Práctica 1 PLN

Enrique Ernesto de Alvear y Javier Alarcón

March 2024

1 Tarea 1

Primero se nos pide cargar los datos del léxico de sentimientos descargado desde `https://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm` y guardarlo en la estructura de datos que creamos necesaria.

Usando pandas hemos cargado el léxico de sentimientos en castellano en python. También hemos eliminado la columna asociada al castellano y hemos eliminado aquellas palabras que no están asociadas a ningún sentimiento. Finalmente hemos creado un diccionario donde las claves son los sentimientos y los valores es un set formado por las emociones de cada sentimiento. Hemos elegido añadir los sentimientos en un set ya que para el siguiente apartado es posible que encontremos una misma palabra asociada a emociones distintas.

2 Tarea 2

En esta tarea se nos pide hacer una ampliación del diccionario encontrando palabras similares a las que contiene nuestro diccionario.

Para ampliar el diccionario lo primero que hemos hecho ha sido cambiar las claves del anterior diccionario, las claves actuales son de la forma $(lemma, POS - tag)$, para conseguir este formato primero hemos tokenizado las palabras y después le hemos aplicado el pos-tag. Por último, para ampliar el diccionario hemos considerado todos los synsets de cada palabra (antes de tokenizarla) y hemos calculado los hiperónimos, hipónimos, holónimos, palabras derivadas, y sinónimos, posteriormente hemos "asociado" a esta nueva palabra el set de la palabra base. Al hacer estas operaciones (incluyendo el tokenizado) es posible encontrar una misma palabra obtenida de distintas formas por ejemplo: al tokenizar dos palabras distintas, como palabra derivada de otra, como sinónimo,... esto supone un problema ya que esa misma palabra está asociada a sentimientos distintos, en nuestro caso lo que hemos hecho ha sido unir los sets asociados.

3 Tarea 3

A continuación, en esta tarea nos dedicamos a cargar desde *Proyect Gutenberg* un conjunto de novelas clásicas utilizando *BeautifulSoup*. Para ello se ha proporcionado un código en el enunciado de la práctica, el cual hemos ejecutado para completar la tarea.

4 Tarea 4

Aquí vamos a hacer el análisis de sentimiento de los datos propiamente. Para ello, nos hemos creado un diccionario que guarda como (clave = emoción, valor = número de apariciones), es decir, se hace un conteo de cuantas veces aparece cada una de las emociones. Vamos a tratar con el conjunto de emociones que está contenido en la base de datos del lexicón original, las cuales son: "anger", "anticipation", "disgust", "fear", "joy", "negative", "positive", "sadness", "surprise", "trust". Pero con esta filosofía no se tienen en cuenta las negaciones, por ejemplo la frase "No me gusta el chocolate", tendría como sentimiento el positivo que le daría las palabras, cosa que no se ajusta con el mensaje que quiere trasmitir la frase. Para tener en cuenta esto al diccionario de emociones mencionado anteriormente, además de las ya consideradas se van a tener en cuenta el conjunto de $-emociones$, lo que quiere denotar como que la frase tiene negaciones de la emoción que quiere trasmitir. Además, también se tienen en cuenta diversos potenciadores de sentimiento como puede ser "very" o "lot" entre otros, por lo que el sentimiento de esa frase se multiplicará por 1.5.

El proceso de análisis consiste en:

1. Preprocesar los textos: primero se leen los textos cargados en la Tarea 3, y se dividen por frases.
2. En cada una de las frases luego se tokeniza por palabras, añadiendo el *pos-tag* correspondiente.
3. Se añade a cada frase una ponderación de 1 inicial, si la frase contiene una negación entonces esta ponderación será -1 y si tiene un potenciador entonces se multiplicará por 1.5 la ponderación.
4. Al acabar la frase entonces se añadirá al diccionario el conteo de cada una de las emociones multiplicada por la ponderación correspondiente a la frase.

Este proceso se realiza para cada uno de los libros disponibles, hay que tener en cuenta que el libro "Fortunata y Jacinta" al estar en español no se tiene en cuenta en el diccionario que tenemos disponible, por lo que se ignorará para el análisis.

5 Tarea 5

Finalmente, se muestra en las siguientes figuras los análisis realizados sobre todos los textos disponibles en forma de histograma. Se muestra que la emoción más frecuente en los textos es positiva, normalmente con valores altos en *trust* y *anticipation*. Para los valores de las emociones en negativo ($-emotion$), normalmente sigue una distribución muy parecida a la de las emociones en positivo, salvo que tiene unas frecuencias menores. Se puede ver que prácticamente todos los libros siguen una distribución muy parecida para las emociones, cambiando obviamente la escala, ya que el número de palabras es muy distinto entre los libros, únicamente destacando la gran diferencia en el libro "Ulysses" donde se ve una gran diferencia entre las "-emociones" y las emociones, esto se podría deber a un uso reducido de negaciones en las frases. Podría decirse algo parecido en el caso de "The adventures of Sherlock Holmes" pero el más notorio sigue siendo el primero.

