Universidad de La Laguna

Sistemas de recomendación. Modelos basados en el contenido

Gestión del conocimiento de las organizaciones

Mariajose Zuloeta Brito
Samuel Lorenzo Sánchez
Adrián Lima García
Andrés Hernández Ortega

Introducción

En esta práctica, implementaremos un sistema de recomendación utilizando el método de filtrado colaborativo. Para comprender su funcionamiento, nos apoyaremos en las diapositivas del curso y desarrollaremos un software con tecnologías web como HTML, CSS3 y Vue. El sistema incluirá los siguientes elementos:

- Fichero de entrada que contendrá los documentos.
- Fichero de entrada que contendrá las stop-words.
- Fichero de entrada que contendrá la información para lematizar.

Como resultado, el sistema deberá generar como salida para cada par de documento:

- Índice del término
- Término
- IDF
- TF-IDF
- Similaridad coseno entre cada par de documentos

Análisis

Los ficheros con la información obtenida en las pruebas se encuentran en el siguiente directorio compartido: Pruebas

Ejemplo de un documento que habla sobre los vinos

Datos de entrada:

- Documento 01.txt
- Palabras de parada.txt
- Fichero de lematización

Resultados: Resultado1.txt

```
Documento 1:

      ÍNDICE |          TÉRMINO |    TF    |    IDF   | TF-IDF

           0 |           aromas | 1.0000 | 0.3010 | 0.2253

           1 |           include | 1.0000 | 1.0000 | 0.2253

           2 |         tropical | 1.0000 | 1.3010 | 0.2253

           3 |            fruit | 1.0000 | 0.6990 | 0.2253

           4 |            broom | 1.0000 | 1.3010 | 0.2253

           5 |        brimstone | 1.0000 | 1.3010 | 0.2253

           6 |            dried | 1.3010 | 1.0000 | 0.2932
```

Conclusiones

En el análisis de los documentos se observan términos recurrentes y temas comunes que permiten identificar la evaluación de vinos. Los términos de alta relevancia, como "acidity" (acidez), "aromas," "fruit" y "wine," aparecen en varios documentos y mantienen valores altos en TF-IDF, indicando su importancia en las descripciones. Esta recurrencia sugiere que los textos están orientados a evaluar vinos en términos de sus perfiles de sabor, aroma y cuerpo.

Cada documento describe perfiles sensoriales particulares, lo que permite distinguir los vinos según sus atributos. Por ejemplo, el Documento 1

menciona aromas tropicales, notas de frutas secas, y un perfil ligeramente ácido, lo que sugiere un vino con matices complejos y frescos. El Documento 2, en cambio, describe un vino "ripe" (maduro) y "fruity" (afrutado)

Además, el análisis revela que aunque algunos términos son comunes en todos los documentos, cada uno presenta un perfil propio que contribuye a la variedad de los vinos.

Ejemplo de un documento que describe la vida diaria en una urbanización

Datos de entrada:

- Documento 02.txt
- Palabras de parada.txt
- Fichero de lematización

Resultados: Resultado 02.txt

```
Documento 1:

 ÍNDICE |      TÉRMINO |   TF   |   IDF  | TF-IDF

      0 |      morning | 1.0000 | 1.0000 | 0.3162

      1 |          sun | 1.0000 | 0.8239 | 0.3162

      2 |        broke | 1.0000 | 1.3010 | 0.3162

      3 |       clouds | 1.0000 | 1.3010 | 0.3162

      4 |      casting | 1.0000 | 1.3010 | 0.3162

      5 |            a | 1.0000 | 0.1549 | 0.3162

      6 |       golden | 1.0000 | 1.3010 | 0.3162
```

Conclusiones:

El análisis de los documentos muestra que ciertos términos como "morning," "sun," "clouds," y "casting" en el Documento 1 tienen un TF alto, lo que indica que son palabras recurrentes dentro del texto. Sin embargo, el valor de IDF (Frecuencia de Documento Inversa) se mantiene relativamente bajo, lo que sugiere que estos términos son comunes en otros textos. El TF-IDF, que combina estos dos valores, proporciona una medida de la relevancia de un

término dentro del contexto general de un corpus, y en este caso, se observa que las palabras con alta frecuencia de aparición y relevancia global, como "morning" y "sun," tienen valores similares, lo que indica que son palabras clave dentro de su documento.

En los documentos 2 y 3, los términos más representativos como "breeze," "flowers," "people," y "traffic" muestran una frecuencia consistente en su aparición, con un valor de IDF más alto en términos menos comunes. Este patrón refleja que, aunque estas palabras pueden aparecer en varios documentos, su relevancia se ve incrementada por la forma en que contribuyen al sentido general del texto. El TF-IDF de estas palabras sugiere que términos como "breeze" o "traffic".

