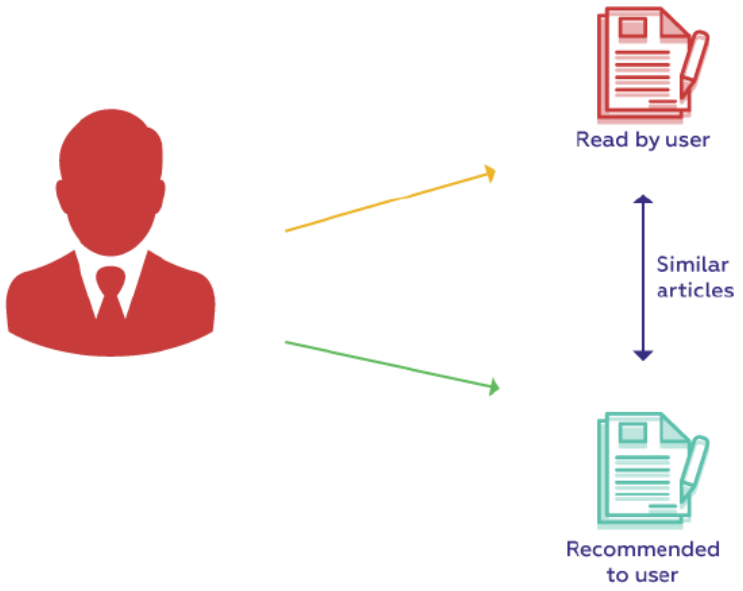
Escuela Superior
de Ingeniería y Tecnología
Universidad de La Laguna

Gestión del conocimiento en las organizaciones

Sistema de recomendación

Modelos Basados en el Contenido



Alberto Mendoza Rodríguez
(alu0101217741@ull.edu.es)

Índice:

1. Introducción

Esta práctica tiene como objetivo implementar un sistema de recomendación siguiendo el modelo basado en el contenido.

El código fuente desarrollado para el sistema de recomendación se encuentra en el siguiente repositorio de GitHub:

https://github.com/alu0101217741/Modelos_Basados_en_el_Contenido_GCO

En este informe se muestran las conclusiones extraídas tras analizar, mediante el sistema desarrollado, el archivo de texto documents-03.txt que se puede encontrar en el repositorio de ejemplos, proporcionado para comprobar el correcto funcionamiento del programa.

Este archivo de texto tiene un total de 20 líneas, cada una de ellas representa un documento que contiene una descripción textual de las características básicas de un determinado producto.

2. Análisis

En este apartado se muestran las conclusiones que se han extraído a partir de los resultados obtenidos de analizar el archivo de texto documents-03.txt con el sistema desarrollado.

2.1. Similaridad coseno entre documentos

La similaridad coseno se basa 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, etc.

Cabe destacar que para calcular la similaridad coseno se ha empleado la frecuencia normalizada (valores TF normalizados) con el objetivo de evitar una predisposición hacia los documentos largos.

A continuación, se muestra la tabla que se ha obtenido para la similaridad coseno entre cada par de documentos.

Cabe destacar que al desarrollar el sistema de recomendación se consideró la visualización de tablas de gran tamaño, para ello se incluyó la opción de **scroll** (desplazarse en las dos direcciones de la tabla). Debido a esto, en la captura que se muestra a continuación se observa únicamente la parte de la tabla que es visible sin desplazarse, si se desea observar todo el resultado acceda al sistema de recomendación e introduzca el archivo de texto.

Similaridad coseno entre documentos

	Documento 1	Documento 2	Documento 3	Documento 4	Documento 5	Documento 6	Documento 7	Documento 8	Documento 9	Documento 10	Doc 11
Documento 1	1	0.224	0.038	0.048	0.156	0.115	0.052	0.144	0.147	0.234	0.05
Documento 2	0.224	1	0.36	0.316	0.303	0.465	0.516	0.43	0.409	0.541	0.18
Documento 3	0.038	0.36	1	0.319	0.259	0.435	0.49	0.348	0.414	0.353	0.33
Documento 4	0.048	0.316	0.319	1	0.331	0.178	0.323	0.407	0.34	0.248	0.19
Documento 5	0.156	0.303	0.259	0.331	1	0.289	0.278	0.301	0.306	0.347	0.28
Documento 6	0.115	0.465	0.435	0.178	0.289	1	0.605	0.311	0.34	0.608	0.18

Como se puede observar los resultados obtenidos varían en el rango de 0 a 1, esto indica que los valores más cercanos a 0 se tratan de documentos que tienen una menor relación entre ellos, mientras que los valores próximos a 1 reflejan pares de documentos que son más similares.

Lo explicado anteriormente se refleja en la tabla cuando se lleva a cabo la similaridad entre el mismo par de documentos, esto hace que el resultado sea 1 debido a que son exactamente iguales, lo que provoca que los valores de la diagonal siempre sean 1.

Sin considerar los valores iguales a 1 obtenidos al analizar el mismo par de documentos, el mayor valor de similaridad que se ha obtenido es **0.608** que se produce entre el par de documentos 6 y 10. El contenido de estos se muestra a continuación:

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**.

Estos textos han tenido el mayor valor de similitud ya que ambos describen las características de un tipo de vino. Además, estas características son parecidas porque tanto en el documento 6 como el 10 se explica el sabor que tienen estos vinos, cómo se ha tratado la fruta empleada para elaborarlos y el gusto que dejan finalmente tras probarlos.

