UNIVERSIDAD ICESI
DIPLOMADO EN ANALÍTICA
Y GRANDES VOLÚMENES DE DATOS

Taller de sistemas de recomendación basado en contenido

Objetivo

El propósito de este taller es crear un modelo de recomendación basado en contenido paso por paso en Excel, para poder entender cada etapa del proceso. Trataremos con una base de 20 documentos (artículos de periódico), cuyo contenido es descrito con respecto a 10 categorías de noticias.

Cada documento es descrito entonces por un vector de 10 posiciones, utilizando un valor binario indicando si trata o no de la categoría correspondiente a cada dimensión del vector.

Tenemos los ratings dados por 2 usuarios, que han calificado positiva (+1) o negativamente (-1) los artículos leídos (columnas O y P). Cabe anotar que los usuarios no interactuaron con los 20 artículos, sino con 5, y no los mismos.

Parte 1: Creación de un modelo simple basado en contenido

Vamos a crear un modelo de recomendación basado en contenido con los datos tal y como los tenemos inicialmente.

1. ¿Qué tan fácil es encontrar artículos tratando de cada una de las categorías de noticias? Lo primero que tenemos que hacer es calcular el DF de cada atributo, la frecuencia (conteo) de cada término en los documentos (fila 23 de la hoja de Excel Orginal). La idea es analizar la influencia de cada dimensión en la recomendación.

2. ¿Cuáles son los gustos de cada usuario con respecto a cada tema? Vamos a crear el **perfil de cada usuario**. Para cada **categoría**, vamos a analizar los **documentos** que han sido notados por cada usuario:
   - Si el usuario dio una nota positiva y el documento trata la categoría en cuestión, esto es una indicación del gusto del usuario.
   - Si el usuario dio una nota negativa y el documento trata la categoría en cuestión, esto es una indicación de aversión del usuario
   - Si el usuario no ha dado una nota al documento, no podemos considerar ninguna evidencia de gusto o aversión
   - Para obtener las evidencias de gusto o aversión dadas por cada documento, basta con calcular el producto interno entre los vectores de la categoría y el vector de ratings del usuario. De esta manera tenemos entonces que llenar los perfiles de usuario que se encuentran en las filas 26 y 27 (utilicen la función SUMAPRODUCTO para facilitar este proceso)
   - ¿Cuáles son las categorías más acordes y desacordes con cada usuario según su perfil?

3. Vamos ahora a establecer qué tan parecidos son los perfiles de cada usuario con cada documento, para poder establecer cuáles serían los mejores documentos para recomendar y para evitar.
   - Vamos a aplicar en esta parte el producto interno como proxy de similitud entre los vectores del perfil de usuario y de descripción de cada documento. Completen las predicciones para cada usuario en las columnas S y T.
   - Realice un ranking de los documentos que se deben recomendar. Consideremos que un documento con una similitud nula o negativa no debe ser recomendado. Complete las casillas de los rankings “Top Inner”.

- En la columna M, calcule el largo de cada vector de descripción de los documentos (la raíz cuadrada de la suma de los cuadrados – distancia euclidiana) ¿Algunos de los documentos influencian más las recomendaciones que otros (cuáles)?

Parte 2: Modelo basado en contenido, normalizando

La idea ahora es aplicar una normalización de los vectores descriptores y de los perfiles de usuario, y analizar su efecto en la recomendación (lo que equivale a utilizar el coseno como medida de similitud). Nos interesan particularmente los cambios en los rankings de los documentos que se recomiendan que se realizan a los usuarios. Se debe ahora utilizar la pestaña de la hoja de cálculo "Norm", donde se deben completar los valores faltantes.

1. Normalizar la descripción de cada documento. Para esto es necesario dividir cada descriptor categórico de cada documento por la norma (largo del vector), calculado en la parte 1 del taller en la columna M de la pestaña "Original". Se debe ahora verificar que el largo de cada descriptor sea de valor 1, completando la columna M de la pestaña "Norm".
2. Realizar los pasos de la parte 1 (puntos 1 y 2), con los nuevos valores descriptores normalizados. Completar los valores de DF de la fila 23 y los perfiles de usuario de las filas 26 y 27. Calcular la norma de los vectores en las casillas M26 y M27
3. De la misma manera que se normalizaron las descripciones de los documentos, normalizar los vectores de los perfiles de los usuarios, completando las filas 29 y 30, verificando la norma de los vectores en las casillas M29 y M30.
4. Vamos a realizar las predicciones de la afinidad de cada usuario con cada documento a partir de la aplicación de la similitud dada por el coseno, rellenando las columnas S y T, utilizando entonces los perfiles normalizados.
5. Realice un ranking de los documentos que se deben recomendar. Consideremos que un documento con una similitud nula o negativa no debe ser recomendado. Complete las casillas de los rankings "Top Conseno".
6. Cómo convertirían uds esas predicciones en ratings de 1 a 5?

Parte 3: Modelo basado en contenido, utilizando TF-IDF

Hasta ahora solo hemos tenido en cuenta los términos de frecuencia absoluta y normalizada para describir los contenidos de los documentos en función de las categorías. Vamos entonces a aplicar la técnica de TF-IDF para tener en cuenta la pertinencia de las categorías que aparecen en menos documentos.

1. Complete la matriz de descripción de los documentos calculando el TF-IDF de cada termino-documento. Utilice los datos de la hoja original como valores de TF, y los DF originales. Aplique la fórmula TF-IDF = TF*Log(20/DF)
2. Una vez tenga la matriz de TF-IDF, calcule el largo de cada vector descriptivo de cada documento en la columna M. Como estos vectores no están normalizados, es necesario copiar sus valores en la columna N para poder normalizar los valores de TF-IDF previamente obtenidos. Hágalo, y compruebe que los vectores tienen norma de 1 en la columna M.
3. Obtenga los perfiles de usuario (filas 26 y 27), y normalícelos (filas 29 y 30)
4. Calcule las afinidades de los usuarios con los documentos en las columnas S y T.

5. Realice un ranking de los documentos que se deben recomendar. Consideremos que un documento con una similitud nula o negativa no debe ser recomendado. Complete las casillas de los rankings “Top TF-IDF”.
6. Analice las diferencias de los resultados de TF-IDF con los demás métodos, y trate de explicarlas teniendo en cuenta la rareza de las categorías y los documentos bien y mal notados por los usuarios.