

Asignatura	
Técnicas de Inteligencia Artificial	Claudia Villalonga Palliser

Actividades resueltas

Recomendación basada en contenidos - VSM

Descripción de la actividad

Un buscador de restaurantes tiene la siguiente base de datos que contiene tres registros con información de restaurantes.

	Nombre	Cocina	Reservas	Precio
R1	Pizzería Roma	Italiana	No	Bajo
R2	La Parrilla Argentina	Argentina	Teléfono	Medio
R3	Bistro de Jacques	Francesa	Online y Teléfono	Alto

Rosa, una usuaria del buscador de restaurantes, cuando se registró en el servicio rellenó su perfil de usuario y definió que está interesada en restaurantes con un precio medio o bajo, que no le interesan los restaurantes donde no se puede hacer reserva y que si se puede efectuar reserva prefiere hacerla online. Además de esa información explícita que Rosa introdujo a través de un formulario, el sistema recomendador ha estado analizando las valoraciones que Rosa ha hecho en la plataforma de los restaurantes que ha visitado. Rosa ha valorado positivamente (le ha dado al botón like) a 3 de los 4 restaurantes italianos que ha visitado y al único restaurante de comida francesa que en el que ha estado hasta ese momento.

¿Qué restaurante se le recomendaría a Rosa? Para resolver esta actividad representa los contenidos (información sobre los restaurantes) y el perfil del usuario en espacio de vectores (VSM) como haría un **recomendador basado en contenido** y calcula la similitud del perfil del usuario con cada uno de los contenidos. Indica claramente los cálculos efectuados para medir las similitudes y responde a la pregunta planteada sobre que restaurante se le recomendaría a Rosa.

Asignatura	
Técnicas de Inteligencia Artificial	Claudia Villalonga Palliser

Resolución de la actividad

Para poder determinar el restaurante que un sistema de recomendación basado contenidos recomendaría a Rosa, el primer paso es representar cada restaurante, que sería el propio contenido, como un vector en el espacio de vectores (VSM), el segundo paso es crear el perfil del usuario Rosa y representarlo también como un vector y por último se tiene que calcular la similitud entre el perfil de usuario y los diferentes contenidos a partir de la medida de similitud del coseno.

Cada uno de los restaurantes se presenta como un registro en la base de datos. Por lo tanto, tenemos información estructurada sobre los metadatos que acompañan y definen los contenidos. Es a partir de estos metadatos que vamos a crear los vectores en el espacio de vectores (VSM).

Existen diversas opciones para la creación de los vectores que modelen los contenidos. Se podría crear un vector con tantas dimensiones como atributos tengan los registros en la base de datos y luego utilizar valores nominales para cada una de las variables. En este caso el contenido se modelaría con un vector del tipo (*cocina, reservas, precio*) donde *cocina*=*{italiana, argentina, francesa}*, *reservas*=*{no, online, teléfono}* y *precio*=*{bajo, medio, alto}*. Sin embargo, en la base de datos tenemos que el segundo restaurante acepta dos tipos de reservas y eso crearía problemas a la hora de definir el atributo *Reservas* de tipo nominal y modelar el vector tal como se ha indicado antes. Es por eso que vamos a optar por una segunda opción en la que creamos vectores binarios con tantas dimensiones como pares de atributos-valores. Entonces el vector que representaría cada contenido (restaurante de la base de datos) tendría la forma (*italiana, argentina, francesa, no_reserva, online, teléfono, bajo, medio, alto*) donde cada elemento sólo puede tomar los valores 0 o 1. Cabe destacar que hemos obviado el nombre del restaurante, aunque éste se podría analizar cómo se hace en el caso que los contenidos son documentos no estructurados. Si aplicáramos procesamiento del lenguaje natural y analizáramos el texto en el nombre de los restaurantes podríamos detectar algunos términos relevantes, por ejemplo, pizzería, parrillada y bistro. Entonces estos términos también se podrían añadir como nuevas dimensiones al vector que define el contenido.

Asignatura	
Técnicas de Inteligencia Artificial	Claudia Villalonga Palliser

Utilizando el modelo de los vectores que hemos descrito anteriormente, los contenidos en la base de datos se representarían en el espacio de vectores (VSM) tal como se indica en la siguiente tabla.

