Estudio de las técnicas de detección de plagio textual y análisis de sinonimia en ensayos y desarrollo de un sistema prototipo

Andrea Elizabeth Flores Vega

Benito Bernardo León Ullauri

Autores

Dirigido por:

Ing. Vladimir Robles Bykbaev

Ingeniero de Sistemas

Docente de la Universidad Politécnica Salesiana

Estudio de las técnicas de detección de plagio textual y análisis de sinonimia en ensayos y desarrollo de un sistema prototipo

Resumen. El presente trabajo se enfoca en el desarrollo de una aplicación que colaborará con la detección de plagio a nivel académico, el mismo que se enfoca en analizar ensayos, para ello aplica técnicas que permiten realizar análisis por sinonimia y análisis textual, además de utilizar técnicas de similitud que permiten determinar el porcentaje de plagio que existe en el documento. Estos resultados pueden ser observados en un archivo que la aplicación entrega al usuario sobre análisis realizado durante el proceso.

1 Introducción.

Plagiar es: "Copiar en lo sustancial obras ajenas, dándolas como propias" [1].

El plagio es una actividad que está presente en diversos ámbitos de la vida, puede ir desde el ámbito laboral hasta el ámbito educativo, sobre este último trataremos a mayor profundidad en este trabajo.

La educación es uno de los principales pilares del desarrollo personal de un ser humano, sin embargo, este se ve afectado debido a la ausencia de valores y respeto por las demás personas en diversos aspectos entre estos el intelectual, ya que en la actualidad el plagio se ha visto arraigado en el ámbito académico, siendo muchas veces poco detectado por los educadores, debido a la habilidad de los estudiantes para llevar a cabo este actividad. Actualmente es una actividad muy común que empieza desde las instituciones de educación básica hasta las grandes universidades, donde se ha ido perfeccionando la habilidad para llevar a cabo el plagio, surgiendo así un sinnúmero de tipos de plagio como son el parafraseo, la sinonimia, la copia textual, etc.

Entre los causantes de este tipo de actitud están el acceso ilimitado a grandes cantidades de información, además que los estudiantes en la actualidad llevan una formación poco investigativa, ya no se molestan en leer, en entender, sino se limitan a copiar y pegar información de internet o de libros sin sus debidas referencias. Otro causante puede ser la falta de tiempo para poder llevar a cabo todas las actividades encomendadas a los estudiantes, además de la ética de la persona y el respeto por lo ajeno son valores que se construyen a lo largo de una vida, es por ello que las instituciones educativas deben formar a las estudiantes de una manera integral y a su vez presentar ciertas normativas que dentro de la institución penalicen el plagio.

Existen ciertas instituciones educativas que incorporan reglamentos en caso de suscitarse situaciones de plagio, como es la Universidad de Londres, que al momento de ingresar el estudiante a la universidad firma un contrato en el que se le especifica que

cualquier tipo de plagio encontrado a lo largo de su vida estudiantil puede incurrir en procesos legales y financieros [2].

Sin embargo, existen casos de plagio que han llegado a tener serios conflictos, un ejemplo de esto es el caso de Karl Theodor Zu Guttenberg quien asistió a la Universidad de Bayreuth para hacer su doctorado en derecho, además de ocupar un excelente cargo como Ministro de Defensa, Zu Guttenberg obtuvo su título de doctorado con honores, pero se realizaron investigaciones que demuestran que su tesis tiene una gran cantidad de plagio, ya que existen ideas que no fueron planteadas por él y que no llevan las debidas referencias como lo indicaba el reglamento de la universidad, ante esta acusación lo que se alegó fue la falta de tiempo y la presión que se tuvo para mantener de manera satisfactoria tanto su carrera como el ejercicio de su profesión. Este hecho tuvo repercusiones a nivel laboral y a nivel académico, ya que le fue quitado su título de doctorado, y en sí su imagen se vio devastada por este suceso [3].

El plagio se ha convertido en un daño que no tiene únicamente repercusiones a nivel académico sino que ha generado un conflicto social ¿Los profesionales que se forman en la actualidad son lo suficientemente aptos para satisfacer las expectativas del sector empleador? La respuesta salta a la vista, para una empresa no basta con saber que una persona tiene un título sobre sus conocimientos adquiridos, sino es necesario corroborar que aquello que se dice sea cierto, es por esto que en la actualidad la mayoría de empresas admiten a sus empleados siempre y cuando aprueben ciertos prerrequisitos tales como pruebas de aptitud, de conocimientos, y es que no es solo una secuela a nivel personal sino a nivel educativo ya que se crea una falta de confianza en las instituciones educativas y la forma en que estas emiten sus conocimientos hacia su población estudiantil.

Esta falta de conciencia social ha traído consigo una diversidad de prácticas de plagio, es por ello que intentamos contrarrestar este tipo de actividades, pero para ello es necesario en primera instancia realizar un análisis sobre los posibles tipos de plagio que se utilizan en la actualidad [6]:

- **Plagio por traducción:** se puede considerar plagio el traducir un texto y tomarlo como propio, por lo general el idioma utilizado es el inglés.
- **Detección por estilometría**: trata de encontrar inconsistencia en el estilo de escritura.
- **Copia textual:** como su nombre indica, consiste en copiar textualmente pensamientos, textos, etc., de otros autores y no realizar las debidas referencias.
- **Copia modificada:** consiste en trabajar sobre una copia textual y sobre esta realizar ciertas modificaciones del texto, estas modificaciones pueden ser reemplazo de palabras por:
  - **Sinónimos**: es el reemplazo de una palabra por un sinónimo.
  - **Antonimia**: reemplazo de una palabra por su antónimo.
  - **Generalización**: se considera el reemplazo de una palabra por una de uso más común.
  - **Sustitución palabra por definición**: Consiste en reemplazar una palabra por su respectivo significado.

