Unidad 2 - Representación Vectorial de Texto

Ejercicio 1 - Escriba ejemplos de frases o busque ejemplos de párrafos de texto con diferentes estructuras, palabras y signos de puntuación. Utilice los códigos de One-Hot encoding en las dos versiones que presenta el material expuesto en clase. Analice los resultados.

Ejercicio 2 - De los recursos propuesto en el último ejercicio de la unidad 1 con el texto obtenido de documento PROYECTO DE LEY TURISMO SOCIAL 2004.pdf y el texto de la metodología en la extracción de webscrapping del Ministerio de Turismo.
Utilizar las bibliotecas de procesamiento de texto en Python para representar estos documentos en forma de matrices numéricas utilizando tanto CountVectorizer como TfidfVectorizer, y luego comparar las diferencias entre las dos representaciones.
Si es necesario eliminar en alguno de los casos las “stop-words”

Ejercicio 3 - Utilice la librería HashingVectorizer para obtener los vectores de características resultantes en los documentos anteriores.

¿Qué quiere decir que esta metodología es una técnica “sin estado”?

¿Cuándo puede ser una ventaja el uso de esta técnica?

¿En qué casos puede generar algunos problemas?

¿Explique el concepto de “matriz dispersa”?

¿Qué rango numérico tienen los valores si se utiliza norm=None?

Ejercicio 4 -¿A qué hace referencia la bibliografía cuando indica que los métodos utilizados en los ejercicios anteriores no capturan la semántica y el contexto de las palabras.? De una explicación breve.

Ejercicio 5 - Genere un código en python que permita tomar el resultado de una matriz de vectores de los ejercicios anteriores y extraiga las 10 similaridades de coseno mayores e imprima las palabras correspondientes en los casos que es posible.

Ejercicio 6 - Cargue el modelo Word2Vec y explore palabras y similitudes en palabras que sean de su interés. Busque posibilidades y combinaciones donde el uso de la aritmética de palabras de resultados semánticamente coherentes.

Ejercicio 7 - Utilice el texto recuperado del autor Hernán Casciari de la práctica anterior para crear un corpus de texto, elimine las stopwords y utilice Word2Vec, GloVe (Global Vectors for Word Representation) y FastText, compare los resultados.

Ejercicio 8 - Separe por párrafos u oraciones el texto recopilado de Casciari. Usar doc2vec sobre este corpus.