TAREA 3 REDES NEURONALES AVANZADAS DE APRENDIZAJE PROFUNDO

Ing. Alejandro Ferreira Vergara
Departamento de Ingeniería Matemática
Universidad de La Frontera

27 de Julio de 2020
(Actividad en parejas)

Pregunta 1

Reúna, ya sea automáticamente o manualmente, un corpus de al menos 30 (máximo 50) documentos. Guárdelos en una base de datos, ordenadamente, y en formato csv o txt (debe adjuntar la base de datos a la resolución de la tarea). Utilizando la codificación bolsa de palabras binaria, mida que tan similares son los documentos y muestre los 2 documentos más similiares (si hay más de 2, también muéstrelos). Para esto, recuerde crear un léxico ad-hoc y utilizar un buen algoritmo de tokenización.

Pregunta 2

Con el mismo corpus de la pregunta anterior, vectorice los documentos con el método TF-IDF y realize una comparación de los documentos en cuanto a su similitud. Muestre los 2 documentos más similares (si hay más de 2, también muéstrelos). Nuevamente, recuerde crear un léxico ad-hoc y utilizar un buen algoritmo de tokenización.

Pregunta 3

Haga un informe en Jupyter Notebook en donde se presenten los resultados de las preguntas anteriores y los códigos con los cuales resolvió dichas preguntas. Además, el informe debe contener una descripción de los datos (tipo de datos, contexto de los datos, etc), una descripción de su método de tokenización y algunas conclusiones acerca de los resultados obtenidos.

Finalmente, exponga en clases los resultados obtenidos (lunes 10 de agosto).