	R1	R2	R3
<i>italiana</i>	1	0	0
<i>argentina</i>	0	1	0
<i>francesa</i>	0	0	1
<i>no_reserva</i>	1	0	0
<i>online</i>	0	0	1
<i>teléfono</i>	0	1	1
<i>bajo</i>	1	0	0
<i>medio</i>	0	1	0
<i>alto</i>	0	0	1

El siguiente paso es crear el perfil del usuario Rosa y representarlo también como un vector en el mismo espacio de vectores (VSM).

En el enunciado se indica que Rosa definió que está interesada en restaurantes con un precio medio o bajo por lo tanto los elementos *bajo* y *medio* del vector que representa el perfil de Rosa tienen un valor 1 y el elemento *alto* tiene un valor 0. Además, se indica que a Rosa no le interesan los restaurantes donde no se puede hacer reserva, por tanto, el elemento *no_reserva* del vector va a tener un valor 0. También sabemos que si se puede efectuar reserva en un restaurante, Rosa prefiere hacerla online. En este caso como dice “prefiere” no vamos a utilizar un valor binario, sino que vamos a darle un peso entre 0 y 1 a los elementos *online* y *teléfono* del vector que representa el perfil de Rosa. El enunciado no nos da información para asignar esos pesos así que vamos a tomar unos valores que nos parezcan razonables (aunque cabe notar que es un ejemplo y no una verdad absoluta). Asignamos un valor de 0.25 para el caso de hacer la reserva

Asignatura	
Técnicas de Inteligencia Artificial	Claudia Villalonga Palliser

por teléfono y 0.75 para el caso de la reserva online para así dar más prioridad al hecho que prefiere Rosa, hacer la reserva online, sin descartar totalmente que pueda reservar por teléfono.

Para completar el perfil de usuario, concretamente los elementos del vector *italiana*, *argentina* y *francesa*, utilizamos la información que el sistema recomendador ha recogido de las valoraciones que ha hecho Rosa de los restaurantes que ha visitado. Sabemos que Rosa ha valorado positivamente a 3 de los 4 restaurantes italianos que ha visitado y le ha dado al botón like del único restaurante de comida francesa en el que ha estado. Entonces el sistema recomendador es capaz de estimar que existe una probabilidad del 60% de que a Rosa le guste la comida italiana porque le han gustado 3 restaurantes italianos del total de los 5 restaurantes que ha visitado. Por lo tanto, el elemento *italiana* del vector que representa el perfil de usuario tendrá un valor 0.60. De la misma forma se puede predecir que existe una probabilidad del 20% de que a Rosa le guste la comida francesa y entonces el elemento *francesa* del vector tendrá un valor 0.20. Por último, sabemos que Rosa no ha valorado ningún restaurante de comida argentina y por ello el recomendador puede asumir que Rosa no ha estado en ningún restaurante argentino porque no le interesa ese tipo de comida. Entonces el elemento *argentina* del vector que representa el perfil de Rosa tendrá un valor 0.

La siguiente tabla recoge la representación del perfil de usuario de Rosa y los contenidos en el espacio de vectores (VSM).

	R1	R2	R3	Rosa
<i>italiana</i>	1	0	0	0.60
<i>argentina</i>	0	1	0	0
<i>francesa</i>	0	0	1	0.20
<i>no_reserva</i>	1	0	0	0
<i>online</i>	0	0	1	0.75
<i>teléfono</i>	0	1	1	0.25

Asignatura	
Técnicas de Inteligencia Artificial	Claudia Villalonga Palliser

<i>bajo</i>	1	0	0	1
<i>medio</i>	0	1	0	1
<i>alto</i>	0	0	1	0

Ahora debemos calcular la similitud entre el perfil de usuario de Rosa (p) y los diferentes contenidos (c_i) o sea los restaurantes a partir de la medida de similitud del coseno que se calcula con la siguiente fórmula:

$$\text{similitud}(p, c_i) = \frac{\vec{p} \cdot \vec{c}_i}{\|\vec{p}\| \cdot \|\vec{c}_i\|} = \frac{\sum_{k=1}^T p_k \cdot c_{k,i}}{\sqrt{\sum_{k=1}^T p_k^2} \cdot \sqrt{\sum_{k=1}^T c_{k,i}^2}}$$

donde $\vec{p} = (p_1 \ p_2 \ \dots \ p_T)$ es el vector de dimensión T (9 en este caso) que representa el perfil de usuario de Rosa y $\vec{c}_i = (c_{1,i} \ c_{2,i} \ \dots \ c_{11,i})$ son los vectores de dimensión T que representan los contenidos (los tres restaurantes en la base de datos).

