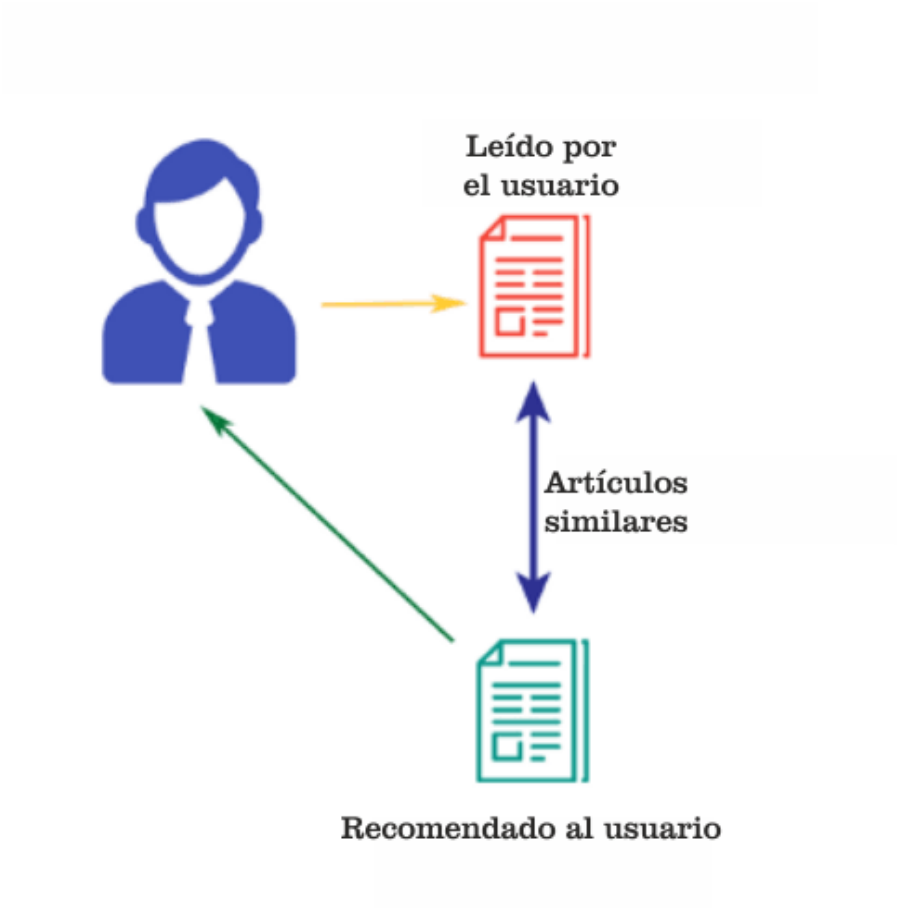
Escuela Superior
de Ingeniería y Tecnología
Universidad de La Laguna

Gestión del Conocimiento en las Organizaciones

Sistema de recomendación - Modelos Basados en el Contenido



Dayana Armas Alonso
(alu0101228020@ull.edu.es)

Índice

1. Introducción

En este informe se realizará una descripción del análisis que se ha llevado a cabo en varios ejemplos y las conclusiones que se han obtenido.

2. Análisis

A continuación, se realizará un análisis del documento 3 que se encuentra en el repositorio que contiene ejemplos de pruebas.

Cabe destacar que debido al gran tamaño de las tablas, se produce el scroll y por lo tanto, no se puede mostrar por completo el contenido mediante capturas de pantalla dado que no se entendería los datos que se muestran. Es por ello que si desea ver este ejemplo deberá acceder a la página web donde se ejecuta el sistema de recomendación que se ha llevado a cabo y una vez en ella, deberá introducir el documento 3 de prueba que estamos analizando para poder mostrar los resultados.

2.1 Tabla de Similitud Coseno

Comenzamos realizando el análisis de la tabla de Similitud de Coseno.

Similitud Coseno

	Documento 1	Documento 2	Documento 3	Documento 4	Documento 5	Documento 6	Documento 7	Documento 8	Documento 9	Documento 10	Documento 11	Docun 12
Documento 1	1	0.224	0.038	0.048	0.156	0.115	0.052	0.144	0.147	0.234	0.051	0.177
Documento 2	0.224	1	0.36	0.316	0.303	0.465	0.516	0.43	0.409	0.541	0.183	0.29
Documento 3	0.038	0.36	1	0.319	0.259	0.435	0.49	0.348	0.414	0.353	0.339	0.32
Documento 4	0.048	0.316	0.319	1	0.331	0.178	0.323	0.407	0.34	0.248	0.196	0.238
Documento 5	0.156	0.303	0.259	0.331	1	0.289	0.278	0.301	0.306	0.347	0.287	0.359
Documento 6	0.115	0.465	0.435	0.178	0.289	1	0.605	0.311	0.34	0.608	0.189	0.449

La Similaridad Coseno consiste en encontrar pares de documentos que sean similares semánticamente y definan el mismo concepto. Encontrar similitudes entre documentos se utiliza en varios dominios, como recomendar libros y artículos similares, identificar documentos plagiados, documentos legales, etc.