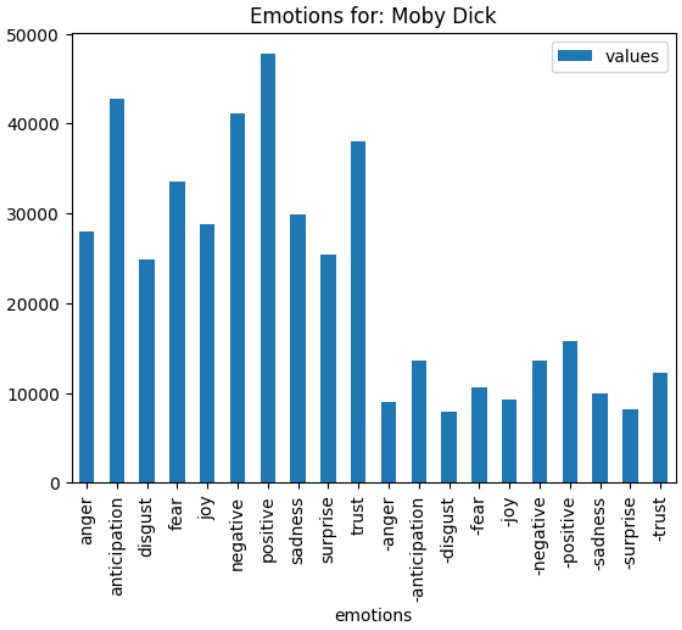

Figure 1: Análisis del libro: Moby Dick

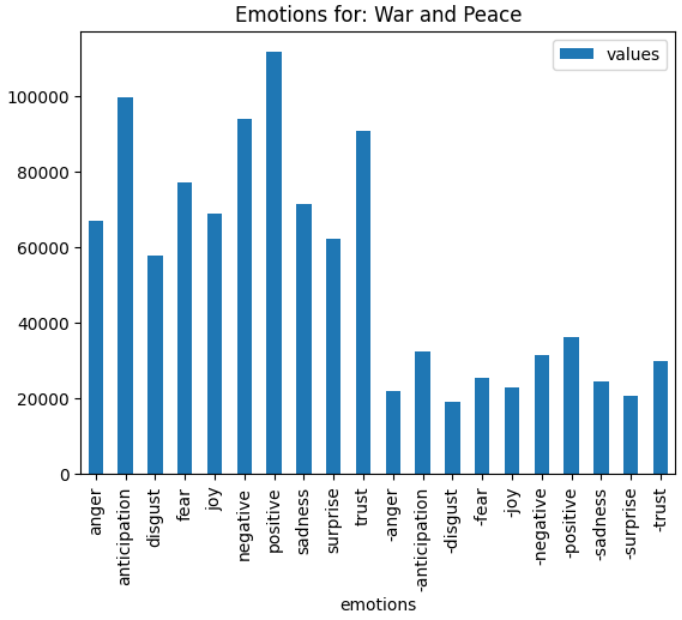

Figure 2: Análisis del libro: War and Peace

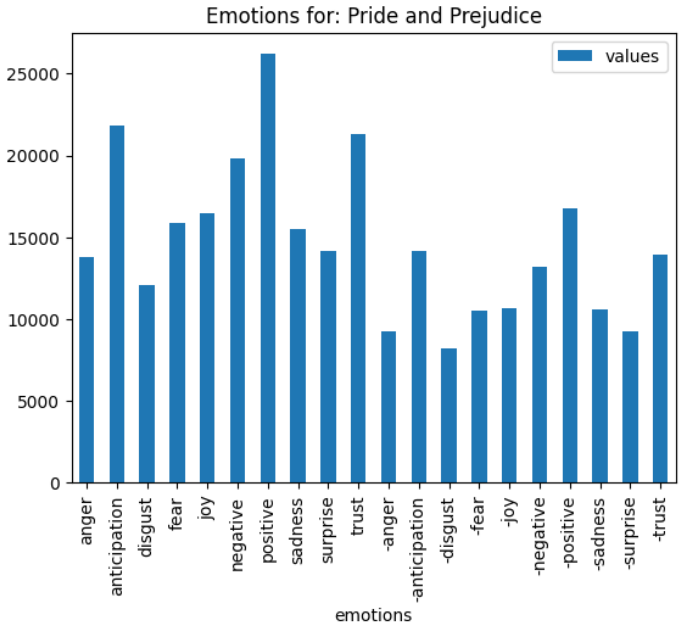

Figure 3: Análisis del libro: Pride and Prejudice