- **Plagio por eliminación**: Consiste en trabajar sobre una copia textual e ir eliminando ciertas palabras.
- **Plagio por segmentación**: está constituido por una copia de plagio textual, sin embargo el cambio radica cuando se utilizan signos de puntuación para separar ciertas ideas.
- **Plagio por paráfrasis**: consiste en cambiar el orden original de la frase, haciendo que esta se re-estructure sin perder el significado.

Como se ha podido observar existen varias formas de plagiar y se ha visto la necesidad de crear herramientas que tengan en consideración los puntos que se mencionaron con anterioridad, sin embargo en la actualidad existen páginas web y programas que son de libre uso que no satisfacen todas las expectativas planteadas.

2 Modelo propuesto

En base a la diversidad de tipos de plagio que existen en la actualidad hemos decidido enfocarnos en dos tipos de plagio como son el plagio de forma textual y el plagio por sinonimia, para ello hemos planteado la siguiente metodología en nuestro sistema:

Plagio Textual.

Para detectar este tipo de plagio hemos decidido dividir el texto a analizar en bloques de palabras y proceder a realizar búsquedas mediante el buscador Bing de Microsoft. Para realizar las búsquedas encerramos a cada bloque entre comillas, si el buscador retorna un resultado almacenamos la URL resultante para que sea mostrada en el informe final junto con el bloque de palabras correspondiente. En caso de que no encuentre un resultado se procede a analizar el siguiente bloque de palabras. Nuestro sistema incluye una opción para "forzar" la búsqueda de un bloque específico. Consiste en eliminar iterativamente la última palabra de un bloque de palabras hasta encontrar plagio. Evidentemente no podemos eliminar todas las palabras del bloque o dejar en un número pequeño de palabras consecutivas puesto que nos devolverían falsos positivos, por lo tanto debemos imponernos un límite: Teóricamente se sugiere que sean bloques de mínimo 4 palabras [7], sin embargo en nuestras pruebas encontramos que un número adecuado de palabras consecutivas que pueden considerarse como plagio es de 16. Por lo tanto, si se elige la opción de "forzar" la búsqueda textual solo eliminará palabras del bloque hasta que éste tenga al menos 16 palabras.

Plagio por Sinonimia.

Al principio, para analizar este tipo de plagio, dividíamos el documento en varias páginas, las cuales no necesariamente debían coincidir con páginas reales, sino más bien eran páginas lógicas conformadas por 50 palabras.

Entonces, para cada página, se obtiene las palabras más importantes y se buscan sinónimos para las mismas, con la finalidad de encontrar las palabras originales, supo-

niendo que dichas palabras sufrieron modificaciones previas mediante sinónimos para eludir cualquier detección de plagio.
Luego se envía a buscar este conjunto de palabras "originales" -a manera de tags- utilizando 5 buscadores[1] (Yandex, Google, Bing, Yahoo y Ask). Si un sitio web posee estas palabras en su interior se lo consideraba relevante y se procede a extraer su contenido para ser comparado con la página actualmente analizada.
Esta comparación se realiza utilizando medidas de similitud. La medida de similitud utilizada en nuestro trabajo fue el coeficiente de Overlap [12], ya que nos permite conocer "que tan" contenido está un texto pequeño (la página que se analiza) dentro de otro texto grande (el contenido web que se extrae). Dicho coeficiente nos indica en una escala de 0 a 1 el nivel de copia entre dos documentos siendo, 0 el valor que indica que no existe copia alguna y 1 el valor que indica que absolutamente todo el texto esta copiado.

De forma similar a lo que realizamos en el análisis textual, almacenamos dicho valor del coeficiente junto con el contenido de la página analizada para añadirlos al informe final. Luego se procede con la siguiente página hasta finalizar el documento.

Ahora, como decíamos, al principio obtuvimos las palabras más importantes de una página y se buscaban sus respectivos sinónimos y se realizaba la búsqueda en base a estos sinónimos. Aunque en principio esto era una buena idea, notamos que en la práctica el resultado no era el esperado, principalmente debido a que EuroWordNet, la base de datos léxica que utilizamos para encontrar los sinónimos, no poseía términos neutrales sino más bien términos españolizados; por ejemplo, para la palabra "trabajo" obteníamos el sinónimo "curro", lo cual no era deseable.

Más importante aún, notamos que las palabras principales no siempre poseían sinónimos alternativos, lo cual, significaba que si el texto sufrió cambios mediante sinónimos, éstos cambios no fueron aplicados sobre las palabras más importantes del texto, puesto que tienen pocos o nulos sinónimos candidatos, sino mas bien fueron aplicados sobre palabras menos importantes. Tomando esto en cuenta, se simplificaban las cosas debido a que no importaba si el texto contenía palabras cambiadas por sus sinónimos, lo que en verdad importaba era que el texto muy probablemente aún mantenía las palabras importantes sin cambios, por lo que se buscaron únicamente las palabras importantes, siendo así más probable dar con el documento original. Luego, el coeficiente de Overlap nos indicaría el nivel de plagio que poseía la página actual.

A medida que se fue indagando descubrimos un mejor método para detectar palabras importantes, el método Zipf [6]. Dicho método nos dice que las palabras más importantes generalmente son las palabras más largas, esto es muy fácil de comprobar especialmente en nuestro idioma en donde por lo general las palabras "importantes" son más largas (poseen mas caracteres) que las palabras "no importantes".

1 Se utilizaron 5 buscadores para balancear las solicitudes ya que si se utilizaba solo uno este detectaba un abuso de nuestra parte y nos bloqueaba sus servicios. Al usar los 5 era menos probable que esto pasara y en caso de llegar a pasar aún se puede realizar búsquedas en los demás buscadores.