Para llevar a cabo el cálculo de la similitud entre par de documentos, se ha realizado anteriormente la frecuencia normalizada, es decir, que se han normalizado los valores TF con el objetivo de evitar una predisposición hacia los documentos largos.

En la tabla que se muestra, se puede observar que tanto las filas como las columnas están compuestas por los documentos que se están analizando, en este caso, va del documento 1 al documento 20.

Por otro lado, podemos apreciar que los valores oscilan entre 0 y 1. Por lo que podemos llegar a la conclusión de que aquellos valores más cercanos a 0 son los pares de documentos que tienen menos similitud y por lo tanto, menor relación entre ellos mientras que aquellos valores más cercanos a 1 son los que tienen mayor relación entre ellos.

Esto se puede demostrar dado que la diagonal que marca desde el documento 1 hasta el documento 10 muestra el valor 1. Esto es debido a que si realizamos la similitud de un documento consigo mismo, al tratar del mismo documento su relación es perfecta y por lo tanto, muestra el valor más alto que es 1.

Ahora podemos fijarnos en aquellos pares de documentos que no se analicen consigo mismo pero que tengan mayor similitud como por ejemplo la similitud entre el **documento 6** y **10** cuyo valor es **0.608**. Este par de documentos es el valor de similitud más alto entre distintos documentos. El contenido de estos documentos es el siguiente:

- **Documento 6**

*As with many of the Erath 2010 vineyard designates, this is strongly herbal. The notes of leaf and herb create somewhat unripe **flavor** impressions, with a touch of bitterness on the **finish**. The **fruit** just passes the ripeness of sweet tomatoes.*

- **Documento 10**

*Part of the natural wine movement, this wine is made from organic grapes, and the label is printed with vegetable ink on recycled paper. The quality of **fruit** is very nice, with a juicy palate and a bright berry **flavor** on the **finish**.*

Si miramos el contenido de estos documentos, podemos ver que ambos tienen relación dado que el primero habla sobre los viñedos designados en Erath 2010 y los sabores y el segundo trata sobre los sabores de la fruta de los vinos.

Para saber la similitud entre estos documentos, se tiene en cuenta las palabras, es decir, conceptos que aparecen en ambos. Para ello, se observa que se han señalado las palabras que se repiten en ambos documentos con dos colores.

Por un lado están las palabras de color violeta que hace referencia a aquella gran cantidad de preposiciones, conjunciones y pronombres que se repiten pero que no son términos que ofrecen algún tipo de significado y por lo tanto, no son relevantes.

Mientras que por otro lado, podemos observar aquellas palabras de color **azul** (**flavor, finish y fruit**) que son términos que tienen un significado dado un contexto y son los más importantes.

Por lo tanto, si traducimos estos términos al español, se puede apreciar que estos conceptos tratan de lo que hablamos anteriormente sobre los sabores de las frutas de los vinos.

Esto se demuestra dado que en ambos textos, la palabra **finish** viene precedida por **on the finish** y por lo tanto, se observa como la relación entre **flavor** y **on the finish** en ambos documentos trata sobre el sabor de las degustaciones de los vinos. Además, al relacionar la palabra **fruit** podemos observar que hacen referencia a los sabores de las frutas de los vinos.

Finalmente, podemos llegar a la conclusión que a través de aquellos términos que aparecen en ambos textos, hemos podido identificar la temática de estos documentos y además, hemos podido ver su relación. Por lo tanto, podemos decir que si un usuario está interesado en uno de estos documentos, posiblemente pueda ser de interés el otro documento dado que ambos tienen una estrecha relación.

2.2 Tabla de documento

A continuación, entre los documentos anteriormente vistos, analizaremos el documento 10 dado que tiene resultados más interesantes que nos pueden servir para realizar un análisis más completo.

Documento 10

Índice	Término	TF	IDF	TFIDF
1	part	1	1	1
2	of	2	0.097	0.194
3	the	4	0.222	0.887
4	natural	1	1.301	1.301
5	wine	2	0.26	0.519
6	movement	1	1.301	1.301
7	this	1	0.046	0.046
8	is	3	0.222	0.666
9	made	1	1	1
10	from	1	1	1
11	organic	1	1.301	1.301
12	grapes	1	1.301	1.301
13	and	2	0.022	0.045
14	label	1	1.301	1.301
15	printed	1	1.301	1.301
16	with	2	0.155	0.31
17	vegetable	1	1.301	1.301
18	ink	1	1.301	1.301
19	on	2	0.699	1.398

20	recycled	1	1.301	1.301
21	paper	1	1.301	1.301
22	quality	1	1	1
23	fruit	1	0.301	0.301
24	very	1	1.301	1.301
25	nice	1	1.301	1.301
26	a	2	0.155	0.31
27	juicy	1	0.824	0.824
28	palate	1	0.602	0.602
29	bright	1	0.824	0.824
30	berry	1	0.699	0.699
31	flavor	1	0.602	0.602
32	finish	1	0.602	0.602

Las tablas de los documentos tienen como columnas: índice, término, valor TF, valor IDF y valor TF-IDF. En los términos, hay que destacar que las palabras repetidas en el mismo documento, solo se muestran una vez. Ahora bien, llevaremos a cabo el análisis de los valores calculables de los términos.