Ejemplo de un documento que describe la variedad de vinos

Datos de entrada:

- Documento 03.txt
- Palabras de parada.txt
- Fichero de lematización

Resultados: resultado 03.txt

```
Documento 1:
```

ÍNDICE	TÉRMINO	TF	IDF	TF-IDF
0	red	1.0000	0.8239	0.2887
1	cherry	1.0000	1.3010	0.2887
2	fruit	1.0000	0.3010	0.2887
3	laced	1.0000	1.3010	0.2887
4	light	1.0000	1.3010	0.2887
5	tannins	1.0000	0.6990	0.2887
6	giving	1.0000	1.0000	0.2887

Conclusiones

El análisis en los primeros documentos, se observan términos recurrentes como "red", "cherry", "fruit", "bright" y "wine", que sugieren una descripción

general de las características de un vino. Estos términos tienen valores altos de TF, lo que indica su presencia dentro de los textos, y valores de IDF que reflejan su relevancia en el conjunto de documentos analizados.

A medida que se avanza en los documentos, los términos más específicos como "merlot", "nero d'Avola", "fettuccine" o "sauvignon", comienzan a dominar, lo que revela una mayor especificidad en las descripciones de los vinos y sus maridajes. En documentos como el 2 y el 5, el TF-IDF resalta la importancia de términos vinculados a variedades de uvas, estilos de vinos y combinaciones culinarias, lo que indica que estos documentos se enfocan en una recomendación más detallada, tanto del producto como de sus posibles acompañamientos.

Ejemplo de un documento que habla sobre las nuevas tecnologías

Datos de entrada:

- Documento 04.txt
- Palabras de parada.txt
- Fichero de lematización

Resultados: Resultado 04.txt

```
Documento 1:
 ÍNDICE |         TÉRMINO |    TF    |    IDF   | TF-IDF
      0 |          todays | 1.0000 | 1.3010 | 0.2182
      1 | interconnected | 1.0000 | 1.3010 | 0.2182
      2 |           world | 1.0000 | 0.6990 | 0.2182
      3 |          advent | 1.0000 | 1.3010 | 0.2182
      4 |        advanced | 1.0000 | 1.3010 | 0.2182
      5 |  communication | 1.0000 | 1.0000 | 0.2182
      6 |   technologies | 1.0000 | 0.6990 | 0.2182
```

Conclusiones

En el análisis del documento aparecen términos como "world," "rise," y "education," que aparecen repetidamente en los textos y refleja la importancia de los avances tecnológicos y educación en la discusión.

En los documentos se encuentran términos relacionados con la inteligencia artificial, el cambio climático, la educación y la evolución del transporte indican que estos son puntos importantes en los debates actuales. A través del análisis de TF-IDF, se evidencia que palabras como "innovation," "climate," "sustainability," y "opportunities" tienen una presencia significativa en las discusiones sobre el futuro y el progreso, lo que refuerza la idea de que la tecnología, el medio ambiente y el desarrollo humano son temas cruciales.

Finalmente, el análisis de los términos muestra que el concepto de "e-commerce" en el quinto documento está relacionado con el cambio en el consumo y las tiendas tradicionales, el sexto documento pone de manifiesto la evolución del transporte, destacando la necesidad de adaptarse a nuevas tecnologías.

Ejemplo de un documento que describe momentos de belleza natural

Datos de entrada:

- Documentos 05.txt
- Palabras de parada.txt
- Fichero de lematización

Resultados: Resultado 05.txt

```
Documento 1:

       ÍNDICE |        TÉRMINO |    TF    |    IDF   | TF-IDF

            0 |    magnificent | 1.0000 | 1.3010 | 0.2294

            1 |         sunset | 1.0000 | 1.3010 | 0.2294

            2 |        painted | 1.0000 | 1.3010 | 0.2294

            3 |            sky | 1.0000 | 0.6990 | 0.2294

            4 |        vibrant | 1.0000 | 1.0000 | 0.2294
```

Conclusiones

Los términos que aparecen con mayor frecuencia en cada documento tienen un valor de TF de 1.0, lo que indica que son palabras clave en su contexto. La diferencia en los valores de IDF resalta los términos que tienen un alto impacto en el documento, aquellos que son poco comunes y, por tanto, más representativos del contenido de cada texto. Esto se refleja en el cálculo del TF-IDF, que combina ambas métricas para ponderar la relevancia de cada término dentro del corpus.

Los documentos en su mayoría tienen términos con valores de TF-IDF cercanos entre sí, aunque algunas palabras destacan ligeramente, como "scent" en el Documento 3, que tiene un TF-IDF de 0.2500, indicando una relevancia significativa. Los términos con valores elevados en IDF, como "library", "ocean" y "storm", indican que estas palabras no se repiten con frecuencia en el corpus general, lo que les otorga un peso mayor en la interpretación del significado de los documentos. La repetición de términos como "a", "feel" y "sky", aunque con un IDF bajo, muestra su papel común en el lenguaje pero sin mucho poder diferenciador.