Una vez que entendimos estas dos últimas ideas cambiamos el modelo que teníamos para realizar análisis por sinonimia. Ahora buscábamos las palabras importantes mediante Zipf y las enviamos como tags a los buscadores sin realizar ningún reemplazo por sinónimos. Al realizar estos cambios los resultados mejoraron mucho y los valores para el Coeficiente de Overlap eran más cercanos a 1 en textos que sabíamos que tenían plagio por sinonimia. Irónicamente el mejor enfoque para detectar plagio por sinonimia fue dejar de utilizar sinónimos.

Detección de citas referenciales.

Como se indicó en la parte introductoria un texto es considerado plagio si no posee una cita referencial que indique quién es el autor de dicho texto. Por tanto, no podemos presentar falsos positivos al ignorar dichas citas y analizar el texto que representan, ya que seguramente obtendremos plagio donde realmente si se reconoce al autor original.

Para evitar este inconveniente hemos incluido la posibilidad de identificar referencias mediante expresiones regulares. Si se identifica que un bloque de texto posee una referencia se excluye dicho bloque del análisis actual y se pasa al siguiente. Mediante la interfaz web un usuario con conocimientos puede agregar una o varias expresiones regulares, que representen como se pueden encontrar las referencias dentro del documento que se va a analizar. En caso de no poseer conocimientos sobre expresiones regulares, nosotros hemos incluido por defecto cuatro de los tipos de expresiones que para nuestro criterio son las más comunes de encontrar en cualquier trabajo, estas son:

- \\[\\d+\\]\\. Es decir un valor numérico contenido entre corchetes y finalizado en punto. Por ejemplo: [2].
- \\.\\d+ Es decir un punto seguido de un valor numérico. Por ejemplo: .4
- \\’\\d+ Es decir comillas seguidas por un valor numérico. Por ejemplo: “3
- \\.\\s*\\[\\d+\\] Es decir un punto seguido de algún o ningún espacio, finalmente seguido por un valor numérico contenido entre corchetes. Por ejemplo: . [1]

3 Experimentación

Con el objetivo de medir la eficacia y la eficiencia del sistema se realizaron pruebas que consisten en presentar varios documentos al sistema, los mismos que a su vez contendrán texto plagiado con y sin referencias. Llamaremos corpus a este conjunto de documentos de prueba. Este corpus esta compuesto por documentos que tratan diferentes temáticas, poseen diferentes extensiones de contenido, etc.

A continuación se presenta el diseño del plan de experimentación:

3.1 Resultados esperados:

Obtener porcentajes de plagio lo más afines a la realidad.
Plan de Experimentación. Variar los siguientes parámetros.

- Extensión del documento.
- Nivel de plagio en un documento.
- Temática de los documentos.
- Número de palabras por página que serán analizadas.
- Tiempo que toma en descargar un archivo.
- Tiempo que toma comparar un archivo.
- Valor de N de los N-gramas.
- Tamaño de descarga de un archivo.

Tiempo disponible:.

- Tiempo que retorne resultados.

Variables de Interés.

- Tiempo de retorno de resultados.
- Exactitud del análisis.

Perturbación.

- Tiempos de procesamiento y descargas demasiado largos.
- Conversión no satisfactoria de los documentos a texto plano.
- Referencias y bibliografía no especificadas.
- Velocidad de Conexión.

Tratamiento estadístico de los resultados.

- Análisis de precisión.
- Análisis de cobertura.
- Análisis de F-measure.
- Calculo del AVP (Average Precision)

Complejidad de la Interfaz.

La interfaz a realizar debe ser lo más sencilla posible de tal manera que sea fácil de utilizar para el usuario.

3.2 Resultados Obtenidos:

El plan de experimentación realizado permite medir el funcionamiento de nuestro prototipo. Los datos obtenidos son útiles para determinar las áreas en las que se puede mejorar el prototipo, así como también determinar la precisión del sistema. A continuación presentamos los resultados del experimento:

Documento	Categoría	Extesión (No. hojas)	N-gramas	No. Palabras /hoja	Análisis			Tiempos de Respuesta (minutos)
					Sinonimia	Textual	con reducción	
Sist_heredados_sinonimia	corto	2	4	50	si	si	si	16
sist_heredados textual	corto	2	4	50	si	si	si	12
Agentes Inteligentes sinonimia	corto	4	4	50	si	si	si	17
Agentes Inteligentes textual	corto	4	4	50	si	si	si	6
DSS_sinonimia	corto	4	4	50	si	si	si	12
DSS_textual.	corto	4	4	50	si	si	si	11
Endodoncia	corto	8	4	50	si	si	si	10

Table 1. Experimentación en Análisis por Textual y Sinonimia

Estos resultados que se pueden observar fueron obtenidos del primer prototipo propuesto, la experimentación posee las siguientes características:

Alteración del valor de n-gramas: A medida que el valor de "N" decrece el sistema encontrará un mayor número de coincidencias, sin embargo, esto puede llevar a un cálculo mayor de palabras coincidentes. Hemos creído conveniente tener un valor N que sea lo más neutral posible, en base a las pruebas y a la documentación existente sobre plagio que indica que cuatro palabras continuas que no posean referencia son consideradas como plagio [7] consideramos que el valor óptimo de N debe ser 4.

Palabras por página:. Considerando como página a un conjunto de palabras y tomando en cuenta que para cada página se realiza una búsqueda en internet, hemos probado con los siguientes números de palabras por página.

- 150
- 100
- 50

Siendo el valor óptimo de 50 palabras, ya que se generan una cantidad suficiente de páginas que a su vez representan más búsquedas en internet que sean precisas y relevantes con el contenido de la página. De esta forma, se aumentan las posibilidades de

encontrar plagio. Todo esto implica que a menor cantidad de palabras por página mayor posibilidad de aciertos.

Tamaño de descarga del archivo: Se ha considerado prudente descargar 5 megabytes tanto para archivos PDF, DOC y DOCX. Para el caso de ficheros de texto plano, los cuales contienen información que es mucho más liviana se limitó su tamaño máximo de descarga a 1 megabyte. Con estos valores hemos equilibrado la eficacia y la eficiencia del prototipo.