Como se puede observar se han destacado las palabras que más coinciden en ambos documentos. Sin embargo, no todos ellas son relevantes para llevar a cabo un análisis de similaridad, ya que los artículos y preposiciones (color azul) son términos muy comunes por lo que no aportan una información relevante.

Por ello, las palabras realmente importantes son las que se han destacado en color verde (**flavor, finish** y **fruit**), debido a que son aquellas que permiten entender realmente la razón de que exista una similaridad tan alta entre ambos documentos.

Analizando estas palabras se observa que se trata el tema del sabor y que este viene provocado por las frutas. En cuanto al término *finish* destaca que en ambos documentos viene precedido de las mismas palabras (**on the finish**). Relacionando esta frase con los otros dos términos podemos concluir que el producto del que tratan deja un sabor con gusto a fruta cuando se termina de probar.

Este análisis coincide con la descripción del contenido de los documentos que se hizo previamente. Por tanto, gracias a la similaridad coseno y analizar las palabras comunes se ha podido determinar la temática que tratan, de esta forma es posible que si un usuario se interesa por uno de los dos documentos recomendarle el otro ya que muy probablemente sea también de su interés.

2.2. Documentos

Debido a que anteriormente se estudió en profundidad la similaridad entre los documentos 6 y 10, en este apartado se analizará el documento 10 debido a que tiene mejores resultados para llevar a cabo el análisis.

Tabla para el documento 10:

Documento 10

Índice	Término	TF	IDF	TF-IDF
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

2.2.1. TF

La columna de valores TF indica el número de veces que cada uno de los términos aparece en el documento 10. Al observar los valores obtenidos podemos ver que muchas de las palabras solamente aparecen una vez, sin embargo hay otras que se repiten. A continuación, se muestran estos términos junto con la cantidad de veces que aparecen en el documento:

- **of:** 2
- **the:** 4
- **wine:** 2
- **is:** 3
- **and:** 2
- **with:** 2
- **on:** 2
- **a:** 2

De la lista anterior se puede concluir que como es habitual las palabras que más se repiten son preposiciones y artículos. Sin embargo, destaca el término **wine** que tiene 2 apariciones, gracias a este se puede deducir la temática del documento.

2.2.2. IDF

La frecuencia inversa del documento (IDF) es una medida que permite determinar si un término es común o no dentro del conjunto de documentos que se analizan. Se calcula con la siguiente fórmula:

$$IDF(x) = \log N/dfx$$

Donde **N** es el número de documentos que pueden ser recomendados, y **dfx** la cantidad de documentos en los que aparece la palabra **x**.

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

Ahora si analizamos los resultados para el documento 10 podemos concluir que los términos más comunes (menor valor de IDF) son:

- **of:** 0.097
- **this:** 0.046
- **and:** 0.022
- **with:** 0.155
- **a:** 0.155

Como era de esperar, las palabras que más se repiten dentro del conjunto de documentos se corresponden con preposiciones, conjunciones, artículos y pronombres. Esto no aporta información relevante acerca del corpus, por ello para poder obtener esta información se han omitido estas palabras para buscar las más comunes fuera de este grupo, son las siguientes:

- **wine:** 0.26
- **fruit:** 0.301
- **palate:** 0.602
- **flavor:** 0.602
- **finish:** 0.602

Con estos términos, ya es posible tener una idea sobre la temática del corpus.

2.2.3. TF - IDF

La medida **TF - IDF** se define como:

$$\textbf{TF-IDF(x) = TF(x) * IDF(x)}$$

Donde **TF(x)** es la frecuencia de la palabra clave x en el documento e **IDF(X)** es el valor IDF para esa palabra.

Por tanto, **TF - IDF** es una medida que indica la importancia de un término en un documento dado. De esta forma da un mayor peso a los términos que son menos comunes en el corpus, por lo que reduce la importancia de términos muy frecuentes.

El valor **TF - IDF** aumenta proporcionalmente con la cantidad de veces que aparece una palabra en el documento, pero es compensada por el número de veces que aparece la palabra dentro de la colección de documentos, lo que permite manejar que algunas palabras sean más comunes que otras.

Si se observa la columna que almacena los valores TF - IDF se puede concluir que los términos más importantes en el documento 10 son:

- **natural:** 1.301
- **movement:** 1.301
- **organic:** 1.301
- **grapes:** 1.301
- **label:** 1.301
- **printed:** 1.301
- **vegetable:** 1.301
- **ink:** 1.301
- **recycled:** 1.301
- **paper:** 1.301
- **very:** 1.301
- **nice:** 1.301

Como se puede observar ya no aparecen preposiciones, conjunciones, artículos o pronombres, en su lugar las palabras más importantes son aquellas que le aportan un significado real al documento.

3. Conclusión

En conclusión, con esta práctica se ha observado la importancia y utilidad de los sistemas de recomendación, en concreto hemos utilizado la recomendación siguiendo el modelo basado en el contenido.

Con ello hemos podido conocer sus características y parámetros tan importantes como TF, IDF, TF-IDF y la similitud coseno, lo que tiene una gran relevancia en la recomendación de documentos con información textual.

Por tanto, con esta práctica hemos adquirido un mayor conocimiento sobre los sistemas que emplean modelos basados en el contenido, lo que es de gran importancia en la actualidad donde estos sistemas se emplean en numerosos ámbitos.
.