Pontificia Universidad Católica de Chile
Escuela de Ingeniería
Departamento de Ciencia de la Computación
Primer Semestre de 2024

IIC 2440 – Procesamiento de Datos Masivos
Actividad BD Vectores

Instrucciones.

Tu tarea va a ser programar un compresor de frases: dada una frase en algun lenguaje, entregar una palabra que sea lo mas parecido a una frase.

La base de tu funcion es la siguiente. Tienes una familia de archivos, que contienen embeddings de palabras. Cada archivo usa vectores mas grandes (con más dimensiones), recomendamos usar la descripcion mas grande siempre que entregue tiempos razonables.

Entonces, dada una frase con palabras $p_1, \ldots, p_n$ (puedes asumir que tu frase solo usa palabras, nada de puntuacion ni similares) definimos el embedding de la frase como la media aritmetica de los vectores $v_1, \ldots, v_n$ correspondientes a cada una de esas palabras. Entonces, el resultado de tu funcion debe recibir una frase y entregar tres palabras cuyo embedding sea suficientemente similar al de la frase, considerando distancia de coseno.

La distancia de coseno (mal llamada distancia, por que no es una distancia realmente, no satisface la desigualdad triangular), se computa simplemente como

$$1 - \frac{\bar{u} \cdot \bar{v}}{\|\bar{u}\|\|\bar{v}\|}$$

Entrega. Esta tarea es individual. Deben subir el notebook al buzón de canvas. La fecha de entrega es el Viernes 10 de Mayo, a las 20:00 hrs. El buzón estará habilitado a partir del jueves.

Preguna 1.

Debes entregar un notebook con dos funciones donde se siguen dos estrategias distintas para realizar esto.

Primero, una función `brute_force_compress(frase)` que entregue las tres palabras cuyo vector sea el más cercano a el vector que computaste de la frase, de acuerdo a la distancia de coseno.

Segundo, una función `approximate_compress(frase)` que extraiga las tres palabras más similares cuando ahora almacenas los vectores de palabras usando la base de datos de vectores Qdrant.

Preguna 2.

Construye 10 ejemplos, y una tabla que te indique el tiempo que se demoró cada función en entregar resultados para cada uno de estos ejemplos. Puedes fijar un maximo de tiempo a esperar (razonable) y omitir los numeros donde alguna función se demora más que ese tiempo.