Tiempo de descarga de un archivo: Limitar el tamaño de descarga no es suficiente ya que pueden surgir problemas con el servidor en donde se encuentra alojado el archivo, es por ello que también nos vemos limitados a controlar el tiempo que le podría tomar al archivo en descargarse.

El tiempo que equilibra velocidad y eficacia es 2 minutos según las pruebas realizadas con los siguientes tiempos:

- 5 minutos
- 2 minutos
- 1 minuto

Tiempo para la comparación de archivos: para el plagio por sinonimia se debe realizar una comparación de archivos, para ello inicialmente se lo realizaba hasta que el proceso termine, sin embargo debido al tiempo que este podía tomar y de acorde a la extensión que tenga cada uno de ellos se ha definido un tiempo máximo de 2 minutos para la comparación entre archivos, evitando de esta manera posibles cuellos de botella que limiten la velocidad de respuesta del sistema.

Temática de los documentos: Con el desarrollo de la experimentación hemos podido notar que la temática del documento también influencia en la capacidad de respuesta que el sistema tenga, ya que entre más científico o elaborado sea un documento, mayor será el tiempo que tome en detectar la existencia de plagio.

Todos los resultados obtenidos con el primer prototipo son buenos, sin embargo, el tiempo de respuesta obtenido es bastante alto, frente a esta problemática surgió la necesidad de implementar hilos, que indudablemente agilizarán el tiempo de respuesta en los procesos, se obtuvieron los siguientes resultados:

Documento	Categoría	Extensión (No. hojas)	No. Palabras/ hoja	Análisis			Tiempos de Respuesta (minutos)
				Sinonimia	Textual	Textual con reducción	
Sist_heredados_sinonimia.doc	corto	2	50	si	si	si	3
Sist_heredados_textual.doc	corto	2	50	si	si	si	2
Agentes Inteligentes_sinonimia.docx	corto	4	50	si	si	si	4
Agentes Inteligentes_textual.docx	corto	4	50	si	si	si	3
DSS_sinonimia.docx	corto	4	50	si	si	si	1
DSS_textual.docx	corto	4	50	si	si	si	2
Motivacion_sinonimia.docx	corto	4	50	si	si	si	3
Motivacion_textual.docx	corto	4	50	si	si	si	3
Endodoncia	corto	8	50	si	si	si	4

Table 2. Análisis Textual y Sinonimia con tiempos mejorados

Como se ha podido observar los tiempos de análisis se han reducido notablemente, eso se debe a la utilización de hilos (Threads) en el programa desarrollado.

Perturbaciones encontradas

- Un factor importante para que la detección eficiente es la velocidad de conexión a internet, ya que de ser lenta al sistema le tomará más tiempo completar el análisis. Asimismo, es probable que los ficheros se descarguen mal desde internet evitando que se puedan analizar de forma adecuada.
- Con respecto a la utilización de bibliografía y referencias, podemos decir que estas no son en su totalidad acertadas, ya que existen diversas formas de anotar referencias, es por ello que se lo ha implementado como un parámetro configurable dentro del sistema.
- Los tiempos de descarga y el tamaño de los archivos demasiado extensos se han contrarrestado a través de las restricciones que se han impuesto para los tamaños

de archivos, buscando en lo posible un funcionamiento eficiente y con resultados ajustados a la temática del documento.

Como se podrá apreciar estos datos son el resultado del proceso de experimentación.

4 Resultados Obtenidos.

Se presentan los resultados en términos de Precisión, Cobertura y F-measure. Los resultados pertenecen al análisis realizado sobre tres documentos, un trabajo relacionado con odontología y dos trabajos relacionados con informática.

Se entiende por **precisión** a "la fracción de casos recuperados que son relevantes" [10], está definida por la siguiente fórmula:

$$precision = \frac{relevant\ documents\ \cap \{retrieved\ document\}}{|\{retrieved\ documents\}|} (1)$$

Cálculo de precisión [10].

Esta medida nos permite conocer en una escala de 0 a 1 el grado de falsos positivos que el sistema pueda entregar, donde 1 significa que la aplicación acertó siempre y no existe ningún falso positivo y 0 significa que todos los resultados que el sistema entregó están errados. Para un análisis de sinonimia obtuvimos los siguientes valores de precisión:

Documento	Documentos Recuperados	Documentos Relevantes	Precisión
Endodoncia	33	33	1
Sistemas heredados	5	2	0.4
Agentes Inteligentes	6	3	0.5
		Total	0.63

Table 3. Resultados de precisión para un análisis de plagio por sinonimia**.**

Como se puede ver, el trabajo sobre endodoncia obtuvo una precisión de 1, sin embargo, al sistema no le fue tan bien al analizar los otros dos trabajos; esto sucede generalmente debido a que los resultados de los buscadores no se encuentran en contexto con el documento original.

Sin embargo, en un análisis textual la precisión fué de 1 para los 3 documentos, puesto que un plagio textual es más fácil de identificar que un plagio basado en sinonimia.

Documento	Documentos Recuperados	Documentos Relevantes	Precisión
Endodoncia	40	40	1
Sistemas heredados	5	5	1
Agentes Inteligentes	6	6	1
		Total	1

Table 4. Resultados de precisión para un análisis de plagio textual

Cobertura es "la fracción de los documentos que son relevantes para la consulta que se ha recuperado correctamente" [10]. En otras palabras, es la relación entre los documentos relevantes y la totalidad de documentos que se esperaba obtener y que posiblemente no se obtuvieron.

$$cobertura = \frac{relevant\ documents\ \cap \{retrieved\ documents\}}{|\ relevant\ documents\ |} (2)$$

Cálculo de cobertura [10].

Ahora presentamos los resultados de cobertura para estos 3 archivos. Como se podrá notar los resultados evidentemente siguen siendo más favorables para un análisis textual que para un análisis por sinonimia.

