagiarism*).

Por tanto, se necesita una restricción adicional:

$$\exists\, g \in P(n_1, d_x) \cap P(n_1, d_s) : \text{member}(g,C) < n_1 - 1 \ \wedge \ \text{maxseq}(g,C) < n_1 - 2$$

Si se sospecha que d_x ha plagiado a d_s (similitud), entonces necesariamente existe un *n_1-grama* g perteneciente a los perfiles de ambos documentos que, dado un conjunto C de los stopwords más frecuentes, hayan menos de n_1-1 stopwords en g que también estén en C. Y la longitud de la secuencia más larga de stopwords en g pertenecientes a C debe ser menor que n_1-2.

2. Detección de límites de fragmentos plagiados

Tenemos un documento sospechoso d_x y su conjunto D_{rx} de documentos fuente d_s recuperado en la fase anterior.

En esta 2ª fase necesitamos que los documentos estén representados por los perfiles $P(n_2, d_x)$ y $P(n_2, d_s)$, siendo $n_2<n_1$, para capturar los detalles de similitud. Diremos que un fragmento representado por un n_2-grama g es un plagio si:

$$g \in P(n_2, d_x) \cap P(n_2, d_s) \bigwedge member(g,C) < n_2$$

Esto es, este n_2-grama pertenece a los perfiles de ambos documentos y tiene menos de n_2 stopwords pertenecientes al conjunto C.

Tras la comparación de cada fragmento de d_x con cada uno en d_s, habremos obtenido un conjunto de n_2-gramas iguales en ambos documentos, al que llamaremos $M(d_x,d_s)$. Para entenderlo mejor, usaremos el siguiente ejemplo:

This came into existence likely from the deviance in the time-period of the particular billet. As the premier is to be nominated for not more than a period of four years, it can infrequently happen that an ample wage, fixed at the embarkation of that period, will not endure to be such to its end.

(a) The plagiarized passage.

This probably arose from the difference in the duration of the respective offices. As the President is to be elected for no more than four years, it can rarely happen that an adequate salary, fixed at the commencement of that period, will not continue to be such to its end.

(b) The original passage.

El primer párrafo es el fragmento del documento sospechoso d_x; el segundo párrafo es el fragmento de d_s. A continuación, sus correspondientes secuencias de stopwords:

this, from, the, in, the, of, the, as, the, is, to, be, for | not, a, of | it, can, that, an, at, the, of, that, will, not, to, be, to

this, from, the, in, the, of, the, as, the, is, to, be, for | it, can, that, an, at, the, of, that, will, not, to, be, to

Podemos apreciar que el fragmento plagiado mantiene bastante bien la estructura del fragmento original, a diferencia de una porción pequeña marcada en rojo. Los 8-gramas correspondientes, una columna por documento, son:

1.	[this, from, the, in, the, of, the, as]	[this, from, the, in, the, of, the, as]
2.	[from, the, in, the, of, the, as, the]	[from, the, in, the, of, the, as, the]
3.	[the, in, the, of, the, as, the, is]	[the, in, the, of, the, as, the, is]
4.	[in, the, of, the, as, the, is, to]	[in, the, of, the, as, the, is, to]

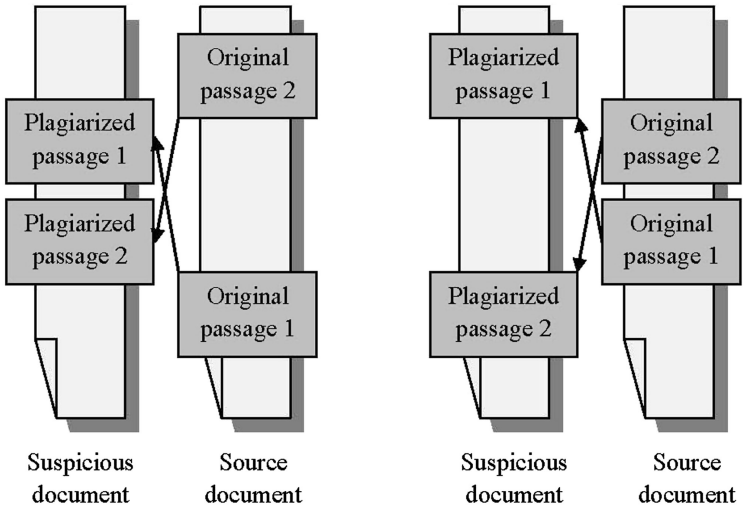
5.	[the, of, the, as, the, is, to, be]	[the, of, the, as, the, is, to, be]
6.	[of, the, as, the, is, to, be, for]	[of, the, as, the, is, to, be, for]
7.	[the, as, the, is, to, be, for, not]	[the, as, the, is, to, be, for, it]
8.	[as, the, is, to, be, for, not, a]	[as, the, is, to, be, for, it, can]
9.	[the, is, to, be, for, not, a, of]	[the, is, to, be, for, it, can, that]
10.	[is, to, be, for, not, a, of, it]	[is, to, be, for, it, can, that, an]
11.	[to, be, for, not, a, of, it, can]	[to, be, for, it, can, that, an, at]
12.	[be, for, not, a, of, it, can, that]	[be, for, it, can, that, an, at, the]
13.	[for, not, a, of, it, can, that, an]	[for, it, can, that, an, at, the, of]
14.	[not, a, of, it, can, that, an, at]	[it, can, that, an, at, the, of, that]
15.	[a, of, it, can, that, an, at, the]	[can, that, an, at, the, of, that, will]
16.	[of, it, can, that, an, at, the, of]	[that, an, at, the, of, that, will, not]
17.	[it, can, that, an, at, the, of, that]	[an, at, the, of, that, will, not, to]
18.	[can, that, an, at, the, of, that, will]	[at, the, of, that, will, not, to, be]
19.	[that, an, at, the, of, that, will, not]	[the, of, that, will, not, to, be, to]
20.	[an, at, the, of, that, will, not, to]	
21.	[at, the, of, that, will, not, to, be]	
22.	[the, of, that, will, not, to, be, to]	

