Pontificia Universidad Católica de Chile
Departamento de Ciencia de la Computación
IIC3633 - Sistemas Recomendadores

Propuesta Proyecto

30 de septiembre de 2019

2º semestre 2019 - Profesor: Denis Parra

Ariel Martínez y Jerónimo Salazar - **pyRecabros**

Contexto

Una tarea muy común realizada por sistemas recomendadores basados en contenido consiste en encontrar ítems similares comparando sus características, ya sea de manera directa, o bien comparando algún vector de dimensiones latentes que represente sus principales características (normalmente obtenido como resultado de una red neuronal). Es posible aplicar este tipo de enfoque a la recomendación de documentos de texto, utilizando un conjunto de palabras clave (*keywords*), que resume el contenido de dicho documento, como elemento principal del algoritmo recomendador.

En este contexto, uno de los algoritmos más populares en el análisis de datos no estructurados, como documentos de texto, es aquel que encuentra el *Text Frequency - Inverse Document Frequency* (TF-IDF) de un *token* (e.g. palabra), valor numérico que expresa cuán relevante es dicho *token* para un documento en particular, dado un *corpus* de documentos del mismo tipo (Rajaraman & Ullman, 2011).

Problema - Ted Talks

Nosotros estamos interesados en aplicar distintas variaciones de algoritmos que computen el TF-IDF con el objetivo de recomendar ítems, cuyo contenido es representado como un texto en lenguaje natural. Las variaciones se lograrán alternando los parámetros que reciben estos algoritmos y las distintas funciones de peso que se utilizan para encontrar los distintos valores, para finalmente evaluar los resultados a través de métricas de *ranking* conocidas. Además, nos interesa proponer nuevas maneras de calcular el TF-IDF y reportar los resultados que estas produzcan.

El *dataset* que utilizaremos se encuentra en Kaggle y contiene información sobre todas las charlas TED subidas a su página web oficial (ted.com) hasta el 21 de septiembre de 2017

obtenidas a través de técnicas de *web scraping*. Esto incluye la transcripción a texto plano de cada una de ellas (Banik, 2017). El esquema de esta base de datos se muestra en la figura de a continuación:

```
ted_main.csv:
{
    comments: int
    description: str
    duration: int
    event: str
    film_date: int
    languages: int
    main_speaker: str
    name: str
    num_speaker: int
    published_date: int
    ratings: json
    related_talks: json
    speaker_occupation: str
    tags: json
    title: str
    url: href (PK)
    views: int
}

transcripts.csv:
{
    transcript: str
    url: href (PK)
}
```

Objetivos

Nuestro objetivo para este proyecto es comparar las distintas técnicas y algoritmos que pueden ser ocupados en una recomendación basada en contenido, mediante TF-IDF. Para esto ocuparemos el *dataset* de TED Talks, comparando las recomendaciones obtenidas con los elementos entregados en *related_talks*.

Solución y experimentos a realizar

En primer lugar, podemos comparar las distintas formulas ocupadas para el calculo de pesos en la frecuencia de frases, la frecuencia de documentos y normalizaciones, como puede ser

visto en la Tabla 1.

Term frecuency	Document frecuency	Normalization
natural $tf_{t,d}$	number 1	none
logarithm $1 + log(tf_{t,d})$	idf $log(\frac{N}{df_f})$	cosine
augmented $0{,}5 + \frac{0{,}5 * tf_{t,d}}{max_t(tf_{t,d})}$	prob idf $max(0, log(\frac{N-df_t}{df_f}))$	pivoted unique
boolean 1 if $tf_{t,d}$>0, else 0		byte size

Tabla 1: variaciones en parámetros de TF-IDF

En segundo lugar, es posible agregar preprocesamientos a los documentos. Los métodos que ocuparemos para probar su eficiencia son *stemming* y *lemmatization*. Ambos procesos, como explican Manning *et al.* (2008), aplican distintas reglas con el objetivo de eliminar la influencia de variaciones o formas verbales en las palabras que serán ocupadas en la recomendación. *Stemming* ocupa un conjunto de reglas para reducir una palabra de sufijos o prefijos (por ejemplo: cars a car, easily a eas). Por otra parte, *lemmatization* corresponde a transformar palabras de acuerdo a un análisis morfológico, es decir, transforma palabras a su concepto elemental, como también las conjugaciones verbales a su forma en infinitivo. Este último proceso puede llevarse a cabo mediante el uso de un diccionario.

Referencias

1. Banik R. (2017). *TED Talks*. Data about TED Talks on the TED.com website until September 21st, 2017. Kaggle. Disponible en: kaggle.com/rounakbanik/ted-talks

2. Manning, C., Raghavan, P. Schütze, H. (2008). *Introduction to Information Retrieval* (pp. 32-34) Cambrigde University Press. ISBN: 0521865719.

3. Rajaraman, A. & Ullman, J. (2011). *Data Mining*. In Mining of Massive Datasets (pp. 1-17). Cambridge University Press. doi: 10.1017/CBO9781139058452.002.