Documento	Documentos Recuperados	Documentos Relevantes	Cobertura
Endodoncia	77	33	0.43
Sistemas heredados	11	2	0.18
Agentes Inteligentes	6	3	0.5
		Total	0.37

Table 5. Resultados de cobertura para un análisis de plagio por sinonimia.

Documento	Documentos Recuperados	Documentos Relevantes	Cobertura
Endodoncia.doc	77	40	0.52
Sistemas heredados	11	5	0.45
Agentes Inteligentes	10	6	0.6
		Total	0.52

Table 6. Resultados de cobertura para un análisis de plagio textual

Finalmente, se considera como **F-measure** a la media armónica entre la precisión y la cobertura y sirve para medir la exactitud de una prueba [10]. Por lo tanto, mediante esta medida podemos conocer el desempeño del sistema para cada uno de los 2 tipos de análisis.

$$F - Measure = 2\frac{precision*recall}{precision+recall}. \qquad (3)$$

Cálculo de F-Measure [10].

Documento	Precision * Recall	Precision + Recall	F-Measure
Endodoncia	0.43	1.43	0.6
Sistemas heredados	0.072	0.58	0.25
Agentes Inteligentes	0.25	1	0.5
		Total	0.45

Table 7. Resultados de F-measure para un análisis de plagio por sinonimia

Documento	Precisión * Recall	Precisión + Recall	F-Measure
Endodoncia	0.52	1.52	0.68
Sistemas heredados	0.45	1.45	0.62
Agentes Inteligentes	0.6	1.6	0.75
		Total	0.70

Table 8. Resultados de F-measure para un análisis de plagio textual.

Según los resultados obtenidos para F-measure comprobamos que se obtienen mejores resultados para plagio textual que para plagio por sinonimia. En el caso de plagio textual vemos que los valores de F-measure son aceptables; sin embargo, en el plagio por sinonimia vemos que nuestro segundo archivo obtuvo un puntaje

inaceptable, como ya se explicó, esto se debe a que los buscadores no siempre devuelven los resultados en un contexto adecuado para el archivo que se está analizando.

5 Conclusiones

Al finalizar el presente trabajo de investigación, hemos podido concluir que el prototipo presentado colaborará de manera eficaz a la detección de plagio. Para mayor comodidad del usuario se lo ha realizado en un entorno Web, de tal manera que el proceso que se lleva a cabo sea transparente al usuario, permitiendo obtener únicamente el informe necesario para la comprobación del mismo.

A lo largo del desarrollo del prototipo se han presentando una serie de inconvenientes, entre estos, la herramienta FreeLing que tiene escasa información con respecto al API de Java, es por ello que nos vimos en la necesidad de consultar con el desarrollador de la herramienta Lluis Padró a través del foro de la página oficial del proyecto, quien nos supo responder y guiar en este proceso.

Una de las grandes limitaciones de nuestro sistema es la dependencia que tenemos de los buscadores, ya que hemos escogido 5 buscadores que nos ayudarán a la obtención de datos, estos son: Bing, Yahoo, Yandex, Ask y Google. Es importante mencionar que fue necesario la utilización de todos estos buscadores ya que al utilizar solo un buscador este presenta complicaciones al detectar el programa y el constante envío de solicitudes, limitando la búsqueda al producir una excepción, bloqueando futuras búsquedas.

En un análisis por sinonimia resulta interesante concluir que se obtienen mejores resultados al no cambiar palabras (al azar) por sinónimos para realizar la búsqueda, sino más bien al obtener una lista de palabras importantes mediante la ley de Zipf y realizar la búsqueda con estas palabras originales sin sinónimos. Por lo tanto podemos concluir que el mejor enfoque para detectar sinonimia no necesariamente debe recurrir a la utilización de sinónimos.

También podemos concluir que por más que se elijan palabras importantes dentro de una hoja gracias a la ley de Zipf [6] los resultados que los buscadores lanzan no siempre son acordes a la temática del documento original ocasionando varios falsos positivos. También debemos resaltar el hecho de que para un análisis de sinonimia generalmente se obtienen mejores resultados al mantener las palabras importantes que Zipf devuelve sin aplicar sinónimos debido a que es poco probable que se cambien por sinónimos a palabras importantes del texto y debido a que los sinónimos pudieron haber sido cambiados en palabras menos importantes que las devueltas por Zipf, de tal forma que al enviar palabras importantes sin ningún cambio podemos dar con el documento original e indirectamente detectar plagio por sinonimia en palabras menos importantes.

Con respecto al estado del arte encontramos razones por las cuales muchos sistemas actuales presentaban resultados nulos al solicitar inclusive análisis textuales y es que se enfocaban en la rapidez para entregar resultados, sin embargo, al desarrollar nuestro sistema nos hemos percatado que realizar análisis de plagio, en especial por

sinonimia, resulta extremadamente pesado por lo tanto dichos sistemas no podrían encontrar un número significativo de resultados sin perder velocidad.

Consideramos nuestros resultados para plagio textual como aceptables en comparación con el estado del arte actual. Sin embargo, no podemos decir lo mismo para nuestros resultados sobre detección de plagio por sinonimia. Es importante mejorar dicho apartado en futuras versiones del sistema. A pesar de esto, el valor agregado que generamos al añadir la posibilidad de programar una revisión múltiple (útil para revisar trabajos de todos los estudiantes de un aula) nos permite concluir que el proyecto es útil y puede ser usado como herramienta frecuente por parte de cualquier docente.

6 Trabajo Futuro

Procedemos a sugerir una serie de mejoras a futuro para que el sistema pueda realizar un mejor trabajo en menos tiempo.

Mejorar la conversión de documentos a texto plano.

