A01369421 Óscar Emilio Reyes Taboada
A01366686 José Israel Quintero Alfaro

28 de abril del 2023

Similitud en textos

Profesor:

Victor Manuel de la Cueva Hernandez

Introducción:

En esta investigación, se aborda la problemática de la detección de plagio en código por medio de tokens, la cual es relevante en el ámbito de la programación. El objetivo principal es evaluar la eficacia de dos métodos de análisis de similitud: TF y TF-IDF.

Hipótesis:

Se plantea la hipótesis de que el método TF- IDF es más efectivo para detectar plagio en código mediante tokens que el método TF debido a que tiene en cuenta la frecuencia de un término en el corpus completo, no sólo en un documento individual. Por lo tanto, los términos menos comunes tendrán más peso en la medición del nivel de similitud entre los documentos, lo que puede ser especialmente útil en la detección de plagio. Además, el método TF-IDF reduce el peso de las palabras que aparecen con mucha frecuencia en todos los documentos, lo que permite centrarse en los términos más importantes.

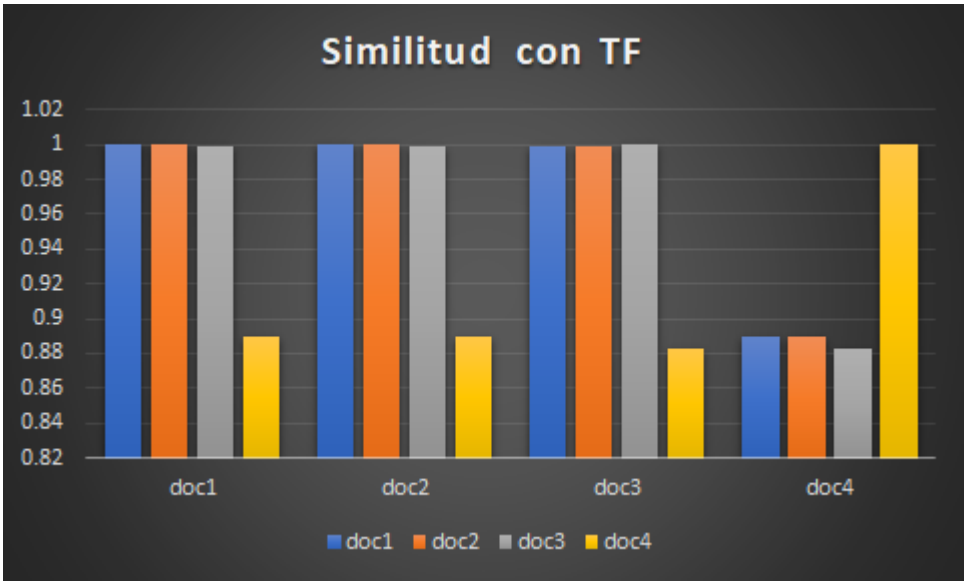
Metodología:

Para evaluar la hipótesis, se realizó un experimento en el cual se compararon los resultados de la matriz de similitud obtenida por los métodos TF y TF-IDF. Se utilizaron cuatro documentos de código de programación para el análisis, los cuales se compararon dos a dos. La matriz de similitud se obtuvo mediante el cálculo de la distancia coseno entre los documentos.

Resultados:

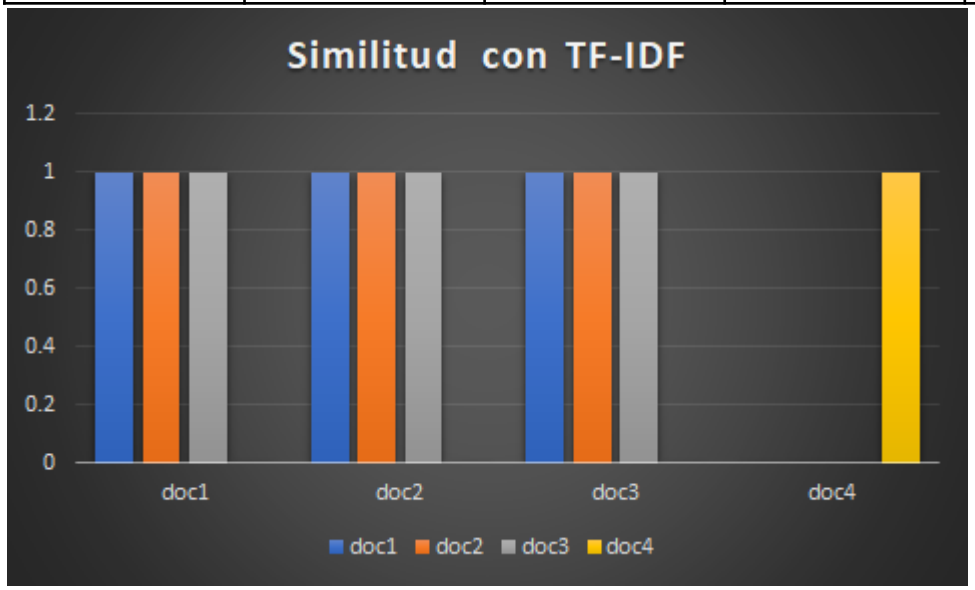
TF

	doc1	doc2	doc3	doc4
doc1	1	1	0.999511	0.88919
doc2	1	1	0.999511	0.88919
doc3	0.999511	0.999511	1	0.882836
doc4	0.88919	0.88919	0.882836	1



TF-IDF

	doc1	doc2	doc3	doc4
doc1	1	1	0.999268	0
doc2	1	1	0.999268	0
doc3	0.999268	0.999268	1	0
doc4	0	0	0	1



Los resultados obtenidos mostraron que la matriz de similitud producida por el método TF fue más alta que la obtenida por el método TF-IDF. Esto se debe a que TF no considera la importancia de cada término en el documento y simplemente cuenta la frecuencia de cada término en cada documento. Por otro lado, TF-IDF tiene en cuenta la importancia de cada término en un documento en relación con la frecuencia del término en todos los documentos. Esto ayuda a reducir la importancia de los términos comunes que aparecen en

muchos documentos y a aumentar la importancia de los términos raros y distintivos que son específicos de un documento en particular.

También como podemos darnos cuenta en el análisis TF-IDF, los primeros 3 archivos no tienen correlación alguna con el último ya que debido a la ponderación de los tokens, el analizador se percata de que la cantidad de veces que se usan los tokens son muy diferentes, y de hecho hay algunos que no aparecen en el segundo código, por esto mismo, el análisis resulta en 0, ya que podemos concluir que como los documentos 1, 2 son iguales (solo con un cambio de nombre de alquinos IDs) y el 3 solo es una variación del 1, el 4 aun cuando ocupa tokens similares, recibe una similitud igual a 0, por la inexistencia de ciertos tokens que sí aparecen en el resto de los documentos, haciendo que su inversa se dispare por la falta de estos términos.

11	tokentypeis	0.004662004662004662	0.004662004662004662	0.004761904761904762	0
12	tokentypele	0.002331002331002331	0.002331002331002331	0.002380952380952381	0
13	tokentypege	0.002331002331002331	0.002331002331002331	0.002380952380952381	0
14	tokentypene	0.002331002331002331	0.002331002331002331	0.002380952380952381	0
15	tokentypeplus	0.02097902097902098	0.02097902097902098	0.023809523809523808	0
16	tokentyperkey	0.03263403263403263	0.03263403263403263	0.03333333333333333	0.0064516129032258064

29	none	0	0	0	0.0064516129032258064

Como se puede observar en la imagen superior donde cada columna es un documento, podemos ver como para TF-IDF consigue un valor de 0 en múltiples palabras del vocabulario que está presente en el resto de los archivos, esto ocurre en múltiples tokens, adicionalmente cuenta con un token “none” el cual es único para el documento 4.

Discusión:

La herramienta TF puede ser útil si se desea encontrar similitudes entre documentos que contienen un gran número de términos distintos, mientras que TF-IDF puede ser más adecuado si se buscan identificar términos distintivos. Por lo tanto, es importante evaluar cuidadosamente la elección de la herramienta de similitud más adecuada para cada situación específica.

Conclusiones:

Como podemos ver en nuestro caso TF nos entregó resultados de similitud más altos para todos los documentos, diciéndonos que el archivo 4 se parece al resto, mientras que el TF-IDF dijo que este no se le acerca en lo absoluto, con esto podemos concluir que para nuestro caso, TF-IDF fue mejor ya que logró resaltar las diferencias entre nuestros casos de prueba, pero no creemos que estos resultados sean satisfactorios, ya que necesitamos una

muestra de documentos mas grande y con mayor diversidad para poder concretar los resultados de una manera más precisa.

Como equipo no creemos que este sea el mejor método de identificación ya que estas 2 formas son ocupadas mayormente para texto, mientras que nosotros las ocupamos para tokens, los cuales son mucho más consistentes y propensos a repetirse, ya que lo que cambia en un código es la estructura lógica del programa, no los tokens. Esto provoca que la similitud (como podemos observar en TF) sea excesivamente alta, o por otro lado que los resultados sean prácticamente binarios.