Escuela Superior
de Ingeniería y Tecnología
Universidad de La Laguna

Sistemas de Recomendación:

Modelos basados en el contenido

Alexander Morales Díaz

(alu0101463018@ull.edu.es)

Daniel Garvi Arvelo

(alu0101501338@ull.edu.es)

Alba Pérez Rodríguez

(alu0101513768@ull.edu.es)

Guillermo Díaz Bricio

(alu0101505688@ull.edu.es)

Índice:

1. Introducción

El objetivo de este informe es analizar los resultados obtenidos al aplicar un sistema de recomendación basado en el contenido, centrándonos principalmente en el cálculo de la similitud entre pares de documentos. Este enfoque de recomendación basado en el contenido es útil para crear sistemas capaces de recomendar documentos similares entre sí en función de sus características textuales y la relación semántica entre términos.

A continuación, se llevará a cabo un análisis detallado de los resultados obtenidos tras probar la aplicación con distintos conjuntos de documentos de prueba. Este análisis permitirá evaluar la eficacia y precisión del sistema de recomendación basado en el contenido, observando en qué medida es capaz de identificar y recomendar documentos similares en función de sus características textuales.

2. Marco teórico

2.1 Frecuencia de Término (TF)

La frecuencia de término (**TF**) mide el número de veces que un término específico aparece en un documento. Cuanto mayor sea el valor de RF, mayor será la probabilidad de que el término sea relevante para el contenido de dicho documento.

2.2 Frecuencia Inversa de Documento (IDF)

La frecuencia inversa de documento (IDF) mide la rareza de un término en el conjunto de documentos. Se calcula como:

$$IDF(t) = \log(\frac{N}{n_t})$$

donde N es el número total de documentos y n_t es el número de documentos que contienen el término t. Un valor alto de IDF indica que el término es poco común en el conjunto de documentos, y por tanto, podría ser más representativo en un contexto específico.

2.3 TF-IDF

La métrica **TF-IDF** es el producto de TF e IDF:

$$TF - IDF(t, d) = TF(t, d) \times IDF(t)$$

TF-IDF es útil para resaltar términos que son relevantes en un documento, pero que no aparecen frecuentemente en otros documentos del conjunto.

3. Análisis de los resultados

Ejemplo 1. Documento-01

Enlace al ejemplo

https://github.com/ull-cs/gestion-conocimiento/blob/main/recommeder-systems/examples-documents/documents-01.txt

Resultado

Simplificando tomaremos como referencia una pequeña parte del resultado:

N.º	Término	TF	IDF	TF-IDF
1	aromas	1	0.301	0.301
2	include	1	1.000	1.000
3	tropical	1	1.301	1.301

Interpretación de Resultados

Analizando la tabla de valores de TF, IDF y TF-IDF, se observa que:

- Los términos con valores altos de **TF-IDF**, como ***tropical***, poseen un alto valor de IDF, lo cual indica que son términos poco comunes en el conjunto de documentos, sugiriendo que son distintivos y potencialmente significativos en el contexto del documento.
- Términos con **IDF** bajo, como ***aromas***, tienen un menor valor de TF-IDF, lo cual sugiere que son menos relevantes para identificar el contenido único del documento ya que aparecen en otros documentos también.
- El término ***include*** muestra un valor medio de TF-IDF, lo cual puede indicar que es relevante dentro del documento, pero que también puede aparecer en otros documentos del conjunto, aportando un contexto general sin ser necesariamente distintivo.