Para la transformación de documentos nos hemos visto en la necesidad de utilizar librerías externas para Java que permitan transformar estos documentos, las librerías que encontramos son "PDFBox", "Apache POI" y "docx4j" que facilitaron la conversión de los archivos PDF, DOC y DOCX respectivamente, sin embargo estos no pudieron recuperar de una manera satisfactoria el texto, ya que incorporaban espacios y saltos de carro donde no existían, tornando dificultoso el proceso de obtención de párrafos al momento de leer un archivo. Se sugiere que en futuras versiones de este prototipo se considere actualizar a librerías mucho más compatibles con los formatos a convertir.

Reconocimiento de imágenes con sus debidas referencias.

Se recomienda la posibilidad de analizar las imágenes que el documento pueda contener con la finalidad de poder identificar uno de los plagios más comunes a un nivel más avanzado.

Tratamiento de idiomas.

Se recomienda que el sistema de detección de plagio vaya más allá del idioma español, de tal forma que tenga la capacidad de abarcar diversos idiomas entre estos el inglés que es uno de los más utilizados y que facilitaría la detección cuando por lo general existe plagio de un documento en inglés traducido al idioma español, escenario en el cual nuestro prototipo no podría detectar plagio.

Mejorar la eficiencia.

El prototipo es bastante eficaz, pero su capacidad de respuesta se ve limitada por diversos factores, como son la velocidad de conexión de Internet, la temática del documento y la extensión del documento, estos factores influyen directamente sobre la

velocidad de respuesta del sistema prototipo, en una próxima versión se recomienda trabajar sobre estos puntos para lograr obtener resultados más eficientes en este aspecto. Somos conscientes de que los principales cuellos de botella que afectan directamente al desempeño de nuestro sistema prototipo son:

- Conectividad con Internet (factor externo).
- Velocidad de conversión de documentos a texto plano (factor interno).

7 Referencias:

[1] Diccionario de la Lengua Española, fecha de recuperación: 18-nov-2011, http://buscon.rae.es/draeI/

[2] PAZMIÑO YCAZA Antonio, Universidad Católica de Santiago de Guayaquil, Revista Jurídica de Propiedad Intelectual, Tomo 4, http://www.revistajuridicaonline.com/images/stories/revistas-juridicas/propiedad-intelectual-tomo-4/propiedad-intelectual-tomo4.pdf

[3] Europapress, "Por plagio la Universidad de Bayreuth retira el doctorado de Derecho al ministro de Defensa" en Europapress, Miércoles, 20 de noviembre 2011, http://www.europapress.es/internacional/noticia-universidad-bayreuth-retira-doctorado-derecho-ministro-defensa-20110223232341.html

[4] NUÑEZ Miguel Ángel, El plagio como amenaza, fecha de recuperación: 20-ene-2011, http://miguelangelnunez.suite101.net/el-plagio-como-amenaza-a8443

[5] GARCÍA G. R and RODRÍGUEZ E.G, Fraude y plagio académico en los ambientes virtuales de aprendizaje, fecha de recuperación: 28-nov-2011, http://www.distancia.unam.mx/contenido/historico/foroeducativodos/Guillermo%20Roquet%20trabajo%20escrito.pdf

[6] CSCAZORLA, La ley de Zipf: El porqué de las palabras cortas y largas, fecha de recuperación: 07-enero-2012, dirección web: http://www.xatakaciencia.com/sabias-que/el-por-que-de-las-palabras-cortas-y-largas

[7] TELLO Estefanía y ZEPEDA Beatríz, El plagio académico, fecha de recuperación: 16-oct-2011, www.flacso.org.ec/docs/plagioacademico.ppt

[8] VILLARDÓN José Luis Vicente, Análisis de coordenadas principales, fecha de recuperación; 04-nov-2011, http://biplot.usal.es/DOCTORADO/3CICLO/BIENIO-04-06/ACP/COORPRIN(apuntes).pdf

[9] Antonio Moreno Ortiz, Wordnet, fecha de recuperación: 12-enero-2012, http://elies.rediris.es/elies9/2-4-2.htm

[10] Wikipedia, Precision and Recall, fecha de recuperación: 22-feb-2012, http://en.wikipedia.org/wiki/Precision_and_recall

[11] ARAUJO Lourdes, Procesamiento de Lenguaje Natural, fecha: 01-nov-2011, http://tabasco.torreingenieria.unam.mx/gch/PLN/cap1.pdf

[12] ELIZALDE Victoria, Estudio y desarrollo de nuevos algoritmos de detección de plagio, fecha de recuperación: 28-nov-2011, http://www.dc.uba.ar/inv/tesis/licenciatura/2011/elizalde

8 ANEXOS

8.1 Elementos del sistema

El sistema posee 2 secciones o funcionalidades las cuales son:

Detección Web:

Es decir, **busca en Internet** mediante el uso de los siguientes buscadores (Google, Bing, Yahoo, Ask y Yandex) fuentes originales como Sitios Web y documentos desde los cuales se pudo haber plagiado. Se considera plagio si encuentra contenido coincidente que no contiene las debidas referencias. Este análisis puede tomar mucho tiempo dependiendo de la extensión y temática del documento a analizar.

Detección Local:

Por otro lado en este tipo de análisis el sistema **compara dos documentos locales en busca de copia entre ellos** por lo que no buscará en Internet fuentes originales, lo que a su vez tiene como efecto un análisis rápido en comparación con la funcionalidad de Detección Web.

Los siguientes elementos pertenecen a la funcionalidad de Detección Web.

Fig. 1. Principales elementos de la funcionalidad Detección Web del sistema.

1. Botón para ir hacia la página que realiza la detección de plagio buscando fuentes en internet, como muestra la Fig. 1.
2. Botón para ir hacia la página que realiza la detección de plagio localmente en el servidor sin necesidad de recurrir a fuentes externas en Internet como por ejemplo Google.
3. Botón para seleccionar el archivo a analizar. Dicho archivo puede ser un ensayo, un trabajo, una tesis o algún otro tipo de documento del cual se sospecha posee contenido plagiado. Este archivo debe poseer alguna de las siguientes extensiones:

DOC, DOCX, PDF o TXT, luego se debe procurar que no posea errores; es decir, que no hayan sido creados, modificados y guardados con editores ineficaces. Para conseguir esto procure que los archivos hayan sido creados con herramientas como Microsoft Office o Libre Office y así mismo evite que los archivos hayan sido creados o modificados con Herramientas como Abiword cuyo desempeño es limitado.