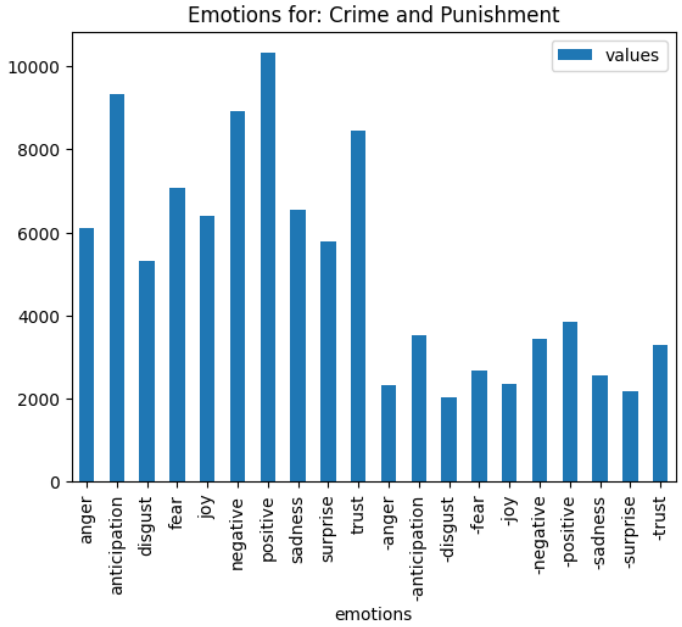

Figure 4: Análisis del libro: Crime and Punishment

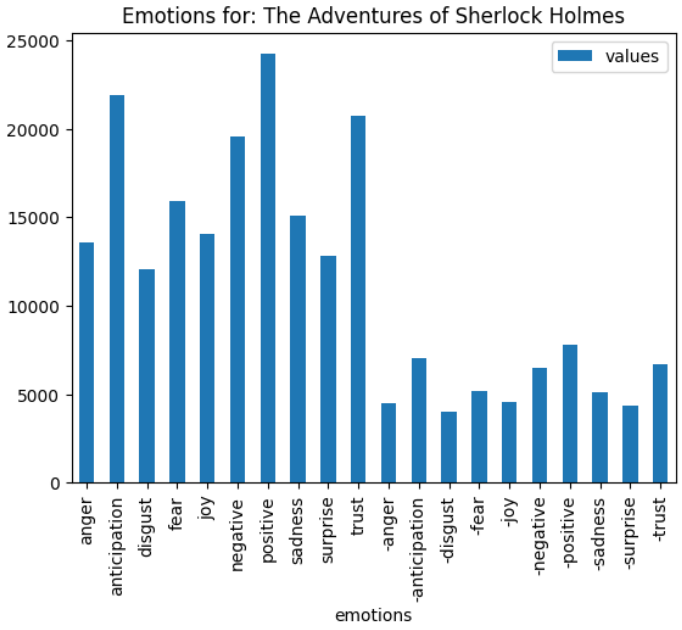

Figure 5: Análisis del libro: The adventures of Sherlock Holmes

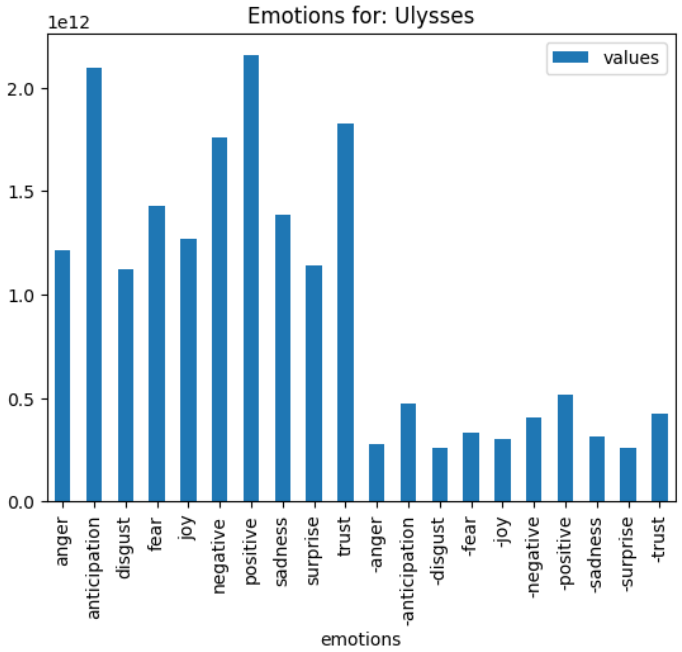

Figure 6: Análisis del libro: Ulysses

