Grado en Ingeniería Informática

Gestión del conocimiento en las organizaciones

Sistemas de recomendación. Modelos basados en el contenido

Lorenzo Román Luca de Tena, Mariam Laaroussi Ramos

ÍNDICE

Descripción del Análisis

El objetivo principal del análisis realizado es calcular la similaridad entre documentos a partir de la frecuencia de términos (TF), la frecuencia inversa de documentos (IDF) y el valor combinado TF-IDF. El análisis se centró en la implementación de un algoritmo que procesa documentos textuales, calcula estos valores y evalúa la similitud entre ellos mediante el uso de la similitud coseno.

El sistema está compuesto por dos archivos principales de código:

1. **main.cc**: Contiene la función principal del programa y coordina las tareas del sistema.

2. **tools.cc**: Contiene funciones auxiliares utilizadas en el proceso de análisis de los documentos.

El proceso comienza con la lectura de los documentos, las palabras vacías (stopwords) y las lematizaciones desde archivos proporcionados por el usuario. Luego, se calculan los valores de TF, IDF y TF-IDF para cada término en los documentos. Finalmente, se calculan las similitudes coseno entre los documentos, las cuales se imprimen junto con los resultados obtenidos.

El algoritmo utiliza un enfoque de preprocesamiento para optimizar los cálculos, como la eliminación de palabras vacías y la lematización, lo que mejora la precisión de los resultados al reducir la variabilidad en los términos utilizados en los documentos.

Análisis realizado

Se han analizado tres frases relacionadas con vinos utilizando la técnica de análisis de texto basada en la frecuencia de términos (TF), la frecuencia inversa de documentos (IDF) y el producto de ambas (TF-IDF). Los valores obtenidos nos permiten identificar cuáles son las palabras más relevantes en cada documento, excluyendo las palabras comunes de la lista de "stop words", y calcular la similitud entre los documentos.

Documentos Analizados
Documento 1: "Aromas include tropical fruit, broom, brimstone and dried herb. The palate isn't overly expressive, offering unripened apple, citrus and dried sage alongside brisk acidity."

Documento 2: "This is ripe and fruity, a wine that is smooth while still structured. Firm tannins are filled out with juicy red berry fruits and freshened with acidity. It's already drinkable, although it will certainly be better from 2016."

Documento 3: "Tart and snappy, the flavors of lime flesh and rind dominate. Some green pineapple pokes through, with crisp acidity underscoring the flavors. The wine was all stainless-steel fermented."

Cálculos de TF, IDF y TF-IDF
El análisis realizado para cada documento ha considerado el cálculo de las frecuencias de términos (TF) y la frecuencia inversa de documentos (IDF) de las palabras clave. Las palabras que son comunes a todos los documentos, como "acidity", "and", "with", se excluyeron de los cálculos porque están en la lista de "stop words".

Documento 1:

Index	Word	TF	IDF	TF-IDF
0	sage	0.042	3	0.12
1	apple,	0.042	3	0.12
2	offering	0.042	2.3	0.096
3	expressive,	0.042	3	0.12
4	Aromas	0.042	3	0.12
5	acidity.	0.042	1.6	0.067
6	citrus	0.042	2.3	0.096
7	include	0.042	2.3	0.096
8	tropical	0.042	3	0.12
9	alongside	0.042	3	0.12

10	broom,	0.042	3	0.12
11	isn't	0.042	3	0.12
12	brisk	0.042	3	0.12
13	overly	0.042	3	0.12
14	palate	0.042	1.4	0.058
15	brimstone	0.042	3	0.12
16	unripened	0.042	3	0.12
17	and	0.083	0	0
18	dried	0.083	2.3	0.19
19	The	0.042	1.4	0.058
20	fruit,	0.042	1.9	0.079
21	herb.	0.042	3	0.12

Documento 2:

Index	Word	TF	IDF	TF-IDF
0	better	0.026	3	0.079
1	be	0.026	2.3	0.061

2	will	0.026	3	0.079
3	while	0.026	1.6	0.042
4	certainly	0.026	3	0.079
5	structured.	0.026	3	0.079
6	wine	0.026	1.4	0.036
7	that	0.026	1.4	0.036
8	filled	0.026	3	0.079
9	smooth	0.026	3	0.079
10	fruity,	0.026	3	0.079
11	with	0.053	0.16	0.0086
12	It's	0.026	1.4	0.036
13	ripe	0.026	2.3	0.061
14	already	0.026	3	0.079
15	is	0.053	0.92	0.048
16	This	0.026	1	0.028

