Sistemas de recomendación

Gestión del Conocimiento en las Organizaciones

Alberto Rios de la Rosa - alu0101235929@ull.edu.es

Alejandro Pérez Álvarez - alu0101215310@ull.edu.es

Daniel Hernández Fajardo - alu0101320489@ull.edu.es

1. Introducción

Este informe corresponde a la práctica de Sistemas de recomendación. Modelos basados en el contenido de la asignatura Gestión del Conocimiento en las Organizaciones (GCO).

Se ha desarrollado un sistema de recomendación con HTML y Javascript siguiendo los modelos basados en el contenido. A este software le entra como parámetros a través del fichero HTML lo siguiente:

- Fichero de texto plano con extensión .txt. Cada documento viene representado en una línea del fichero.
- Fichero con palabras de parada, stop words, a descartar durante el proceso de recomendación.
- Fichero de lematización de términos.

Cada documento viene representado en una línea del archivo. Y muestra como salida:

- Para cada documento, tabla con las siguientes columnas:
    - Índice del término.
    - Término.
    - TF.
    - IDF.
    - TF-IDF.

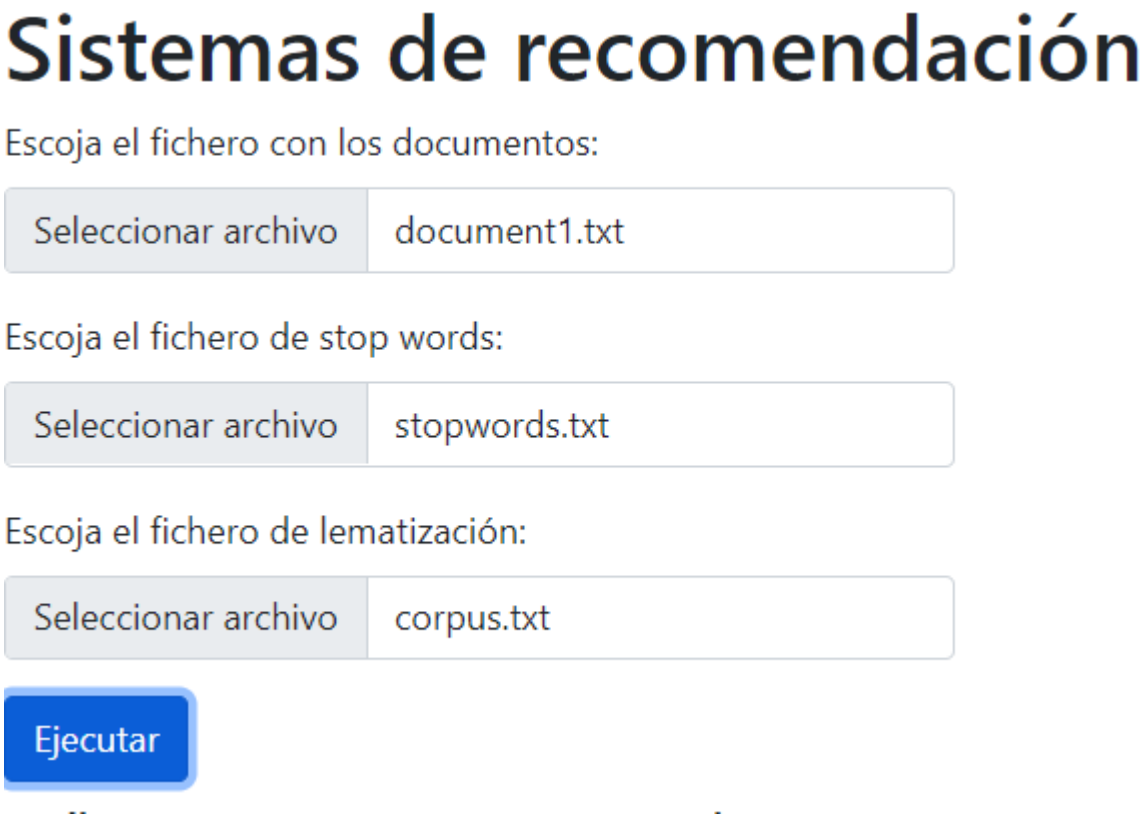
- Similaridad coseno entre cada par de documentos.

2. Ejemplos

A continuación mostraremos la salida de varios documentos, a modo de ejemplo.

Los archivos introducidos son los que encontramos en nuestro repositorio github.

Datos de entrada:



Salida documento1:

Indice	Término	TF	IDF	TF-IDF
Documento 1				
0	aromas	1	1	1
1	include	1	1	1
2	tropical	1	1	1
3	fruit	1	1	1
4	broom	1	1	1
5	brimstone	1	1	1
6	dried	1.3010299956639813	1	1.3010299956639813
7	herb	1	1	1
8	palate	1	1	1
9	overly	1	1	1
10	expressive	1	1	1
11	offer	1	1	1
12	unripened	1	1	1
13	apple	1	1	1
14	citrus	1	1	1
15	dried	1.3010299956639813	1	1.3010299956639813
16	sage	1	1	1
17	brisk	1	1	1
18	acidity	1	1	1

Similitudes del documento 1

2: 0.8600

3: 0.9331

4: 0.9047

5: 0.8113

6: 0.7995

7: 0.9455

8: 0.8228

9: 0.9320

10: 0.9525

Salida documento2:

Documento 2				
0	be	1.3010299956639813	1	1.3010299956639813
1	ripe	1	1	1
2	fruity	1	1	1
3	a	1	1	1
4	wine	1	1	1
5	be	1.3010299956639813	1	1.3010299956639813
6	smooth	1	1	1
7	still	1	1	1
8	structured	1	1	1
9	firm	1	1	1
10	tannins	1	1	1
11	fill	1	1	1
12	with	1	1	1
13	juicy	1	1	1
14	red	1	1	1
15	berry	1	1	1
16	fruits	1	1	1
17	freshened	1	1	1
18	acidity	1	0.6989700043360189	0.6989700043360189
19	already	1	1	1
20	drinkable	1	1	1
21	it	1	1	1
22	certainly	1	1	1
23	better	1	1	1
24	2016	1	1	1

Similitudes del documento 2

1: 0.8600

3: 0.8905

4: 0.9205

5: 0.9282

6: 0.9106

7: 0.8804

8: 0.7235

9: 0.8147

10: 0.8325

3. Conclusiones

Las conclusiones que se han extraído de analizar los resultados de los ejemplos anteriores son las siguientes:

- La similitud coseno ajustada realizada está estrechamente relacionada con el coeficiente de correlación de Pearson.

- Una vez se han normalizado los TFs, la longitud del vector de los TFs de los términos (Raíz cuadrada de la suma de los valores al cuadrado de cada TF del vector de términos) debe ser 1.