Asignatura	
Técnicas de Inteligencia Artificial	Claudia Villalonga Palliser

Actividades resueltas

Recomendación basada en contenidos - VSM

Descripción de la actividad

Una base de datos de un periódico online contiene tres documentos que representan noticias y cada uno de ellos tiene el siguiente contenido[1]:
d_1: "Shipment of gold damaged in a fire"
d_2: "Delivery of silver arrived in a silver truck"
d_3: "Shipment of gold arrived in a truck"

Si Rosa está navegando por la web del periódico online y accede a la noticia del documento d_1 **¿qué otra noticia se le recomendaría?** Para resolver esta actividad utiliza una representación de los documentos basada en el espacio de vectores (VSM) como la que aplican los **recomendadores basados en contenidos** y calcula la matriz de similitudes indicando claramente los cálculos efectuados para medir la similitud entre contenidos.

Resolución de la actividad

Para poder determinar la noticia que un sistema de recomendación basado contenidos recomendaría a un Rosa cuando ésta accede al documento d_1, el primer paso es representar cada documento como un vector de pesos y el segundo calcular la similitud entre documentos a partir de la medida de similitud del coseno.

En esta actividad no se proporcionan metadatos que acompañen a los documentos y por lo tanto éstos se deben extraer del propio contenido del documento para poder definir el vector de pesos.

[1] Ejemplo de base de datos extraído del material de clase de los profesores David Grossman y Ophir Frieder (Illinois Institute of Technology) que son los autores del libro Grossman, David A. and Frieder, Ophir. *Information Retrieval: Algorithms and Heuristics*. Volume 15 of The Information Retrieval Series, Springer Science & Business Media, 2012.

Se va a utilizar la ponderación TF-IDF (Term Frequency-Inverse Document Frequency) para calcular los pesos del vector en base a los términos que aparecen en el documento.

El primer paso es extraer todos los términos que aparecen en la colección de tres documentos ($D = \{d_1, d_2, d_3\}$). Para la base de datos de documentos los términos serían: shipment, of, gold, damaged, in, a, fire, delivery, silver, arrived, truck. Por tanto tenemos un diccionario o conjunto de términos con 11 elementos ($T = \{t_1, t_2, \ldots, t_{11}\}$).

De hecho, en un sistema recomendador más eficiente se eliminarían todas las palabras que sean artículos y preposiciones, se utilizaría el lexema o infinitivo de los verbos y los nombres en singular. Esto nos permitiría quedarnos realmente con los términos realmente relevantes y evitar duplicidades. Sin embargo, para evitar realizar un análisis lingüístico en esta actividad vamos a crear los vectores utilizando los 11 términos en la forma en que aparecen en el documento.

Para calcular los pesos primero debemos calcular la frecuencia con que cada término (t_k) aparece en cada uno de los documentos ($d_j \in D = \{d_1, d_2, d_3\}$). La siguiente tabla recoge las frecuencias ($f_{k,j}$) o número de veces que cada término aparece en los documentos. Además, en la última columna de la tabla se presenta n_k, el número de documentos de la colección en el que el término t_k ocurre al menos una vez.

	$f_{k,1}$	$f_{k,2}$	$f_{k,3}$	n_k
a	1	1	1	3
arrived	0	1	1	2
damaged	1	0	0	1
delivery	0	1	0	1
fire	1	0	0	1
gold	1	0	1	2
in	1	1	1	3
of	1	1	1	3

shipment	1	0	1	2
silver	0	2	0	1
truck	0	1	1	2

A partir de n_k, número de documentos de la colección en el que el término t_k ocurre al menos una vez, y teniendo en cuenta que la colección contiene tres documentos ($N=3$) podemos calcular IDF (Inverse Document Frequency) con la siguiente fórmula:

$$IDF(t_k) = \log \frac{N}{n_k}$$

Los resultados del cálculo de $IDF(t_k)$ se presentan en la siguiente tabla:

	n_k	$IDF(t_k)$
a	3	0
arrived	2	0.176
damaged	1	0.477
delivery	1	0.477
fire	1	0.477
gold	2	0.176
in	3	0
of	3	0
shipment	2	0.176
silver	1	0.477
truck	2	0.176

