Escuela Superior
de Ingeniería y Tecnología
Universidad de La Laguna

Gestión del Conocimiento en las Organizaciones:

Sistemas de recomendación

Iteración 02 - Modelos Basados en el Contenido.

Hugo Fernández Solís
(alu0101112664@ull.edu.es)

Índice

1. Objetivos

El objetivo de esta práctica es avanzar con los sistemas de recomendación y aprender el funcionamiento de nuevos modelos y sistemas dentro de los mismos. Así como familiarizarnos con ellos y sus utilidades.

2. Análisis realizado

El primer paso es analizar los ficheros que se nos presentan. En esta práctica nos centraremos en un tipo de archivos concretos en el que cada línea contiene un documento distinto. Y estos documentos serán los que analizaremos.

Lo segundo es analizar los pasos que debemos seguir. Primero debemos separar el texto en cada uno de los documentos, luego realizaremos un preprocesamiento de cada documento, esto es, quitar todas las palabras sin relevancia, signos de puntuación, etc... Acto seguido debemos calcular con cada documento su term frequency, su inverted document frequency y su tf+idf.

El term frequency es una medida que nos devuelve lo que se repite una palabra dentro de un documento.

El inverted term frequency nos devuelve también una proporción de lo que se repite cada término pero en el conjunto de los artículos.

$$IDF(x) = log \frac{N}{df_x}$$

El tf+idf es una medida estándar sobre las dos medidas anteriores.

$$w_{x,y} = tf_{x,y} \times \log\left(\frac{N}{df_x}\right)$$

Una vez que ya tenemos todas las medidas sobre cada uno de los términos del documento, vamos a proceder a calcular la similitud entre los mismos.

Para ello vamos a utilizar la distancia coseno respecto a tf+idf.

$$sim(\vec{a}, \vec{b}) = \frac{\sum_{u \in U}(r_{u,a} - \bar{r}_u)(r_{u,b} - \bar{r}_u)}{\sqrt{\sum_{u \in U}(r_{u,a} - \bar{r}_u)^2}\sqrt{\sum_{u \in U}(r_{u,b} - \bar{r}_u)^2}}$$

Esto nos devolverá un valor numérico que contiene la similitud entre dos documentos. Si repetimos este proceso con todos los documentos, obtendremos cuáles son los que más se parecen y los que menos.

3. Ejemplos

Como ejemplo vamos a analizar estos dos documentos:

Doc1:

```
This dry and restrained wine offers spice in profusion. Balanced with acidity and a firm texture, it's very much for food.
```

Doc2:

```
Savory dried thyme notes accent sunnier flavors of preserved peach in this brisk, off-dry wine. It's fruity and fresh, with an elegant, sprightly footprint.
```

Si eliminamos las palabras que no nos aportan información y los signos de puntuación, nos quedan así:

Doc1:

```
dry restrained wine offers spice profusion balanced acidity firm texture much food.
```

Doc2:

```
savory dried thyme notes accent sunnier flavors preserved peach brisk off-dry wine fruity fresh elegant sprightly footprint
```

Nótese que también se han sustituido las mayúsculas por minúsculas, pues para nosotros tiene el mismo valor la palabra *Banana* que *banana*.

El siguiente paso es realizar todas las transformaciones anteriores. Podemos observar algunos resultados.

---------- Article 8 ----------

	Term_freq	Doc_freq	tf + idf
dry	0.08	0.92	0.08
restrained	0.08	2.30	0.19
wine	0.08	0.69	0.06
offers	0.08	2.30	0.19
spice	0.08	1.61	0.13
profusion	0.08	2.30	0.19
balanced	0.08	1.20	0.10
acidity	0.08	0.36	0.03
firm	0.08	1.61	0.13
texture	0.08	1.61	0.13
much	0.08	1.61	0.13
food	0.08	2.30	0.19

---------- Article 9 ----------

	Term_freq	Doc_freq	tf + idf
savory	0.06	1.61	0.09
dried	0.06	1.20	0.07
thyme	0.06	2.30	0.14
notes	0.06	1.61	0.09
accent	0.06	2.30	0.14
sunnier	0.06	2.30	0.14
flavors	0.06	0.92	0.05
preserved	0.06	2.30	0.14
peach	0.06	2.30	0.14
brisk	0.06	1.61	0.09
dry	0.06	0.92	0.05
wine	0.06	0.69	0.04
fruity	0.06	1.61	0.09
fresh	0.06	0.92	0.05
elegant	0.06	2.30	0.14
sprightly	0.06	2.30	0.14
footprint	0.06	2.30	0.14

** Nótese que estas capturas están sacadas de un archivo con más documentos, por lo que el resultado puede variar.*

Una vez que ya tenemos toda la información sobre los documentos y sus términos, podemos pasar a comprobar las similitudes. para ello buscamos ambos tf+idf y calculamos la similitud.

Un resultado podría ser:

```
--------    Cosine relation    ---------

cos(A1, A2) = 0.15084593691843545
cos(A1, A3) = 0.15921455660831899
cos(A1, A4) = 0.1559649950254621
cos(A1, A5) = 0.15378756419511078
cos(A1, A6) = 0.12979527182433503
cos(A1, A7) = 0.15876110888643066
cos(A1, A8) = 0.14925627120171014
cos(A1, A9) = 0.17008645499628683
cos(A1, A10) = 0.154449631798639
```

Donde vemos la similitud entre el documento 1 y todos los demás.

4. Conclusiones

Este tipo de sistemas de recomendación son muy útiles cuando tenemos contenido muy extenso que no se puede cuantificar de primeras. Además nos ayudan a clasificar los archivos por su contenido y no por la valoración que hayan dado otros usuarios, lo que puede ser muy útil cuando tienes que valorar archivos que acaban de publicarse o que tienen muy pocas valoraciones.