Finalmente, los TF-IDF también reflejan las temáticas principales de los documentos. En los textos donde se describen paisajes o ambientes naturales (como el Documento 1 y el Documento 5), los términos relacionados con la naturaleza, como "ocean", "sky" y "breeze", muestran un TF-IDF notablemente alto. Por otro lado, en documentos que aparecen escenas más dinámicas o urbanas, como el Documento 6, términos como "city", "cars" y "honking" presentan un patrón similar.

Ejemplo de un documento que muestra momentos de belleza de una ciudad bulliciosa

Datos de entrada:

- Documento 06.txt
- Palabras de parada.txt
- Fichero de lematización

Resultados: Resultado 06.txt

```
Documento 1:

ÍNDICE |       TÉRMINO |   TF   |   IDF   | TF-IDF

     0 |           sky | 1.0000 | 0.8239 | 0.2123

     1 |       morning | 1.0000 | 1.3010 | 0.2123
```

```
2 |             a | 1.4771 | 0.1871 | 0.3136

3 |     brilliant | 1.0000 | 1.3010 | 0.2123

4 |         shade | 1.0000 | 1.0000 | 0.2123

5 |          blue | 1.0000 | 1.3010 | 0.2123

6 |   interrupted | 1.0000 | 1.3010 | 0.2123
```

Conclusiones

El análisis de estos valores, se puede observar que términos como "a", "city", "train" o "mountain" tienen valores de TF elevados, lo que indica su presencia frecuente en los textos, pero sus valores de IDF son relativamente bajos, lo que sugiere que estos términos son comunes en muchos documentos y, por lo tanto, no aportan una gran distinción o especificidad.

Algunas palabras específicas como "nostalgia", "dream", "library" y "secrets" tienen valores de IDF más altos, lo que indica que son más exclusivas de ciertos documentos y, por lo tanto, juegan un papel más significativo en la identificación. Estos términos, al tener una mayor presencia en un único contexto, se convierten en indicadores clave que podrían ayudar a distinguir o categorizar los documentos de manera más efectiva.

Ejemplo de un documento que habla sobre una serie de escenas vívidas

Datos de entrada:

- Documento 07.txt
- Palabras de parada.txt
- Fichero de lematización

Resultados: Resultado 07.txt

```
Documento 1:

        ÍNDICE |        TÉRMINO |    TF    |    IDF   | TF-IDF

             0 |          early | 1.0000 | 1.3010 | 0.2582

             1 |        morning | 1.0000 | 1.3010 | 0.2582

             2 |            sun | 1.0000 | 1.3010 | 0.2582

             3 |       filtered | 1.0000 | 1.3010 | 0.2582

             4 |         leaves | 1.0000 | 1.3010 | 0.2582

             5 |        casting | 1.0000 | 1.3010 | 0.2582

             6 |        dappled | 1.0000 | 1.3010 | 0.2582
```

Conclusiones

En los documentos analizados, se observa que la frecuencia de los términos (TF) es constante, con un valor de 1.0000 para todas las palabras, lo que indica que cada término aparece una sola vez en su respectivo documento. Sin embargo, la variabilidad se encuentra en los valores de IDF (Inversa de la Frecuencia de Documentos), que reflejan la importancia relativa de cada término en el corpus total. Los términos con mayor IDF, como "sun", "edge" o "chef", son aquellos que aparecen en pocos documentos y, por lo tanto, se consideran más distintivos en sus respectivos textos.

Los términos con valores más altos de TF-IDF, como "chef" o "cliff", son más representativos de los documentos en los que aparecen, ya que tienen una mayor importancia contextual dentro de esos textos.

Ejemplo de un documento que describe escenas de la vida cotidiana

Datos de entrada:

- Documento 08.txt
- Palabras de parada.txt
- Fichero de lematización

Resultados: Resultado 08.txt

Documento 1:

ÍNDICE	TÉRMINO	TF	IDF	TF-IDF
0	sun	1.0000	1.3010	0.3162
1	set	1.0000	1.3010	0.3162
2	slowly	1.0000	1.3010	0.3162
3	mountains	1.0000	1.3010	0.3162
4	painting	1.0000	1.0000	0.3162
5	sky	1.0000	0.8239	0.3162
6	hues	1.0000	1.3010	0.3162

Conclusiones

Se observa que las palabras con una alta frecuencia en cada documento, como "sun", "bustling", "oak", "ship", y "opened", se destacan por tener un impacto considerable dentro de su contexto específico..

Los términos con una baja frecuencia en varios documentos (por ejemplo, "oak" o "breeze") pueden tener un valor de TF-IDF relativamente alto, sugiriendo que estos términos son significativos dentro de sus respectivos textos, a pesar de su presencia más aislada.