Valor TF (Term frequency)

En primer lugar, tenemos el valor TF que hace referencia a la frecuencia con la que aparece un término en el documento.

Las palabras que aparecen más de una vez, es decir, con una frecuencia mayor que 1, son:

- **of:** con un valor de frecuencia igual a 2.
- **the:** con un valor de frecuencia igual a 4
- **wine:** con un valor de frecuencia igual a 2.
- **is:** con un valor de frecuencia igual a 3.

- **and:** con un valor de frecuencia igual a 2.
- **with:** con un valor de frecuencia igual a 2.
- **on:** con un valor de frecuencia igual a 2.
- **a:** con un valor de frecuencia igual a 2.

De todas estas palabras que tienen un índice de frecuencia mayor que 1, podemos destacar **wine** dado que tiene un significado y trata de un concepto mientras que las demás son preposiciones, conjunciones y pronombres que permiten unir las palabras, por lo tanto, no son relevantes.

Valor IDF (Inverse document frequency)

A continuación, tenemos el valor IDF que mide el significado de un término no por su frecuencia en un documento determinado, sino por su distribución en los demás.

La fórmula del IDF viene dada por lo siguiente:

$$\mathbf{IDF(x) = log10\ (N/dfx)}$$

Donde **N** significa la cantidad de documentos que hay y **dfx** en cuantos documentos aparece la palabra **x**.

De esta forma, podemos decir que si un término aparece en todos los documentos del corpus, IDF es igual a 0. Mientras en menos documentos aparezca el término, mayor será el valor de IDF.

Por lo tanto, entre los valores que se observan en la tabla, destacaremos aquellos valores con menor valor IDF dado que hace referencia a aquellos valores que se repiten en más documentos, que son:

- **of:** con un valor igual a 0.097.
- **this:** con un valor igual a 0.046.
- **and:** con un valor igual a 0.022.
- **with:** con un valor igual a 0.155.
- **a:** con un valor igual a 0.155.

Podemos observar como las palabras que aparecen en más documentos son preposiciones o conjunciones dado que estás nos permiten unir frases pero no tienen ningún significado relevante.

Por otro lado, también podemos observar que aquellas palabras con significado que tienen menor índice de valor IDF son:

- **wine:** con un valor igual a 0.26.
- **fruit:** con un valor igual a 0.301.
- **palate:** con un valor igual a 0.602.
- **flavor:** con un valor igual a 0.602.
- **finish:** con un valor igual a 0.602.

Esto quiere decir que estas palabras de este documento tienen alguna relación con otros documentos que engloban dentro del índice del valor IDF.

Valor TF-IDF

El valor TF-IDF trata de la frecuencia de ocurrencia de un término en la colección de documentos, esta es una medida numérica que expresa cuán relevante es una palabra para un documento en una colección.

De esta manera, nos permite saber con qué frecuencia aparece un término en un documento y permite compararlo con el número de documentos en los que aparece ese término.

La fórmula que permite hallar este valor es:

TF-IDF(x) = TF(x) * IDF(x)

Donde se multiplican los valores TF e IDF, siendo **TF(x)** la frecuencia de la palabra clave x en el documento actual, e **IDF(x)** el valor IDF para la palabra x.

De esta manera, al ser una medida que nos indica la importancia de un término en un documento, permite dar un mayor peso a los términos que son menos comunes en el corpus, y así, reduce la importancia de términos muy frecuentes.

Por lo tanto, los términos con mayor importancia son:

- **natural:** con un valor igual a 1.301.
- **movement:** con un valor igual a 1.301.
- **organic:** con un valor igual a 1.301.
- **grapes:** con un valor igual a 1.301.
- **label:** con un valor igual a 1.301.
- **printed:** con un valor igual a 1.301.
- **vegetable:** con un valor igual a 1.301.
- **ink:** con un valor igual a 1.301.
- **on:** con un valor igual a 1.398.
- **recycled:** con un valor igual a 1.301.
- **paper:** con un valor igual a 1.301.
- **very:** con un valor igual a 1.301.

- **nice:** con un valor igual a 1.301.

3. Conclusiones

En conclusión, esta práctica nos ha servido para aprender sobre los sistemas de recomendación sobre modelos basados en el contenido. Este tipo de sistemas es uno de los que tiene mayor presencia en la actualidad.

Esto es porque al tener en cuenta algunos datos del historial de un usuario se intenta predecir que busca el usuario y que sugerencias similares puede mostrar. Con estos sistemas, podemos descubrir opciones que se ajusten a las características de los productos o contenidos que hemos disfrutado con anterioridad y elegir elementos similares nuevos.