Análisis General de la Matriz de Similitud

SIMILITUD ENTRE DOCUMENTOS

	[D.0]	[D.1]	[D.2]	[D.3]	[D.4]	[D.5]	[D.6]	[D.7]	[D.8]	[D.9]	[D.10]	[D.11]	[D.12]	[D.13]	[D.14]	[D.15]	[D.16]	[D.17]	[D.18]	[D.19]
[Doc 0] →	1.000	0.009	0.008	0.011	0.000	0.010	0.077	0.050	0.146	0.040	0.052	0.025	0.059	0.017	0.029	0.015	0.004	0.006	0.009	0.033
[Doc 1] →	0.009	1.000	0.019	0.000	0.008	0.007	0.116	0.089	0.065	0.059	0.019	0.090	0.037	0.033	0.000	0.007	0.064	0.000	0.000	0.035
[Doc 2] →	0.008	0.019	1.000	0.083	0.008	0.053	0.010	0.027	0.038	0.056	0.058	0.009	0.035	0.096	0.000	0.047	0.000	0.048	0.000	0.033
[Doc 3] →	0.011	0.000	0.083	1.000	0.000	0.014	0.013	0.000	0.017	0.000	0.000	0.033	0.069	0.022	0.065	0.018	0.014	0.015	0.056	0.084
[Doc 4] →	0.000	0.008	0.008	0.000	1.000	0.038	0.000	0.011	0.007	0.000	0.015	0.040	0.000	0.027	0.000	0.000	0.026	0.022	0.000	0.000
[Doc 5] →	0.010	0.007	0.053	0.014	0.038	1.000	0.035	0.011	0.032	0.034	0.045	0.000	0.035	0.022	0.000	0.070	0.013	0.104	0.038	0.042
[Doc 6] →	0.077	0.116	0.010	0.013	0.000	0.035	1.000	0.058	0.084	0.081	0.030	0.000	0.088	0.000	0.008	0.037	0.004	0.035	0.071	0.038
[Doc 7] →	0.050	0.089	0.027	0.000	0.011	0.011	0.058	1.000	0.013	0.283	0.214	0.112	0.000	0.009	0.036	0.038	0.000	0.000	0.058	0.036
[Doc 8] →	0.146	0.065	0.038	0.017	0.007	0.032	0.084	0.013	1.000	0.031	0.017	0.008	0.000	0.015	0.063	0.010	0.000	0.010	0.026	0.034
[Doc 9] →	0.040	0.059	0.056	0.000	0.000	0.034	0.081	0.283	0.031	1.000	0.126	0.000	0.000	0.022	0.056	0.062	0.072	0.000	0.000	0.000
[Doc 10] →	0.052	0.019	0.058	0.000	0.015	0.045	0.030	0.214	0.017	0.126	1.000	0.018	0.000	0.012	0.024	0.000	0.046	0.000	0.019	0.024
[Doc 11] →	0.025	0.090	0.009	0.033	0.040	0.000	0.000	0.112	0.008	0.000	0.018	1.000	0.025	0.006	0.000	0.019	0.074	0.026	0.140	0.072
[Doc 12] →	0.059	0.037	0.035	0.069	0.000	0.035	0.088	0.000	0.000	0.000	0.000	0.025	1.000	0.000	0.007	0.072	0.055	0.065	0.026	0.101
[Doc 13] →	0.017	0.033	0.096	0.022	0.027	0.022	0.000	0.009	0.015	0.022	0.012	0.006	0.000	1.000	0.016	0.020	0.020	0.007	0.025	0.044
[Doc 14] →	0.029	0.000	0.000	0.065	0.000	0.000	0.008	0.036	0.063	0.056	0.024	0.000	0.007	0.016	1.000	0.005	0.000	0.000	0.029	0.037
[Doc 15] →	0.015	0.007	0.047	0.018	0.000	0.070	0.037	0.038	0.010	0.062	0.000	0.019	0.072	0.020	0.005	1.000	0.063	0.020	0.033	0.043
[Doc 16] →	0.004	0.064	0.000	0.014	0.026	0.013	0.004	0.000	0.000	0.072	0.046	0.074	0.055	0.020	0.000	0.063	1.000	0.014	0.003	0.004
[Doc 17] →	0.006	0.000	0.048	0.015	0.022	0.104	0.035	0.000	0.010	0.000	0.000	0.026	0.065	0.007	0.000	0.020	0.014	1.000	0.049	0.006
[Doc 18] →	0.009	0.000	0.000	0.056	0.000	0.038	0.071	0.058	0.026	0.000	0.019	0.140	0.026	0.025	0.029	0.033	0.003	0.049	1.000	0.073
[Doc 19] →	0.033	0.035	0.033	0.084	0.000	0.042	0.038	0.036	0.034	0.000	0.024	0.072	0.101	0.044	0.037	0.043	0.004	0.006	0.073	1.000

Observamos que la mayoría de los documentos tienen una baja similitud con otros documentos, con valores de similitud que oscilan en gran medida por debajo de 0.1. Esto implica que cada documento posee características bastantes distintivas, sin que exista un solapamiento significativo de términos o temas entre ellos.

Sin embargo, algunos pares de documentos tienen similitudes destacables, que analizaremos en detalle en la siguiente sección.

Identificación de Pares de Documentos con Alta Similitud

A continuación, se comentan los pares de documentos con mayor valor de similitud (mayor que 0.1) en la matriz.

- **Documento 7** y **Documento 9**: Tienen una similitud de 0.283. Esto sugiere que estos documentos comparten una cantidad significativa de términos o

temas en común. Es posible que ambos documenten contenido o ideas relacionadas o, potencialmente, un campo de especialización similar.
- **Documento 7** y **Documento 10**: Tienen una similitud de 0.214, lo cual también es elevado. La similitud entre estos documentos es indicativa de una correlación fuerte en el vocabulario y/o en la estructura de los temas tratados, similar al caso anterior.
- **Documento 9** y **Documento 10**: Presentan una similitud de 0.126, aunque no tan alta como en el caso anterior, es suficiente para sugerir una relación o continuidad en temas entre estos documentos.
- **Documento 11** y **Documento 18**: Con una similitud de 0.140, este par de documentos parece compartir ciertas palabras clave, siendo así más relacionados que otros pares.
- **Documento 6** y **Documento 1**: La similitud entre estos documentos es 0.116, lo que indica un cierto solapamiento, aunque no tan marcado como en los casos previos.

