HOJA DE EJERCICIOS

PROCESAMIENTO DEL TEXTO UTILIZANDO NLTK

Se proponen diferentes ejercicios para procesar el texto y practicar con la librería NLTK.

** Los alumnos opcionalmente podrán utilizar sus propios ficheros de texto para realizar las comprobaciones que crean necesarias además de los indicados por defecto.*

1. Dado un fichero de texto con contenido en inglés ("Data_Science.txt") se quiere eliminar todos los signos de puntuación que aparezcan en el mismo.

2. Dado un fichero de texto con contenido en inglés ("Data_Science.txt") se quiere saber cuáles son las 5 palabras más frecuentes en el texto, antes y después de eliminar los signos de puntuación. Además, se quiere saber cuántas veces aparece la palabra "data" en el texto.

3. Dado un fichero de texto con contenido en inglés ("Data_Science.txt") se quiere obtener su contenido, dividirlo en frases y por cada una de ellas eliminar las palabras vacías (stop words) que contengan.

4. Se pide lo mismo que en el ejercicio anterior, salvo que hay que hacerlo para un texto en español ("Ciencia_de_datos.txt").

5. Dado un fichero de texto con contenido en inglés ("Data_Science.txt") se quiere hacer uso de expresiones regulares para encontrar palabras que cumplan unas determinadas condiciones en el texto, por ejemplo:
    a. Buscar las palabras que comiencen por 'd'
    b. Buscar las palabras que comiencen por 's', terminen por 'e' y cuya longitud total sea de 7 caracteres.

6. Dado un fichero de texto con contenido en inglés ("Data_Science.txt") se pide obtener los diferentes tokens del texto y por cada uno de ellos realizar stemming.

7. Se pide lo mismo que en el ejercicio anterior, salvo que hay que hacerlo para un texto en español ("Ciencia_de_datos.txt").

8. Dado un fichero de texto con contenido en inglés ("Data_Science.txt") se pide obtener los diferentes tokens del texto y por cada uno de ellos lematizar

9. Se pide lo mismo que en el ejercicio anterior, salvo que hay que hacerlo para un texto en español (“Ciencia_de_datos.txt”).

10. Dado un fichero de texto con contenido en inglés (“Data_Science.txt”) se quiere obtener la categoría gramatical de cada uno de los tokens que componen el texto. Se quiere conocer la categoría gramatical con la notación por defecto y con la notación universal. La diferencia entre una y otra es que la primera presenta diferentes etiquetas para las diferentes variaciones de una misma categoría, mientras que la segunda agrupa en una sola etiqueta todas esas variaciones. Por ejemplo, en el caso por defecto se puede etiquetar un nombre de muchas formas diferentes (NNP para anotar que es un nombre propio, NN para anotar un nombre común, NNS para anotar un nombre común en plural, etc.). Sin embargo, todos los casos anteriores se etiquetarían como NOUN en el caso de la notación universal.

11. Se pide lo mismo que en el ejercicio anterior, salvo que hay que hacerlo para un texto en español (“Ciencia_de_datos.txt”).
    * Se recomienda utilizar el Stanford tagger para ello. http://nlp.stanford.edu/software/tagger.shtml

12. Dados textos en inglés de tres ficheros diferentes (“Data_Science.txt”, “Science.txt” y “Data.txt”) se pide construir un corpus (colección de textos) con ellos. Después se quiere saber lo siguiente:
    a. Conocer cuál es la frecuencia de aparición de la palabra “data” en uno de los textos de la colección, es decir, se pide el valor TF (Term Frequency), que indica cuál es la frecuencia de aparición de la palabra “data” en el texto elegido.
    b. Se quiere conocer cuál es el valor TF-IDF para la misma palabra en el mismo texto, es decir, se quiere conocer cuál es la frecuencia de aparición de la palabra en un documento con respecto a la frecuencia de aparición en la colección.

13. Dado un texto en inglés (“Data_Science.txt”) se quiere obtener el “vocabulario” de dicho texto. Normalmente el vocabulario lo forman los diferentes términos del texto sin repeticiones y habiendo eliminado signos de puntuación, stopwords, e incluso dependiendo de la tarea habiendo hecho stemming o lematización.

14. Dado un conjunto de textos en inglés (“Data_Science.txt”, “Data.txt”, “Credit_card.txt” y “Science.txt”) se quiere obtener lo siguiente:
    a. El número total de palabras diferentes juntando todos los textos.
    b. El tamaño del vocabulario para cada texto.

c. El vocabulario completo juntando todos los textos.
d. Palabras que están en un solo texto.
e. Palabras que están en más de un texto.

15. A partir de un conjunto de textos en inglés (“Data_Science.txt”, “Data.txt”, “Credit_card.txt” y “Science.txt”) se desea obtener la siguiente información:
    a. El número total de palabras diferentes juntando todos los textos.
    b. Qué palabras aparecen en cada texto mediante TF (Term Frequency). Se debe llevar a cabo mediante una matriz (palabras x textos) que muestre 0 o 1 en la posición según si aparece o no la palabra en el texto.

* *Para mejorar la visualización del resultado es recomendable utilizar un DataFrame (librería “pandas”).*