Universidad
Internacional
Menéndez Pelayo

Máster Universitario en Investigación en Inteligencia Artificial

Curso 2020-2021

Recuperación y extracción de información, grafos y redes sociales

Práctica Bloque II: Recuperación de información y minería de texto

29 de marzo de 2021

Laura Rodríguez Navas
DNI: 43630508Z
e-mail: rodrigueznavas@posgrado.uimp.es

Índice

1. Resumen

2. Rastreador web (crawler)

Informe práctica de REIN de la UPC.

spider books_toscrape.py items settings DOWNLOAD_DELAY = 3

resultado en books.json

3. K-means

K-Means es un algoritmo no supervisado de Clustering. Se utiliza cuando tenemos un montón de datos sin etiquetar. El objetivo de este algoritmo es el de encontrar "K" grupos (clústers) entre los datos.

El algoritmo trabaja iterativamente para asignar a cada "punto" (las filas de nuestro conjunto de entrada forman una coordenada) uno de los "K" grupos basado en sus características. Son agrupados en base a la similitud de sus features (las columnas). Como resultado de ejecutar el algoritmo tendremos:

- Los "centroids" de cada grupo que serán unas "coordenadas" de cada uno de los K conjuntos que se utilizarán para poder etiquetar nuevas muestras.
- Etiquetas para el conjunto de datos de entrenamiento. Cada etiqueta perteneciente a uno de los K grupos formados.

Los grupos se van definiendo de manera "orgánica", es decir que se va ajustando su posición en cada iteración del proceso, hasta que converge el algoritmo. Una vez hallados los "centroids" deberemos analizarlos para ver cuales son sus características únicas, frente a la de los otros grupos. Estos grupos son las etiquetas que genera el algoritmo.

Casos de Uso de K-Means

El algoritmo de Clustering K-means es uno de los más usados para encontrar grupos ocultos, o sospechados en teoría sobre un conjunto de datos no etiquetado. Esto puede servir para confirmar -o desterrar- alguna teoría que teníamos asumida de nuestros datos. Y también puede ayudarnos a descubrir relaciones asombrosas entre conjuntos de datos, que de manera manual, no hubiéramos reconocido. Una vez que el algoritmo ha ejecutado y obtenido las etiquetas, será fácil clasificar nuevos valores o muestras entre los grupos obtenidos.

Nuestro caso de uso es -> Categorización de Inventario: agrupar los libros por categorías.

3.1. Datos de Entrada para K-Means

Las "features" o características que utilizaremos como entradas para aplicar el algoritmo k-means deberán ser de valores numéricos, continuos en lo posible. En caso de valores categóricos (por ej. Ciencia Ficción, Terror,etc) se puede intentar pasarlo a valor numérico.

Categorización de Ciencia Ficción, Terror...

4. HOLA

El conjunto de datos contiene diferenciadas 50 categorías - temáticas de libros. Pero alguna de las categorías solo aparece una vez en el conjunto de datos. Así pues, no se considera para el uso del clustering ya que no se podrán formar un grupo de más de un libro. Se eliminan del conjunto de datos.

Las temáticas a eliminar son: ["Academic", "Adult Fiction", "Crime", "Cultural", "Erotica", "Novels", "Paranormal", "Parenting", "Short Stories", "Suspense"]

Una vez eliminadas tenemos

Bibliografía