CURSO: MINERÍA DE DATOS

MAESTRÍA EN EXPLOTACIÓN DE DATOS Y DESCUBRIMIENTO DE CONOCIMIENTO

Locality Sensitive Hashing

Introducción:

Esta práctica de laboratorio tiene como objetivo abordar la técnica de Locality Sensitive Hashing para encontrar documentos similares o parecidos.

Para la exploración de estos temas, se utilizará el IDE R-Studio del lenguaje de programación R, a efectos de ejercitar los conceptos abordados en las clases teóricas.

CONSIGNAS

A partir de un subconjunto de letras de canciones que figuraron en los Spotify Charts (link de descarga), se solicita trabajar sobre las siguientes consignas:

1. **SIMILITUD DE JACCARD ENTRE PARES**
    a. Cargue los datos provistos por la cátedra y explore si existen duplicados en base al *track_name*
    b. Genere una nueva variable que contenga la cantidad de caracteres incluídos en el campo *Lyrics*
    c. Con el comando View, observe los datos ordenados por esta nueva variable y el nombre del track. Que se observa ? Qué ocurre con temas como *Aftertaste* o *All The Lovers*?
    d. Calcule la similitud de Jaccard sobre sus *lyrics* (4 registros, 6 comparaciones).

2. **LOCALITY SENSITIVE HASHING**
    a. Configure la función de minHash utilizando un n de 200 y un seed de 318
    b. Genere un corpus con la función TextReuseCorpus donde haga referencia a la función de minhasing recíen creada. Utilice el método de tokenize_ngrams, donde el n sea de 3. Cúantas documentos fueron capturados en el Corpus? Qué contienen los *warnings*?

c. Con la cantidad de permutaciones que configuramos en la función de MinHash y utilizando 20 Bandas, cuál es la probabilidad de que para 2 documentos con una similitud de Jaccard de 70% todas las filas de al menos una banda produzcan el mismo valor de MinHash?
d. Qué ocurre con la probabilidad si se incrementa a 40 bandas?
e. Utilizando 40 bandas, genere los buckets, los documentos candidatos y calcule su score. Existen documentos iguales? Y parecidos?
f. Proponga una posible solución al problema que plantean casos como en 2.c