Realizando los cálculos, la similitud entre el perfil de usuario de Rosa p representado por el vector (0.60, 0, 0.20, 0, 0.75, 0.25, 1, 1, 0) y el contenido c_1 representado por el vector (1, 0, 0, 1, 0, 0, 1, 0, 0) es:

$$\begin{aligned} \text{similitud}(p, c_1) &= \frac{(0.60 \ 0 \ 0.20 \ 0 \ 0.75 \ 0.25 \ 1 \ 1 \ 0) \cdot (1 \ 0 \ 0 \ 1 \ 0 \ 0 \ 1 \ 0 \ 0)}{\|(0.60 \ 0 \ 0.20 \ 0 \ 0.75 \ 0.25 \ 1 \ 1 \ 0)\| \cdot \|(1 \ 0 \ 0 \ 1 \ 0 \ 0 \ 1 \ 0 \ 0)\|} \\ &= \frac{0.60 + 1}{\sqrt{0.60^2 + 0.20^2 + 0.75^2 + 0.25^2 + 1^2 + 1^2} \cdot \sqrt{3}} = \frac{1.60}{\sqrt{3.025} \cdot \sqrt{3}} = 0.531 \end{aligned}$$

La similitud entre el perfil de usuario de Rosa p y el contenido c_2 representado por el vector (0, 1, 0, 0, 0, 1, 0, 1, 0) es:

$$\begin{aligned} \text{similitud}(p, c_2) &= \frac{(0.60 \ 0 \ 0.20 \ 0 \ 0.75 \ 0.25 \ 1 \ 1 \ 0) \cdot (0 \ 1 \ 0 \ 0 \ 0 \ 1 \ 0 \ 1 \ 0)}{\|(0.60 \ 0 \ 0.20 \ 0 \ 0.75 \ 0.25 \ 1 \ 1 \ 0)\| \cdot \|(0 \ 1 \ 0 \ 0 \ 0 \ 1 \ 0 \ 1 \ 0)\|} = \frac{0.25 + 1}{\sqrt{3.025} \cdot \sqrt{3}} \\ &= \frac{1.25}{\sqrt{3.025} \cdot \sqrt{3}} = 0.415 \end{aligned}$$

Asignatura	
Técnicas de Inteligencia Artificial	Claudia Villalonga Palliser

La similitud entre el perfil de usuario de Rosa p y el contenido c_3 representado por el vector (0, 0, 1, 0, 1, 1, 0, 0, 1) es:

$$\begin{aligned}
 similitud(p, c_3) &= \frac{(0.60 \ 0 \ 0.20 \ 0 \ 0.75 \ 0.25 \ 1 \ 1 \ 0) \cdot (0 \ 0 \ 1 \ 0 \ 1 \ 1 \ 0 \ 0 \ 1)}{\|(0.60 \ 0 \ 0.20 \ 0 \ 0.75 \ 0.25 \ 1 \ 1 \ 0)\| \cdot \|(0 \ 0 \ 1 \ 0 \ 1 \ 1 \ 0 \ 0 \ 1)\|} \\
 &= \frac{0.20 + 0.75 + 0.25}{\sqrt{3.025} \cdot \sqrt{4}} = \frac{1.20}{\sqrt{3.025} \cdot \sqrt{4}} = 0.345
 \end{aligned}$$

La siguiente tabla recoge los valores de similitud entre el perfil de usuario de Rosa (p) y los diferentes contenidos (c_i) sobre los restaurantes que se le pueden recomendar:

	c_1	c_2	c_3
p	0.531	0.415	0.345

Observando los valores de la similitud se puede contestar a la pregunta de esta actividad. El servicio de búsqueda de restaurantes le recomendaría a Rosa la Pizzería Roma porque el contenido c_1 es el que tiene mayor similitud (0.531) con el perfil de usuario de Rosa.