project 2

October 14, 2019

1 Proyecto 2

Computación II, IES-2019 Profesor: Eduardo Jorquera (eduardo.jorquera@postgrado.uv.cl)

Fecha/hora tope de entrega: 21/10/2019, 23:55 hrs

Formato de entrega: html y ipynb

Usando jupyter notebook, desarrolle lo siguiente:

I

1.- Usando la libraría `gutenbergr` descargue la obra número 31 de la indexación como lo hemos visto en clase, que corresponde a "Plays of Sophocles: Oedipus the King; Oedipus at Colonus; Antigone".

2.- Quite todo lo que está antes de "ARGUMENT". Luego, de todas las palabras enteras que estén en mayúsculas, muestre las únicas (es decir, sin repetirse al ser mostradas).

3.- Construya una nube de palabras con los términos más frecuentes de la obra, donde NO se muestren palabras vacías, y tampoco las palabras que estén completamente en mayúsculas; quite también donde dice "(Str. x)" y "(Ant. x)", donde x es un número.

4.- Compare el análisis de sentimiento usando los tres léxicos vistos en clase, para el texto separado en trozos de 80 líneas. Visualice los resultados.

5.- £Cuáles son las palabras positivas y negativas más comunes (frecuentes) según los léxicos de Bing *et.al.* y NRC?

6.- Para ambos resultados de la pregunta anterior, grafique una nube de palabras por cada uno de los dos léxicos mencionadoes, donde por un lado se muestre la parte positiva del texto, y por el otro la parte negativa.

II

7.- Para el libro La máquina del tiempo, en la indexación número 35, halle según el léxico de Bing *et.al.* los capítulos más positivos y más negativos del libro. Al igual que para el libro de Sófocles, construya una nube de palabras que separe los términos positivos y los negativos, sin las palabras vacías.

8.- Luego de responder a la pregúnta número 7, £cuál capítulo tiene la mayor proporción de palabras negativas?

8.- Para el mismo libro, usando el léxico NRC, halle para cada emoción, los capítulos más representativos para cada una.

9.- Usando el léxico AFINN, responda a la pregunta, cuál es el capítulo más negativo?

10.- Dividiendo el texto en trozos de 30 líneas, compare nuevamente los tres léxicos en un gráfico.