4. Botón para subir el archivo al servidor y dar inicio a la ejecución del sistema.
5. Enlace a las configuraciones avanzadas del sistema. Posteriormente se detallarán las configuraciones que se pueden modificar y que significan.
6. Opción para indicar al sistema que se desea analizar plagio textual en el archivo que se va a procesar. Toma menor tiempo que analizar plagio por sinonimia.
7. Opción para indicar al sistema que se desea analizar plagio por sinonimia en el archivo que se va a procesar. Toma más tiempo que analizar plagio textual.

Se puede observar que las opciones 6 y 7, es decir, los tipos de análisis que se van a realizar no son excluyentes, por lo tanto se puede indicar que para un archivo se realice análisis textual y también análisis por sinonimia.

Los siguientes elementos pertenecen a la funcionalidad Detección Local:

Fig. 2. Elementos de la funcionalidad Detección Local del sistema**.**

1. Es el primer archivo a ser comparado.
2. Es el segundo archivo a ser comparado.
3. El botón que sube los archivos al servidor y ejecuta el sistema para realizar la comparación entre estos 2 archivos.

Como se ha visto el sistema posee pocos elementos los cuales además son simples de entender y utilizar. Aun así, debemos explicar los dos tipos de análisis con los que se cuenta para realizar la Detección Web.

8.2 Configuraciones

Si se desea ajustar el funcionamiento de la sección “Detección Web” del sistema, se deberá hacer clic sobre configuraciones presente en dicha sección. Recuerde que no

es necesario modificarlas. A continuación se explican cada una de éstas configuraciones:

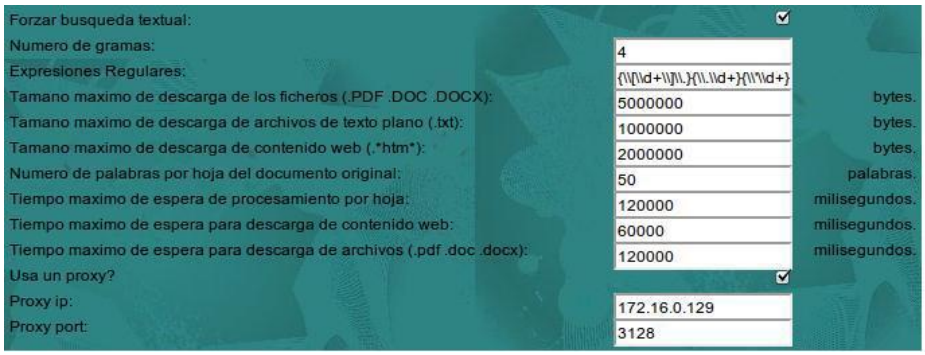

Fig. 3. Configuraciones avanzadas disponibles en la sección Detección Web del sistema.

Forzar busqueda textual: .

Cuando se realiza analisis de plagio textual muchas veces los buscadores no encuentran coincidencias debido a que el texto a buscar es demasiado largo. Es por esto que si activamos esta opcion el sistema ira eliminando cada vez una palabra al final de la busqueda y, mientras la busqueda tenga al menos 10 palabras, seguira buscando con la esperanza de encontrar alguna coincidencia. Evidentemente realizar esta tarea puede hacer que el analisis tome mas tiempo, sin embargo, permite identificar plagio con mayor grado de aciertos. Se recomienda activar esta opcion para textos que no posean muchas referencias, por otro lado se recomienda desactivar esta opcion cuando el texto a analizar posee una gran cantidad de referencias. Esta opcion viene activada por defecto. (Se utiliza en analisis textual).

Número de gramas: .

Son utilizados para saber el numero de palabras que conformaran cada grama, esto servira como punto de comparacion entre los documentos para el proceso de sinonimia, el sistema utiliza este valor para obtener el coeficiente de similitud.

Un ejemplo del uso de N-gramas es el siguiente:

Ejemplo: “Esto ejemplifica N-gramas” con N=2

El N-grama quedaría así: Esto ejemplifica | ejemplifica N-gramas.

En el sistema, el valor que se ha definido por defecto para el N-grama es de 4, pudiendo variar este valor; si es menor encontrara mas coincidencias y si aumenta el valor es posible que no se encuentren gramas iguales entre documentos (se utiliza en análisis por sinonimia).

Expresiones regulares: .
El sistema detecta si existen referencias basándose en expresiones regulares que las definan. Por defecto vienen incluidas 4 expresiones regulares, es decir, el sistema verificara los 4 tipos de referencias siguientes:

- \\[\\d+\\]\\. Valor numérico contenido entre corchetes y finalizado en punto. Por ejemplo: [2].
- \\.\\d+ Punto seguido de un valor numérico. Por ejemplo: .4
- \\ • f\\d+ Comillas seguidas por un valor numérico. Por ejemplo: • g3
- \\.\\s*\\[\\d+\\] Punto seguido de varios o ningún espacio y todo esto seguido de un valor numérico contenido entre corchetes. Por ejemplo: .[1]

Si sabe como escribir una expresión regular que defina el formato de una referencia que usted necesite puede añadir su expresión regular junto a las demás. Para hacerlo deberá encerrar su expresión entre llaves así: {\\d+\\.}.

No olvide que el sistema verifica toda la lista de expresiones regulares. Si usted desea que se busque un solo tipo de referencia deberá borrar todas las demás expresiones regulares y quedarse solo con la que le interesa. No se olvide de encerrar entre llaves aun cuando exista solo una expresión regular.

