IMPLEMENTACIÓN DEL ALGORITMO RABIN-KARP PARA DETECCIÓN DE PLAGIO EN DOCUMENTOS DE TEXTO

DOCENTE: BORIS CHULLO LLAVE

ALUMNOS:

- HUANCARA CCOLQQUE ALEX HELDER QUISPE YAHUIRA RONALDO SARCON JACINTO DANIEL EDUARDO

- VEGA CENTENO OLIVERA RONALDINHO

FUNCIONAMIENTO DEL RABIN-KARP

El hash es el valor más importante del algoritmo de Rabin-Karp. El resultado del hash letras de k-gramo con un cierto número de bases se obtiene multiplicando el valor ASCII por números predeterminados donde la base es prima. El método Rabin-Karp tiene provisiones si dos cadenas son lo mismo, entonces el valor hash debe ser el mismo. Aquí hay un ejemplo de cálculo en Rabin-Karp algoritmo. Suponga que el texto es MEDAN.

K-gramo = 5

Base = 7 (Se puede usar cualquier número primo)

texto = "MEDAN"

• Primero hallamos el valor ASCII de cada carácter que se encuentra en el texto texto(1) = 77

$$texto(2) = 69$$

texto(3) = 68

$$texto(4) = 65$$

texto(5) = 78

Hallamos el hash del texto

$$Hash = (77*7^4 + 69*7^3 + 68*7^2 + 65*7^1 + 78*7^0)$$

Hash = 235599

El resultado del cálculo hash es 235599. Esta acción se realiza hasta que se cumplen todas las palabras de la lista.

Las siguientes tablas 1 y 2 son ejemplos de comparación de documentos después de que los valores hash son adquirido. El valor hash de la primera tabla se calculará mediante el valor hash de la segunda tabla.

Tabla 1. Valor hash del documento uno

19875 16830 23124 17433 20546 21489 26753 13498 23846 16528 21848 28447 29994 10301 13009 18832 27217 23157 25854 22492 14952 14337 29348 19978 28809 13485 14188 13131 21215 12053 25669 13809 26508 19455 25356 29964 17723 26633 17445 11803 19477 27142 24814 15155 26266 28432 19007 21896 16625 20681

Tabla 2. Valor hash del documento dos

28432 26406 28424 13930 19187 18049 10867 18516 26753 19975 10152 13053 24120 21896 18351 12605 25101 21215 20750 15513 22949 26006 25045 25932 10695 13254 21504 20286 22492 10615 25565 29941 17403 23018 22666 19744 19769 19877 29535 13139 25669 16830 14297 20916 24640 16960 20681 13131 13009 18947 Hay diez piezas del mismo hash que tienen ambas tablas. Después de calcular el valor hash similar, el siguiente paso es calcular el porcentaje de similitud de los dos documentos. La fórmula utilizada es la sigue:

$$P = (2*SH)/(THA + THB)*100\%$$

Dónde:

P = Tasa de plagio

SH = Hash idéntico

THA = Hash total en el documento A

THB = Hash total en el documento B

En el cálculo anterior hay diez valores que tienen un valor similar. Entonces el nivel de plagio el cálculo es el siguiente.

$$P = (2*10)/(50 + 50)*100\%$$

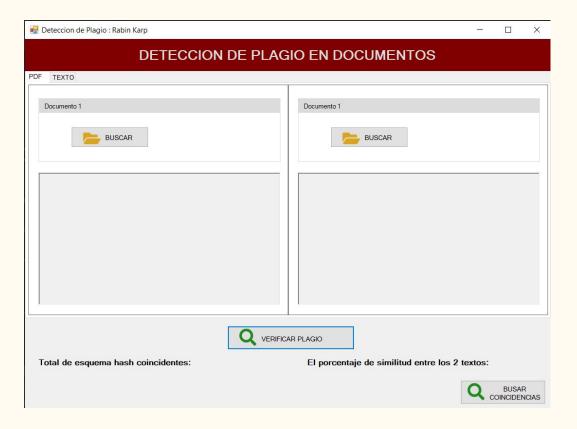
$$P = (20/100)*100\%$$

$$P = 0.5*100\%$$

$$P = 20\%$$

El porcentaje de plagio que tienen ambos documentos es del 20%

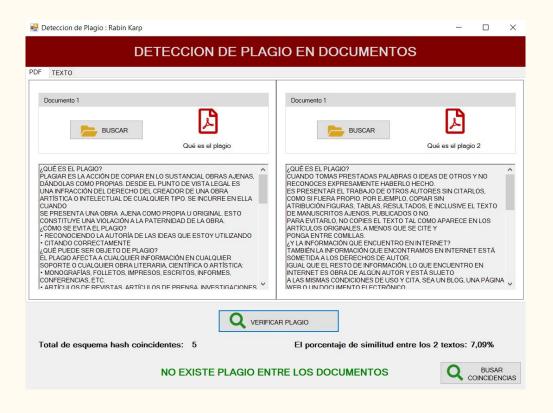
INTERFAZ DEL APLICATIVO



Primeramente subimos los archivos .pdf dando click en el botón Buscar Deteccion de Plagio: Rabin Karp



Posteriormente damos click en el botón VERIFICAR PLAGIO



- Nos bota las coincidencias entre ambos documentos
- El porcentaje de plagio
- así, como también hay un boton que nos indica las cadenas coincidentes