A partir de las frecuencias ($f_{k,j}$) de los términos en los documentos calculamos TF (Term Frequancy) aplicando la siguiente fórmula:

$$TF(t_k, d_j) = \frac{f_{k,j}}{\max_z f_{z,j}}$$

Para calcular $TF(t_k, d_j)$ se computa el máximo de las frecuencias $f_{z,j}$ para todos los términos t_z que aparecen en el documento d_j. El máximo de las frecuencias de los términos en el documento d_1 es uno ($\max_z f_{z,1} = 1$), en el documento d_2 es dos ($\max_z f_{z,2} = 2$) y en el documento d_3 es uno ($\max_z f_{z,3} = 1$). Los resultados del cálculo de $TF(t_k, d_j)$ de acuerdo con la fórmula anterior se presentan en la siguiente tabla:

	$f_{k,1}$	$f_{k,2}$	$f_{k,3}$	$TF(t_k, d_1)$	$TF(t_k, d_2)$	$TF(t_k, d_3)$
a	1	1	1	1	0.5	1
arrived	0	1	1	0	0.5	1
damaged	1	0	0	1	0	0
delivery	0	1	0	0	0.5	0
fire	1	0	0	1	0	0
gold	1	0	1	1	0	1
in	1	1	1	1	0.5	1
of	1	1	1	1	0.5	1
shipment	1	0	1	1	0	1
silver	0	2	0	0	1	0
truck	0	1	1	0	0.5	1

Una vez tenemos TF (Term Frequancy) y IDF (Inverse Document Frequency) podemos calcular fácilmente TF-IDF (Term Frequancy - Inverse Document Frequency) como:

$$TF - IDF(t_k, d_j) = TF(t_k, d_j) \cdot IDF(t_k)$$

La métrica TF-IDF mide la relevancia de los términos en cada uno de los documentos de la colección y se basa en el principio de que aquellos términos que ocurren frecuentemente en un documento (TF) pero que ocurren rara vez en otros documentos (IDF), serán más relevantes en el tema del documento en concreto.

Los resultados del cálculo de *TF-IDF(t_k, d_j)* se presentan en la siguiente tabla:

	TF ***(t_k, d_1)***	***TF*** ***(t_k, d_2)***	***TF*** ***(t_k, d_3)***	***IDF(t_k)***	***TF-IDF*** ***(t_k, d_1)***	***TF-IDF*** ***(t_k, d_2)***	***TF-IDF*** ***(t_k, d_3)***
a	1	0.5	1	0	0	0	0
arrived	0	0.5	1	0.176	0	0.088	0.176
damaged	1	0	0	0.477	0.477	0	0
delivery	0	0.5	0	0.477	0	0.239	0
fire	1	0	0	0.477	0.477	0	0
gold	1	0	1	0.176	0.176	0	0.176
in	1	0.5	1	0	0	0	0
of	1	0.5	1	0	0	0	0
shipment	1	0	1	0.176	0.176	0	0.176
silver	0	1	0	0.477	0	0.477	0
truck	0	0.5	1	0.176	0	0.088	0.176

El último paso para calcular los pesos consiste en normalizar los valores calculados de TF-IDF para no favorecer los documentos más largos frente a los más cortos. Con la normalización vamos a obtener pesos en el intervalo [0, 1] y además el vector de pesos resultante a tener módulo 1. Entonces los pesos se calculan con la siguiente fórmula:

$$w_{k,j} = \frac{TF - IDF(t_k, d_j)}{\sqrt{\sum_{s=1}^{|T|}(TF - IDF(t_s, d_j))^2}}$$

El denominador de esta fórmula sería el factor de normalización y se debe calcular para cada documento ($d_j \in D = \{d_1, d_2, d_3\}$). Teniendo en cuenta que el diccionario está compuesto de 11 términos ($T = \{t_1, t_2, \ldots, t_{11}\}$) y por lo tanto |T|=11 podemos calcular el factor de normalización para el documento d_1 de la siguiente manera:

$$\sqrt{\sum_{s=1}^{|T|}(TF - IDF(t_s, d_1))^2}$$
$$= \sqrt{(TF - IDF(t_1, d_1))^2 + (TF - IDF(t_2, d_1))^2 + \cdots (TF - IDF(t_{11}, d_1))^2}$$
$$= \sqrt{0^2 + 0^2 + 0.4771^2 + 0^2 + 0.4771^2 + 0.1761^2 + 0^2 + 0^2 + 0.1761^2 + 0^2 + 0^2}$$
$$= 0.7192$$