Si desea que se analice todo el texto sin importar si existen o no referencias puede borrar todo el contenido de este parámetro (se utiliza en análisis textual y análisis por sinonimia).

Tamaño máximo de la descarga de ficheros (PDF, DOC Y DOCX).
Cuando se realizan búsquedas en la web es muy probable que los resultados que devuelva la búsqueda no sean solo paginas web, pueden existir diversidad de formatos. El sistema esta enfocado en los ficheros con extensión .PDF, .DOC y .DOCX. Se ha limitado el tamaño máximo que debe poseer el archivo, esto con el objetivo de agilizar el proceso de descarga, por defecto se ha definido que el tamaño máximo del archivo sea de 5 megabytes. Si un fichero sobrepasa los 5 megabytes será ignorado y no se descargara continuando así con el proceso de Detección de Plagio. Si considera que este tipo de archivos son muy importantes puede aumentar este valor. Por otro lado, si considera que la descarga de este tipo de archivos no es relevante y desea ganar velocidad en el análisis puede reducir el valor. El valor debe ser escrito en bytes (se utiliza en análisis por sinonimia).

Tamaño máximo de archivos de texto plano.
Al igual que en el punto anterior este parámetro indica el tamaño máximo que deberá poseer un archivo de texto plano como pueden ser aquellos con formato TXT. Si el fichero posee un tamaño mayor al establecido no será descargado y se ignorara. Puede aumentar el valor de este parámetro si desea que se descarguen archivos más grandes. Este valor también se encuentra dado en bytes (se utiliza en análisis por sinonimia).

Tamaño máximo de descarga de contenido web.

Igual a los 2 puntos anteriores. Se puede especificar en bytes el tamaño máximo que debe poseer un sitio web para empezar a extraer texto desde el mismo (se utiliza en análisis por sinonimia).

Numero de palabras por página del documento original.

Al realizar análisis por sinonimia se divide el documento que se esta analizando en paginas y de estas se extraen las palabras mas importantes para ser reemplazadas con sinónimos y realizar las búsquedas. Por tanto, si se eligen pocas palabras el sistema será mas preciso en su análisis y así también demorara más en terminar de procesar el documento. Por otro lado, si se eligen muchas palabras por página el análisis será menos preciso a favor de una mayor velocidad en la ejecución del sistema. El valor mínimo de palabras por página deberá ser de 16 y el valor máximo deberá ser como mucho igual al número de palabras que posea el documento (se utiliza en análisis por sinonimia).

Tiempo máximo de espera de procesamiento por página.

Igualmente, en el análisis por sinonimia, una vez que se ha encontrado un resultado y se lo ha descargado se deberá comparar dicho documento con la pagina que genero la búsqueda inicial. Si el documento descargado es muy extenso, por ejemplo de mas de 500 hojas, puede que procesarlo tome mucho tiempo. Si desea puede reducir este tiempo máximo para agilizar la ejecución del sistema, mantener igual o aumentar el tiempo en caso de que considere que es muy importante analizar archivos extensos que el sistema descarga. Este tiempo debe darse en milisegundos (se utiliza en análisis por sinonimia).

Tiempo máximo de espera para descarga de contenido web.

Si el servidor tiene problemas de conexión o presenta una velocidad lenta de conexión a Internet es posible que al extraer texto desde sitios web, en especial de aquellos con mucho texto en su interior, el sistema se bloquee a la espera de que termine la extracción. En caso de presentarse tal escenario, este parámetro limita el tiempo que el sistema permanecerá bloqueado. En caso de sobrepasar este límite de tiempo se ignora la extracción y el análisis continúa sobre las siguientes paginas. Este valor también esta dado en milisegundos (se utiliza en análisis por sinonimia).

Tiempo máximo de espera para descarga de archivos (PDF, DOC y DOCX).

Al igual que en el parámetro anterior si el sistema debe descargar algún fichero y dicha descarga toma mucho tiempo bloqueando al sistema entonces este parámetro establece un limite de tiempo, el cual si es sobrepasado anula la descarga y procede con las siguientes paginas a analizar. Recuerde que es posible que al analizar muchas paginas el sistema proceda a realizar muchas descargas de archivos, por lo tanto, si este valor es muy alto y la conexión muy lenta el sistema puede tomar mucho tiempo en completar el análisis. Este tiempo también se encuentra dado en milisegundos (se utiliza en análisis por sinonimia).

Los siguientes 3 parámetros están relacionados con el servidor. Por tanto, si esta instalando el sistema en otro servidor o si conoce que el sistema ha cambiado de servidor

estos parámetros puede ser útiles (se utilizan en análisis textual y análisis por sinonimia).

Usa un proxy?.
Si el nuevo servidor en donde se aloja el sistema atraviesa un proxy para salir a internet debe activar esta casilla.

Proxy IP .
En caso de que el servidor atraviese un proxy, se deberá indicar la dirección IP de dicho proxy.

Proxy port.
Al igual que en el parámetro anterior si el servidor atraviesa un proxy, mediante este parámetro se

8.3 Coeficiente de similitud

Se analizaron varios coeficientes para medir similitud entre estos tenemos: Jaccard, Overlap, Dice, Rogger y Tanimoto, sin embargo, el que se aplica más a nuestras necesidades es el Coeficiente de Overlap

Coeficiente de Overlap

Esta es una medida que se encuentra muy relacionada con el índice de Jaccard, lo que hace es calcular la similitud en base a los conjuntos A y B, para ello realiza una operación de intersección de los conjuntos y los divide para el valor mínimo de los conjuntos, es decir, escoge el conjunto menor para realizar la operación. En otras palabras lo que trata de hacer el coeficiente de Overlap es demostrar que tan contenido esta un conjunto dentro del otro.

$$O = \frac{a}{\min(|b|, |c|)} \quad (4)$$

Coeficiente de Overlap [12].

Dónde [8]:
a: Número de caracteres presentes en los dos individuos,
b: Número de caracteres presentes en i y ausentes k.