Estos pares indican agrupaciones temáticas o estilísticas entre los documentos, lo cual puede ser útil para identificar temas o clasificar los documentos en grupos de contenido.

Interpretación de Pares No Relacionados

Los documentos con valores de similitud cercanos a 0.0 (como [Doc 4] y [Doc 3], o [Doc 17] y [Doc 1]) tienen un contenido claramente diferenciado entre sí. Estos valores indican que los términos en un documento son, en su mayoría, únicos respecto a los otros, sugiriendo que cada documento trata sobre temas o contenidos diferentes, sin un vocabulario común significativo.

Ejemplo 2. Documento-05

Enlace al ejemplo

https://github.com/ull-cs/gestion-conocimiento/blob/main/recommeder-systems/examples-documents/documents-05.txt

Resultado

Simplificando tomaremos como referencia una pequeña parte del resultado:

N.º	Término	TF	IDF	TF-IDF
1	magnificent	1	1.301	1.301
2	sunset	1	1.301	1.301
3	paint	1	1.301	1.301
4	sky	1	0.699	0.699
5	vibrant	1	1.000	1.000
6	shades	1	1.301	1.301
7	orange	1	1.301	1.301
8	pink	1	1.301	1.301
9	purple	1	1.301	1.301
10	reflect	1	1.301	1.301
11	calm	1	1.000	1.000
12	waters	1	1.301	1.301
13	ocean	1	1.000	1.000
14	gentle	1	1.000	1.000
15	breeze	1	1.000	1.000
16	carry	1	1.000	1.000
17	salty	1	1.301	1.301
18	scent	1	0.824	0.824
19	sea	1	1.301	1.301

Interpretación de Resultados

- Los términos con valores altos de **TF-IDF**, como **magnificent**, **sunset**, y **vibrant** (aparecen 1 vez juntando todos los documentos), poseen un alto valor de **IDF**, lo cual indica que son términos relativamente raros en el conjunto de documentos. Esto sugiere que son términos distintivos y potencialmente significativos dentro de cada documento, ayudando a captar la esencia o los detalles específicos del texto.
- Términos con **IDF** más bajos, como **scent** y **sky** (aparecen 3 y 4 veces juntando todos los documentos), tienen un menor valor de **TF-IDF**, lo cual indica que estos términos son más comunes y aparecen con frecuencia en otros documentos también. Estos términos son menos útiles para identificar el contenido único de cada documento, ya que no son lo suficientemente distintivos.
- Algunos términos, como **gentle**, **breeze**, y **calm** (aparecen 2 veces juntando todos los documentos), muestran valores de **TF-IDF** intermedios, lo cual puede sugerir que son relevantes dentro del contexto del documento, pero también pueden aparecer en otros textos dentro del conjunto. Esto puede implicar que estos términos tienen importancia en el contexto del documento, pero no son tan exclusivos o representativos para diferenciar el documento de otros.

En general, los términos con **TF-IDF** más altos son aquellos que pueden ser considerados como claves o elementos representativos del documento, mientras que los términos con

TF-IDF más bajos contribuyen menos a la diferenciación del contenido específico de cada documento.