El factor de normalización para el documento d_1 sería 0.7192, para d_2 sería 0.5478 y para d_3 sería 0.3522 y con estos valores podemos aplicar la fórmula descrita antes para calcular los pesos $w_{k,j}$. La siguiente tabla recoge el valor de los pesos:

	TF-IDF (t_k, d_1)	***TF-IDF*** (t_k, d_2)	***TF-IDF*** (t_k, d_3)	$w_{k,1}$	$w_{k,2}$	$w_{k,3}$
a	0	0	0	0	0	0
arrived	0	0.088	0.176	0	0.161	0.500
damaged	0.477	0	0	0.663	0	0
delivery	0	0.239	0	0	0.436	0
fire	0.477	0	0	0.663	0	0

gold	0.176	0	0.176	0.245	0	0.500
in	0	0	0	0	0	0
of	0	0	0	0	0	0
shipment	0.176	0	0.176	0.245	0	0.500
silver	0	0.477	0	0	0.871	0
truck	0	0.088	0.176	0	0.161	0.500

Nota: En los ejercicios de examen nos van a proporcionar directamente los vectores de pesos ($w_{k,j}$) que representan cada uno de los documentos, tal como se muestra en la tabla presentada a continuación. Por lo tanto, en ese caso no sería necesario realizar todos los cálculos descritos anteriormente y podríamos partir directamente de esos vectores para calcular la similitud entre contenidos.

$w_{k,1}$	$w_{k,2}$	$w_{k,3}$
0	0	0
0	0.161	0.500
0.663	0	0
0	0.436	0
0.663	0	0
0.245	0	0.500
0	0	0
0	0	0
0.245	0	0.500
0	0.871	0
0	0.161	0.500

Partiendo de que $\vec{w}_i = (w_{1,i}\ w_{2,i}\ \dots\ w_{T,i}\ )$ es el vector de pesos que representa el contenido del documento d_i, podemos calcular la similitud entre dos contenidos d_i y d_j a partir de la medida de similitud del coseno del ángulo formado por los vectores $\vec{w}_i$ y $\vec{w}_j$. Estos vectores de pesos tienen dimensión T (11 en este caso), valor que se corresponde con el número de términos que definen el vocabulario que se ha utilizado para extraer los metadatos a partir del propio texto del documento.

La similitud de los contenidos d_i y d_j se puede calcular con la siguiente fórmula de la similitud del coseno:

$$similitud\ (d_i, d_j) = cos(\vec{w}_i, \vec{w}_j) = \frac{\vec{w}_i \cdot \vec{w}_j}{\|\vec{w}_i\| \cdot \|\vec{w}_j\|} = \frac{\sum_{k=1}^{T} w_{k,i} \cdot w_{k,j}}{\sqrt{\sum_{k=1}^{T} {w_{k,i}}^2} \cdot \sqrt{\sum_{k=1}^{T} {w_{k,j}}^2}}$$

Para esta actividad en concreto, hemos normalizado los vectores de pesos y por lo tanto los módulos de los todos los vectores serán iguales a uno ($\|\vec{w}_i\| = \|\vec{w}_j\| = 1$). Entonces se van a simplificar los cálculos de la similitud del coseno. Si no se hubieran normalizado los vectores de pesos deberíamos calcular los módulos de los vectores.

Los valores de similitud nos sirven para crear la correspondiente matriz de similitud entre contenidos. La matriz de similitud tiene una fila y una columna para cada uno de los contenidos, es decir para cada uno de los documentos disponibles en la base de datos. Por tanto, es una matriz cuadrada de dimensiones kxk, donde k es el número de documentos. En este caso la matriz de similitud tiene unas dimensiones de 3x3.

Cada elemento de la matriz de similitud es el valor de la similitud entre el contenido representado en la correspondiente fila y el contenido representado en la correspondiente columna. La matriz de similitud es simétrica y en su diagonal tiene como valores unos porque la similitud de un contenido consigo mismo es máxima e igual a 1 en base a la fórmula anterior.