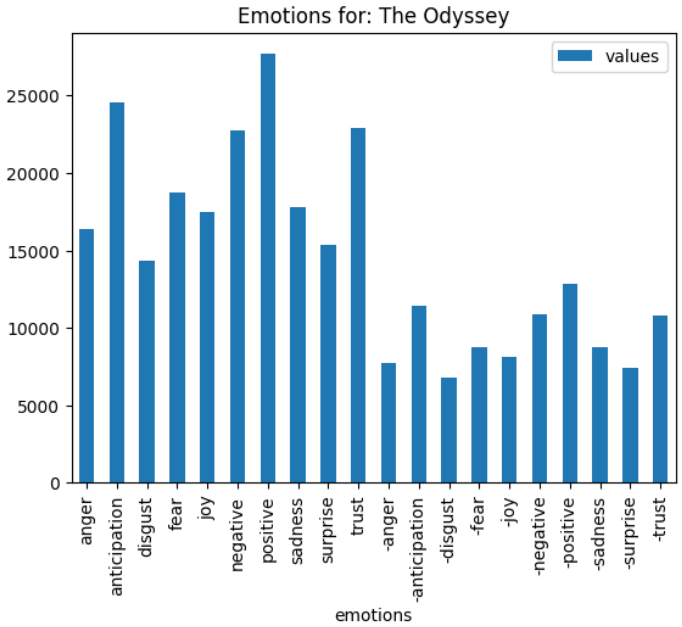

Figure 7: Análisis del libro: The Odyssey

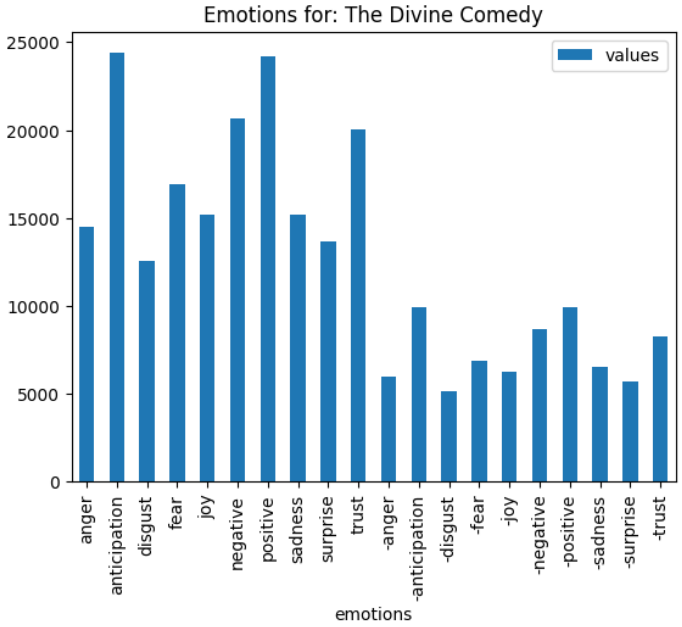

Figure 8: Análisis del libro: The Divine Comedy

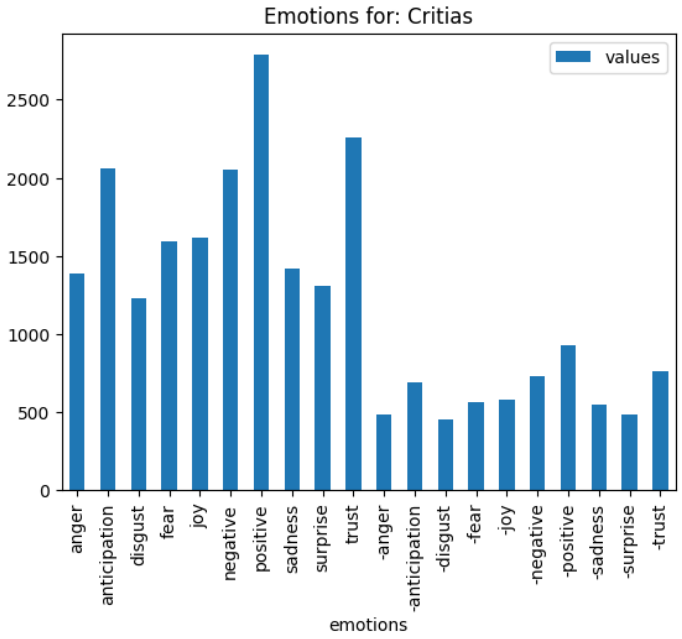

Figure 9: Análisis del libro: Critias