Análisis General de la Matriz de Similitud

SIMILITUD ENTRE DOCUMENTOS

	[D.0]	[D.1]	[D.2]	[D.3]	[D.4]	[D.5]	[D.6]	[D.7]	[D.8]	[D.9]	[D.10]	[D.11]	[D.12]	[D.13]	[D.14]	[D.15]	[D.16]	[D.17]	[D.18]	[D.19]
[Doc 0] ->	1.000	0.000	0.028	0.021	0.043	0.000	0.000	0.000	0.000	0.000	0.113	0.000	0.000	0.023	0.039	0.000	0.042	0.000	0.019	0.040
[Doc 1] ->	0.000	1.000	0.000	0.000	0.000	0.027	0.000	0.000	0.026	0.025	0.069	0.027	0.025	0.000	0.026	0.000	0.028	0.044	0.000	0.000
[Doc 2] ->	0.028	0.000	1.000	0.045	0.000	0.000	0.000	0.000	0.000	0.040	0.031	0.000	0.000	0.000	0.000	0.044	0.000	0.000	0.000	0.000
[Doc 3] ->	0.021	0.000	0.045	1.000	0.000	0.059	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.043	0.000	0.060	0.045	0.000	0.021	0.016
[Doc 4] ->	0.043	0.000	0.000	0.000	1.000	0.000	0.000	0.045	0.000	0.000	0.031	0.030	0.028	0.000	0.000	0.000	0.031	0.000	0.000	0.044
[Doc 5] ->	0.000	0.027	0.000	0.059	0.000	1.000	0.026	0.042	0.066	0.000	0.000	0.000	0.000	0.065	0.000	0.015	0.000	0.000	0.000	0.043
[Doc 6] ->	0.000	0.000	0.000	0.000	0.000	0.026	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.044	0.000	0.000	0.000	0.000	0.000	0.026
[Doc 7] ->	0.000	0.000	0.000	0.000	0.045	0.042	0.000	1.000	0.000	0.000	0.000	0.000	0.040	0.000	0.000	0.043	0.088	0.000	0.083	0.000
[Doc 8] ->	0.000	0.026	0.000	0.000	0.000	0.066	0.000	0.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.041	0.000	0.000	0.000
[Doc 9] ->	0.000	0.025	0.040	0.000	0.000	0.000	0.000	0.000	0.000	1.000	0.000	0.027	0.037	0.000	0.038	0.000	0.000	0.000	0.000	0.000
[Doc 10] ->0.113		0.069	0.031	0.000	0.031	0.000	0.000	0.000	0.000	0.000	1.000	0.030	0.000	0.000	0.000	0.000	0.031	0.000	0.000	0.000
[Doc 11] ->0.000		0.027	0.000	0.000	0.030	0.000	0.000	0.000	0.000	0.027	0.030	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
[Doc 12] ->0.000		0.025	0.000	0.000	0.028	0.000	0.000	0.040	0.000	0.037	0.000	0.000	1.000	0.000	0.026	0.000	0.028	0.000	0.000	0.000
[Doc 13] ->0.023		0.000	0.000	0.043	0.000	0.065	0.044	0.000	0.000	0.000	0.000	0.000	0.000	1.000	0.000	0.018	0.000	0.000	0.023	0.017
[Doc 14] ->0.039		0.026	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.038	0.000	0.000	0.026	0.000	1.000	0.000	0.029	0.000	0.027	0.000
[Doc 15] ->0.000		0.000	0.044	0.060	0.000	0.015	0.000	0.043	0.000	0.000	0.000	0.000	0.000	0.018	0.000	1.000	0.000	0.000	0.000	0.015
[Doc 16] ->0.042		0.028	0.000	0.045	0.031	0.000	0.000	0.088	0.041	0.000	0.031	0.000	0.028	0.000	0.029	0.000	1.000	0.000	0.029	0.000
[Doc 17] ->0.000		0.044	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	1.000	0.000	0.047
[Doc 18] ->0.019		0.000	0.000	0.021	0.000	0.000	0.000	0.083	0.000	0.000	0.000	0.000	0.000	0.023	0.027	0.000	0.029	0.000	1.000	0.000
[Doc 19] ->0.040		0.000	0.000	0.016	0.044	0.043	0.026	0.000	0.000	0.000	0.000	0.000	0.000	0.017	0.000	0.015	0.000	0.047	0.000	1.000

Observamos que la mayoría de los documentos tienen una similitud baja con otros documentos, con valores que oscilan generalmente por debajo de 0.1. Esto indica que los documentos en su conjunto son relativamente diversos y no comparten muchos términos o ideas en común.

Identificación de Pares de Documentos con Alta Similitud

A continuación, se destacan los pares de documentos con los valores de similitud más altos:

- **Documento 0** y **Documento 10**: Tienen una similitud de 0.113, lo que sugiere que ambos comparten ciertos temas o vocabulario. Aunque no es una similitud extremadamente alta, es suficiente para indicar que podría tener alguna relación en términos de contenido o enfoque.
- **Documento 7** y **Documento 16**: Con una similitud de 0.088, este par también muestra una relación baja. Los documentos podrían estar relacionados a través de algunos términos, aunque no se superpongan en gran medida.
- **Documento 5** y **Documento 7**: Con una similitud de 0.066, la relación entre estos documentos es baja. Esto podría sugerir que ambos tratan temas similares pero con diferencias notables en el vocabulario o el enfoque.
- **Documento 6** y **Documento 9**: La similitud entre estos documentos es 0.040, lo que es relativamente bajo, por lo que sugiere que se comparten pocos términos.

Interpretación de Pares No Relacionados

Los documentos con valores de similitud cercanos a 0.0 (como [Doc 0] y [Doc 1], o [Doc 6] y [Doc 13]) tienen un contenido claramente diferenciado entre sí. Estos valores indican que los términos en un documento son, en su mayoría, únicos respecto a los otros, sugiriendo que cada documento trata sobre temas o contenidos diferentes, sin un vocabulario común significativo.