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:

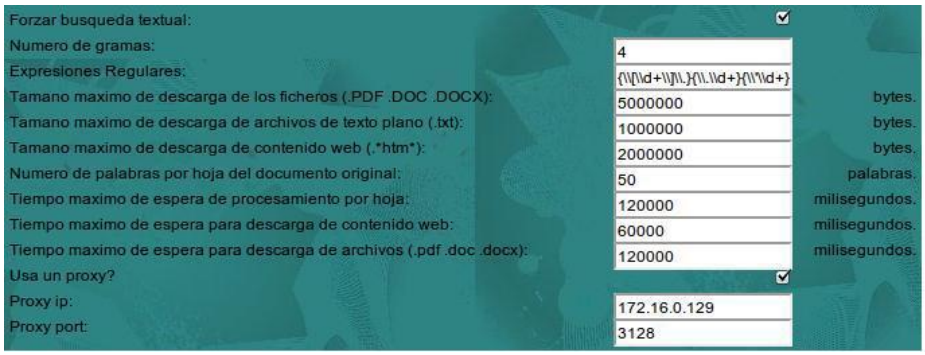

Fig. 3. Configuraciones avanzadas disponibles en la sección Detección Web del sistema.

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.