

Práctica 1B

Guillermo Ruiz

Crea un notebook para preprocesar el texto y convertirlo a minúsculas, quita acentos y los siguientes caracteres:

" ; : , . \ \ - \ " ' / () [] ¿ ? ¡ ! { } ~ < > | « » -- ' "

Reemplaza las tabulaciones y saltos de línea con un espacio.

2. Calcula las frecuencias a nivel de palabras de los siguientes documentos.

Libro 1 Libro 2

3. Genera los histogramas para cada archivo.
4. En el notebook responde las siguientes preguntas:

¿Cuántas palabras hay en cada texto?

¿Cuántas palabras diferentes (palabras únicas) hay en el texto?

¿Cuántas son las palabras con frecuencia 1 (hapax legomena) en el texto?

Genera los histogramas de todos los documentos. Para generar los histogramas ordena las palabras de acuerdo a la frecuencia de mayor a menor. Si el número de datos es excesivo y difícil de visualizar, presente los datos que considere para que se aprecien adecuadamente los datos graficados.

Archivos:

- libro_1.txt
- libro_2.txt

Palabras totales