Y los índices de los 8-gramas que pertenecen al conjunto M(d_x,d_s) son:

M(d_x,d_s) = {(1, 1), (2, 2), (3, 3), (4, 4), (5, 5), (6, 6), (17, 14), (18, 15), (19, 16), (20, 17), (21, 18), (22, 19)}

También podemos observar que el fragmento plagiado se divide en dos sub-fragmentos bastante cercanos, dado que la estructura del texto ha cambiado ligeramente.

Esto nos sugiere que, si bien los fragmentos de cada documento fuente d_s han sido altamente modificados en el documento sospechoso d_x, varios fragmentos en d_x pueden estar plagiando a varios fragmentos de d_s, dándose también los casos en que dichos sub-fragmentos sean muy cercanos entre sí o no:



En el primer caso, para reconocer esos sub-fragmentos cercanos en d_x como un fragmento global (límites inicio-fin), se cumple:

$$m_i \in M1(d_x, d_s) : abs(diff(m_i)) > \theta_g$$

Siendo m_i cada fragmento en $M1$, que es el subconjunto de M que corresponde a d_x, entonces el valor absoluto de la diferencia derivativa de m_i debe ser mayor a un umbral θ_g que permite pequeños cambios de sintaxis en dicho fragmento.

Dichos fragmentos m_i también pertenecen al subconjunto $M2$ correspondiente a cada d_s, sin embargo, puede que estos estén muy separados entre sí en el documento original, por lo que dado $M2_i \subseteq M2$, se busca el gran fragmento en el documento original que alberga el conjunto de fragmentos cercanos de $M1$:

$$m_i \in M2_i(d_x, d_s): abs(diff(m_i)) > \theta_g$$

3. Postprocesamiento

Una vez detectados todos los fragmentos sospechosos de plagio en d_x, cada uno de ellos emparejados con los fragmentos originales en d_s, se necesita un postprocesamiento para detectar el grado de similitud de cada par asignándoles una puntuación y asegurar que sean muy similares para finalmente declararlo plagio.

$$Sim(t_x, t_s) = \frac{|P_c(n_c, t_x) \cap P_c(n_c, t_s)|}{\max(|P_c(n_c, t_x)|, |P_c(n_c, t_s)|)}$$

Siendo t_x uno de esos fragmentos en d_x y t_s uno de esos en d_s, para cada fragmento se extraen los n-gramas de caracteres (en minúsculas y sin signos de puntuación), creando los perfiles $P_c(n_c, t_x)$ y $P_c(n_c, t_s)$. La similitud entre estos dos fragmentos viene dada por el cociente de la longitud de la intersección de sus perfiles y la longitud del perfil más largo.

3- Conclusión

A modo de conclusión, consideramos que la detección del plagio representa un problema para la informática, debido al amplio abanico de posibilidades en las que este podría presentarse. Por ello, es necesario conocer de la existencia de los tipos de plagio en los que un texto sospechoso podría haber incurrido, y así analizar los correspondientes métodos para detectarlos.

Apreciamos que dentro de la tipología de plagio no suele haber un método definitivo de detección, y esto es especialmente estimable en el caso de la paráfrasis, donde hay diferentes aproximaciones dependiendo del caso específico usado en el plagio.

También debemos valorar los detectores más modernos, como aquellos enfocados en los términos de los documentos o incluso en los stopwords, los cuales han dado muy buenos resultados en diferentes análisis.

Finalmente, remarcar que si bien ya se ha realizado mucho trabajo en este campo, los métodos actuales no son infalibles y se sigue haciendo todo lo posible para poder conseguir los mejores detectores de plagios posibles en un futuro próximo.

4- Resumen

Al inicio se presenta qué es un plagio, además de la existencia de detectores automáticos de este. Posteriormente se presentan los tipos de plagio, mientras observamos el funcionamiento habitual de un detector de plagios. A continuación observamos las distintas tipologías de plagios comunes, así como varias formas de detectarlas con ejemplos. Finalmente

descubrimos la existencia de detectores de plagio alternativos, siendo estos o bien enfocados en el contenido del documento, o bien en su estructura.

5- Bibliografía:

Algoritmo de Rabin-Karp usando hash polinomial y aritmética modular. (n.d.). ICHI.PRO. Retrieved 6 1, 2021, from https://ichi.pro/es/algoritmo-de-rabin-karp-usando-hash-polinomial-y-aritmetica-modular-74174751276470

Barrón Cedeño, A., Vila, M., & Rosso, P. (n.d.). *Detección automática de plagio: de la copia exacta a la paráfrasis **. personales.upv.es. Retrieved 6 2, 2021, from http://personales.upv.es/prosso/resources/BarronEtAl_JLF10.pdf

Stamatatos, E. (n.d.). *Plagiarism Detection Using Stopword n-grams*. Wiley Online Library. Retrieved 6 1, 2021, from http://onlinelibrary.wiley.com/doi/10.1002/asi.21630/pdf