Integrantes: Fabian Pallares, Daniel Mejia, Laura Mora

Agosto 29/2022

**Taller 1 NLP.**

1. Crear un programa en el que se genere un archivo (fichero\_suma.txt) que sea el resultado de la suma de los valores contenidos en dos archivos (fichero1.txt, fichero2.txt). Estos ficheros contendrán 5 valores (1 por línea).

Inicialmente se investigó sobre el manejo de archivos en python y se encontró la librería OS, con esta se permite leer y escribir en archivos. Con el fin de dar solución del problema se decidió sumar el contenido de ambos archivos txt, inicialmente se lee el fichero1 y se van guardando sus datos en una suma, posteriormente se realiza el mismo proceso con fichero2; tras tener ambos datos acumulados se suman y se guardan en un nuevo archivo fichero\_suma.txt con la función write de la librería OS.

1. Utilizar el sistema de ayuda de Python para obtener información sobre las librerías “nltk” y “sklearn” que serán utilizadas en las siguientes prácticas.

El sistema de ayuda de Python para las librerías viene con la función help(), para conocer el contenido y las descripciones de cada una de estas se ingresó help(“nltk”) y help(“sklearn”).

1. Producir un diagrama de dispersión de los cuatro protagonistas Elinor, Marianne, Edward y Willoughby en “text2: Sense and Sensibility by Jane Austen 1811”:
   1. ¿Qué puedes observar sobre los diferentes roles que juegan los hombres y las mujeres en esta novela?
   2. ¿Puedes identificar a las parejas?

Con el fin de generar el diagrama de dispersión, inicialmente se importa la librería nltk y todos los libros que esta trae, entre estos está el text2: Sense and Sensibility by Jane Austen 1811, con el fin de crear el diagrama de dispersión se utiliza la función asociada y se indican como paramentros los nombres de los personajes que estamos buscando (text2.dispersion\_plot(["Elinor", "Marianne", "Edward", "Willoughby"]) ) obteniendo el siguiente diagrama (Ilustración 1).

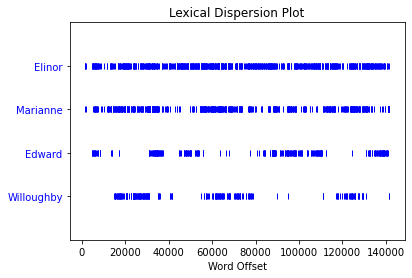


Ilustración 1. Diagrama de dispersión

Se puede observar que Elinor y Marianne (las mujeres) son las protagonistas de las historia, debido a la cantidad de ocurrencias que tienen, por su parte, Edward y Willioghby (los hombres) juegan un papel secundario, debido a que si bien tienen varias apariciones no son tantas como las de las mujeres.

Por otro lado, no pueden identificarse las parejas claramente ya que por momentos las cuatro interactúan en el mismo tiempo, para esto sería mejor utilizar otro comando para ver el tipo de interacciones entre ambos a través de los textos que los rodean.

1. En la primera oración del texto3 se crea la variablesent3que la contiene. Ahora en el índice 1de este arreglo sent3 nos da 'the', se expresaría sent3[1].
   1. ¿Cuáles son los índices de las otras dos ocurrencias de esta palabra en sent3?
   2. ¿Qué otros índices nos servirían para recuperar la misma información?

Para dar solución a este punto se realizó a ojo, y se decidió imprimir el arreglo sent3, obteniendo el siguiente resultado ['In', 'the', 'beginning', 'God', 'created', 'the', 'heaven', 'and', 'the', 'earth', '.'].

Se puede observar que los siguientes ‘the’, se encuentran en los índices 5 y 8.

Por otro lado, otra forma de obtener esos mismos índices seríasn utilizando el [-3], [-6] y [-10], es decir empezando a recorrer el arreglo de atrás hacia delante.

1. ¿Cuál es la diferencia entre las líneas a y b? ¿Cuál dará un mayor valor? ¿Este caso es generalizable?
   1. >>> sorted(set(w.lower() for w in text1))
   2. >>> sorted(w.lower() for w in set(text1))

Finalmente, para este punto se corren ambas líneas de código, si bien viendo el resultado al ser tan largo es difícil notar las diferencias, si analizamos el orden en el que se realiza cada proceso se obtiene que: la primera línea (a) convierte la cadena a minúsculas, luego remueve duplicados y ordena, mientras que la segunda línea (b) elimina duplicados, luego convierte a minúsculas y finalmente organiza. La segunda entonces va a dar un mayor valor ya que el valor del carácter minúscula es diferente a la mayúscula (no sería duplicado).