Ejemplo de un documento que muestra los avances sociales y tecnológicos

Datos de entrada:

- Documento 09.txt
- Palabras de parada.txt
- Fichero de lematización

Resultados: [Resultado 09.txt]

```
Documento 1:
```

ÍNDICE	TÉRMINO	TF	IDF	TF-IDF
0	serene	1.0000	1.3010	0.2236
1	afternoon	1.0000	1.3010	0.2236
2	sky	1.0000	1.3010	0.2236
3	painted	1.0000	1.3010	0.2236
4	a	1.0000	0.1549	0.2236
5	thousand	1.0000	1.3010	0.2236
6	shades	1.0000	1.3010	0.2236

Conclusiones

En el Documento 1, los términos relacionados con la naturaleza, como "serene," "sky," "sun," y "mountains," presentan un TF-IDF muy similar, lo que indica una alta frecuencia de aparición en el documento, pero también una distribución bastante equitativa de su importancia dentro del texto. Esto sugiere que el enfoque principal del documento es una descripción detallada y armoniosa de un paisaje

Las palabras clave relacionadas con la economía y la colaboración, como "challenging," "economic," "environment," "initiative," y "fostering," muestran un patrón en el que el TF-IDF es moderadamente bajo, indicando que el texto es más técnico y tiene una alta frecuencia de términos comunes, pero con una importancia algo más dispersa en su estructura.

Finalmente, otros documentos muestran un patrón más diverso, donde los términos asociados con la interacción social ("laughter," "conversation," "families," "children") y los avances tecnológicos ("artificial," "intelligence," "industries," "efficiency") tienen una TF-IDF relativamente más equilibrada, reflejando temas de dinamismo social en el primer caso y de innovación en el segundo.

Ejemplo de un documento que habla sobre los avances tecnológicos

Datos de entrada:

- Documento 10.txt
- Palabras de parada.txt
- Fichero de lematización

Resultados: Resultado 10.txt

```
Documento 1:

       ÍNDICE |             TÉRMINO |    TF    |    IDF   | TF-IDF

            0 |               world | 1.0000 | 0.6021 | 0.3015

            1 |            changing | 1.0000 | 1.3010 | 0.3015

            2 |      unprecedented | 1.0000 | 1.3010 | 0.3015

            3 |                pace | 1.0000 | 1.3010 | 0.3015

            4 |                 due | 1.0000 | 1.3010 | 0.3015

            5 |            advances | 1.0000 | 1.3010 | 0.3015

            6 |          technology | 1.0000 | 1.3010 | 0.3015
```

Conclusiones

En el Documento 1, las palabras como "world," "changing," "advances," y "technology" dominan el contenido, lo que indica una reflexión sobre los cambios rápidos y significativos impulsados por la tecnología en la sociedad global.

En el Documento 2, el tema principal es el cambio climático, con términos como "climate," "change," "challenges," y "action" que resaltan la urgencia de abordar los impactos del cambio climático a través de acciones humanas.

Por otro lado, el resto de documentos abordan la educación en el contexto de la evolución digital, con términos como "education," "evolving," "online," y "learning platforms" que indican un cambio hacia métodos de enseñanza más accesibles y digitales. Centra su atención en los avances en inteligencia artificial, resaltando tanto sus oportunidades como los dilemas éticos que surgen con su integración en diversas industrias.

Conclusiones

El análisis realizado sobre los diferentes documentos revela aspectos significativos en la frecuencia y relevancia de los términos. En general, el cálculo del TF-IDF nos ha sido de utilidad para determinar los términos más destacados de cada documento. Así, se han podido identificar términos claves relacionados con vinos, tecnología, paisajes naturales, entre otros, que varían en su frecuencia según el tema tratado en cada documento. Los términos con valores altos en TF-IDF reflejan el enfoque principal de cada texto.

En cuanto a los documentos relacionados con el vino y la naturaleza, se observó que las palabras que describen características sensoriales y paisajísticas, como "aromas," "acidity," "sun," "sky" y "morning," son recurrentes. El análisis también mostró que, mientras algunos términos se repiten a través de los documentos, otros son más específicos de ciertos contextos, como las variedades de vino o los elementos naturales, lo que ayuda a diferenciar los textos en función de su enfoque particular.

Finalmente, el estudio de documentos relacionados con la tecnología y los avances sociales demuestra cómo términos clave como "innovación," "sostenibilidad" y "cambio climático" se destacan dentro de los textos. Los documentos sobre la evolución digital y los avances tecnológicos. Estos resultados demuestran cómo el análisis de frecuencias y relevancia de los términos puede ayudar a identificar no solo las temáticas centrales, sino también las relaciones subyacentes entre los diferentes aspectos tratados en los textos analizados.