17	from	0.026	1.9	0.05
18	tannins	0.026	2.3	0.061
19	red	0.026	1.9	0.05
20	are	0.026	2.3	0.061
21	still	0.026	3	0.079
22	juicy	0.026	1.9	0.05
23	2016.	0.026	3	0.079
24	berry	0.026	3	0.079
25	fruits	0.026	1.9	0.05
26	a	0.026	0.29	0.0076
27	freshened	0.026	3	0.079
28	although	0.026	3	0.079
29	Firm	0.026	3	0.079
30	acidity.	0.026	1.6	0.042
31	and	0.053	0	0

32	drinkable,	0.026	3	0.079
33	out	0.026	3	0.079
34	it	0.026	1.9	0.05

Documento 3:

Index	Word	TF	IDF	TF-IDF
0	all	0.036	3	0.11
1	was	0.036	3	0.11
2	wine	0.036	1.4	0.05
3	The	0.036	1.4	0.05
4	fermented.	0.036	3	0.11
5	acidity	0.036	1.6	0.057
6	crisp	0.036	1.9	0.068
7	pokes	0.036	3	0.11
8	Some	0.036	3	0.11
9	pineapple	0.036	2.3	0.082

10	with	0.036	0.16	0.0058
11	and	0.071	0	0
12	underscoring	0.036	3	0.11
13	Tart	0.036	3	0.11
14	flesh	0.036	3	0.11
15	snappy,	0.036	3	0.11
16	the	0.071	0.51	0.036
17	flavors	0.036	1.4	0.05
18	through,	0.036	2.3	0.082
19	of	0.036	0.43	0.015
20	flavors.	0.036	3	0.11
21	lime	0.036	3	0.11
22	rind	0.036	3	0.11
23	stainless-steel	0.036	3	0.11

24	dominate.	0.036	3	0.11
25	green	0.036	1.6	0.057

Palabras Relevantes: En cada documento, las palabras más relevantes tienen un valor TF-IDF alto. Por ejemplo, en el Documento 1, las palabras "sage", "apple", "tropical" y "acidity" tienen un TF-IDF significativo, lo que indica que son cruciales para entender el contenido del documento. Similarmente, en el Documento 2, términos como "better", "smooth", "tannins" y "fruity" son destacadas.

Similitudes y Diferencias: A pesar de que los tres documentos comparten algunas palabras (por ejemplo, "acidity", "wine"), las palabras más características varían entre los documentos, lo que refleja la diferencia en los temas tratados. El Documento 3 se centra más en la descripción de sabores específicos, mientras que los Documentos 1 y 2 destacan más las características generales del vino.

Uso de TF-IDF: El valor TF-IDF es útil para identificar las palabras más importantes en cada documento y resaltar las diferencias clave entre ellos. Las palabras con valores más altos de TF-IDF son aquellas que mejor definen el contenido del documento, lo que ayuda a comparar la similitud y los temas abordados.

Conclusiones Extraídas

A partir de los resultados obtenidos, se puede concluir que el cálculo de las similitudes entre documentos es una tarea eficiente para evaluar la relación entre textos. Los pasos de preprocesamiento, como la eliminación de las palabras vacías y la lematización, contribuyen significativamente a mejorar la precisión de los resultados. Al utilizar el valor de TF-IDF, se logra un análisis más representativo del contenido relevante de cada documento.

En cuanto a la similitud coseno, este método resultó ser efectivo para medir la relación entre los documentos, proporcionando una medida cuantitativa de similitud entre los textos. Un valor cercano a 1 indica que los documentos son muy similares, mientras que un valor cercano a 0 indica que son muy diferentes.

Es importante destacar que los valores calculados de TF, IDF y TF-IDF son fundamentales para cualquier sistema de búsqueda o análisis de textos, ya que permiten identificar los términos más relevantes en un conjunto de documentos. El uso de estas métricas es común en tareas de minería de texto, recuperación de información y análisis semántico.

Finalmente, el algoritmo proporciona un enfoque robusto para calcular y visualizar la similitud entre los

documentos, lo que puede ser utilizado en una variedad de aplicaciones, desde motores de búsqueda hasta sistemas de recomendación basados en contenido.