Reporte de SI

Juan David Menéndez del Cueto[1] and Karl Lewis Sosa Justiz[1]

Universidad de la Habana, Facultadad de Matemática y Computación

Abstract. El siguiente trabajo consiste en el diseño, implementación, evaluación y análisis de un Sistema de Recuperación de Información. Desde la consulta hecha por el usuario, la representación de los documentos y consultas, el funcionamiento del motor de búsqueda y la obtención de resultados. El modelo escogido fue el Vectorial.

1 Colecciones

Se escogieron 2 coleccciones de prueba: Cranfield y CISI, el principal motivo fue que se pudieron conseguir en json, un formato habitual y que es fácilmente trabajable. En el caso de Cranfield, donde las evaluciones estan puntuadas, solo se consideró su relevancia de manera binaria, o sea es relevante o no, y se desechó la escala.

2 Interfaz para usuarios

Es una interfaz por consola que guía al usuario, dando opciones como son escoger la colección para realizar la query, escoger si se tienen en cuenta solo sustantivos y verbos en el procesamiento de los textos, realizar una evalución del sistema o finalmente solo realizar una consulta.

3 Modelo

El modelo implementado es uno de los 3 clásicos, el modelo vectorial, porque a pesar de ningún modelo ser superior a otro, este modelo tiene peso no binario, es decir tiene en cuenta la frecuencia de los términos, siguiendo el esquema tf-idf , y esta característica terminó por mover la balanza en sentido del modelo vectorial, además de otras ventajas, como ser de uso sencillo por un usuario inexperto y el uso del coseno para ordenar lo elementos de acuerdo al grado de similitud, que también resultaron atractivas.

4 Procesamiento de Texto

Al procesar las colecciones de prueba, específicamente los documentos, se utilizaron varias técnicas, con el objetivo de capturar mejor la esencia de los documentos, primeramente todas las palabras fueron llevadas a minúsculas, ya que para el modelo carece de importancia si una palabra cominenza en mayúscula o no, se eliminaron tanto urls, como signos de puntuación que son elementos que no aportan nada a la semántica, ya que este modelo asume que los elementos son mutuamente independientes, siendo esta su principal debilidad.

Se utilizó una herramienta externa, el módulo nltk[1] para 3 procesos en concreto, eliminar las llamadas **stopwords** que son palabras muy comunes como

nexos gramaticales o artículos que no aportan realmente a la semántica y son causantes de ruido en el sistema comprometiendo su efectividad. Además se utilizó también para hacer **lemmatizing** que no es más que llevar las palabras de formas flexionadas a el lemma que no es mas que la palabra que encontraríamos en un diccionario.También se da la opción de solo analizar sustantivos y verbos, esta característica se da como opción ya que no siempre trae consigo mejores resultados.

Las consultas realizadas por el usuario fueron procesadas de manera similar, es decir fue llevada a minúscula , se eliminaros urls, signos de puntuación además se da la opción de expandir la consulta o no usando un diccionario de sinónimos proporcionado por el módulo nltk, se da la opción de escoger usarlo o no ya que en determinadas querys puede ser contraproducente su uso, más cuando se tienen documentos centrados en un tema en particular como es el caso en ambas collecciones

5 Retroalimentación

La idea de la retroalimentación en el modelo vectorial consiste en encontrar un vector consulta que maximice la similitud con los elementos relevantes mientras minimice la similitud con los documentos no relevantes. El algoritmo utilizado para cumplir este objetivo es el algoritmo de Rocchio, dado que su implementación es sencilla y que considera la influencia de la consulta original en la formación de la nueva consulta. Los pesos usados para cada término de la nueva consulta fueron $\alpha = 1$, $\beta = 0.75$ y $\gamma = 0.15$.

6 Evaluación

Medidas Objetivas: Entre las medidas objetivas se tomaron la precisión, recobrado, f1 y la proporción de fallo, siendo estos los resultados en ambas colecciones de datos. El modelo vectorial no da una una lista de relevancia, sino que los ordena a todos, así tomamos los primeros 10 de mayor coincidencia.

Table 1. Usando todas las palabras

Metric	*CRAN*	*CISI*
10-Recobrado	0.4372	0.1288
10-F1	0.3540	0.17118
10-Precisión	0.3062	0.3053
10-Fallout	0.0050	0.0049

Table 2. Usando sustantivos

Metric	*CRAN*	*CISI*
10-Recobrado	0.3638	0.0981
10-F1	0.3090	0.1616
10-Presición	0.2484	0.2789
10-Fallout	0.0054	0.0051

Podemos ver que si bien no son resultados especialmente buenos, tampoco son del todo malos, podría decirse que están en la media teniendo en cuenta que es un modelo clásico.

Medidas Subjetivas Independientemente de los resultados de las medidas objetivas, la satisfacción del usuario es al final la mejor medida de evaluación. En el esfuerzo del usuario, si bien la realización de consultas no es complicada, sigue siendo poco atractiva visualmente y es un tanto complicada la búsqueda de resultados. El tiempo de respuesta es bastante bueno pero al final no son colecciones tan grandes así que era esperado.

7 Posibles mejoras

Entre las posibles mejoras en primera instancia está la creación de una mejor interfaz de usuario, podría ser un bot en telegram una aplicación de escritorio o bien una página web.
En lo que respecta al modelo la principal limitante es que asume que los términos son mutuamente independientes, esto podría ser resuelto aplicando redes neuronales o alguna herramienta de IA para tener en cuenta estas relaciones.

References

1. https://www.nltk.org/