Hashing y LSH 2020-2C

Contamos con una colección de 1300 millones de tweets con fake news. Queremos usarlos para detectar otros tweets que puedan contener mensajes parecidos para poder filtrarlos.

El objetivo final de nuestra aplicación es dado un tweet buscar cuáles son tweets parecidos en nuestra base de 1300 millones de tweets usando LSH para luego decidir si el tweet es o no un fake news. Si hay tweets muy similares al nuevo en nuestra base de datos, lo categorizamos como fake news; caso contrario, no. El criterio que fijamos es que si el tweet actual tiene una semejanza mayor o igual a 0.73 con alguno de nuestra base de fake news entonces es un fake news.

Se decide usar la distancia de Jaccard como métrica para la construcción de nuestro esquema de LSH.

Queremos que si la **semejanza** entre tweets es mayor o igual a **0.55** tengamos más de un **70%** de probabilidad de que sean candidatos. Por otro lado, si la semejanza es menor o igual a **0.05** queremos tener menos de un **1%** de probabilidad de que sean candidatos a ser comparados.

En base a esta información le pedimos que responda las siguientes preguntas:

1. ¿Cuántos minhash hacen falta y con qué tipo de esquema (b y r) ? Detalle los cálculos realizados y estime **cantidad** de falsos positivos y falsos negativos que vamos a tener sobre una base de 10000 tweets nuevos. (30 puntos)
2. Describa la etapa de pre-procesamiento en la cual tiene que recorrer los 1300 millones de tweets y crear una única tabla de hash. Recuerde que puede usar diagramas, esquemas y pseudocódigo para hacer más clara su explicación. Debe ser claro indicando cómo queda conformado el esquema LSH que va a usar en el punto 3. (35 puntos)
3. Describir cómo se hace la predicción de si un tweet es fake news o no utilizando el esquema LSH propuesto. (35 puntos)