Para responder a la pregunta formulada nos va a interesar calcular la similitud del contenido d_1 (la noticia que está consultando Rosa) con respecto al resto de contenidos (d_2 y d_3). Si embargo en el enunciado se nos indica que proporcionemos la matriz de similitudes completa por lo que vamos a calcular todos los valores de esta.

Realizando los cálculos, la similitud entre el contenido d_1 representado por el vector (0, 0, 0.663, 0, 0.663, 0.245, 0, 0, 0.245, 0, 0) y el contenido d_2 representado por el vector (0, 0.161, 0, 0.436, 0, 0, 0, 0, 0, 0.871, 0. 161) es:

$$similitud(d_1, d_2) = \frac{(0\ 0\ 0.663\ 0\ 0.663\ 0.245\ 0\ 0\ 0.245\ 0\ 0) \cdot (0\ 0.161\ 0\ 0.436\ 0\ 0\ 0\ 0\ 0\ 0.871\ 0.161)}{\|(0\ 0\ 0.663\ 0\ 0.663\ 0.245\ 0\ 0\ 0.245\ 0\ 0)\| \cdot \|(0\ 0.161\ 0\ 0.436\ 0\ 0\ 0\ 0\ 0\ 0.871\ 0.161)\|} = \frac{0}{1 \cdot 1} = 0$$

La similitud entre el contenido d_1 representado por el vector (0, 0, 0.663, 0, 0.663, 0.245, 0, 0, 0.245, 0, 0) y el contenido d_3 representado por el vector (0, 0.500, 0, 0, 0, 0.500, 0, 0, 0.500, 0, 0.500) es:

$$similitud(d_1, d_3) = \frac{(0\ 0\ 0.663\ 0\ 0.663\ 0.245\ 0\ 0\ 0.245\ 0\ 0) \cdot (0\ 0.500\ 0\ 0\ 0\ 0.500\ 0\ 0\ 0.500\ 0\ 0.500)}{\|(0\ 0\ 0.663\ 0\ 0.663\ 0.245\ 0\ 0\ 0.245\ 0\ 0)\| \cdot \|(0\ 0.500\ 0\ 0\ 0\ 0.500\ 0\ 0\ 0.500\ 0\ 0.500)\|} = \frac{0.245 \cdot 0.500 + 0.245 \cdot 0.500}{1 \cdot 1} = 0.245$$

La similitud entre el contenido d_2 representado por el vector (0, 0.161, 0, 0.436, 0, 0, 0, 0, 0, 0.871, 0. 161) y el contenido d_3 representado por el vector (0, 0.500, 0, 0, 0, 0.500, 0, 0, 0.500, 0, 0.500)

$$similitud(d_1, d_3) = \frac{(0\ 0.161\ 0\ 0.436\ 0\ 0\ 0\ 0\ 0\ 0.871\ 0.161) \cdot (0\ 0.500\ 0\ 0\ 0\ 0.500\ 0\ 0\ 0.500\ 0\ 0.500)}{\|(0\ 0.161\ 0\ 0.436\ 0\ 0\ 0\ 0\ 0\ 0.871\ 0.161)\| \cdot \|(0\ 0.500\ 0\ 0\ 0\ 0.500\ 0\ 0\ 0.500\ 0\ 0.500)\|} = \frac{0.161 \cdot 0.500 + 0.161 \cdot 0.500}{1 \cdot 1} = 0.161$$

Habiendo calculado todos los valores de similitud se puede construir la matriz de similitud que quedaría tal como se muestra a continuación:

	d_1	d_2	d_3
d_1	1	0	0.245
d_2	0	1	0.161
d_3	0.245	0.161	1

Una vez se ha calculado la matriz de similitudes y observando las similitudes entre el contenido d_1 y los otros dos contenidos se puede contestar a la pregunta de esta actividad. Si Rosa está navegando por la web del periódico online y accede a la noticia del documento d_1 se le recomendaría el contenido d_3 porque el contenido d_1 tiene mayor similitud con el contenido d_3 que con el contenido d_2. Concretamente la similitud del coseno entre el vector que modela el documento d_1 y el vector que modela el documento d_3 es de 0.245.