Computación – Curso Servetto

Problema que involucra el uso de listas, conjuntos, diccionarios y archivos de texto

Objetivo

Desarrollar un programa que solicite al usuario el nombre de un archivo de texto a procesar y el nombre de otro archivo de texto donde reportar la siguiente información sobre el texto del primer archivo:

1. Cantidad de líneas (incluyendo vacías)
2. Cantidad de palabras relevantes
3. Cantidad de caracteres representables
4. Lista de palabras relevantes distintas (vocabulario) con la cantidad de veces que aparecen en el texto (frecuencia)
5. Lista de palabras más importantes del texto en orden descendente de importancia

Se consideran palabras relevantes que conforman el vocabulario de un texto a las que no aparezcan en listas definidas en otro archivo de texto que se provee (exclusiones.txt): preposiciones, pronombres, artículos y adverbios.

Precisiones

Las palabras que se deben ignorar o excluir del vocabulario de un texto se pueden cargar desde el archivo exclusiones.txt en uno o más conjuntos para, al momento de decidir si una palabra se incluye o busca en el vocabulario para computar una nueva aparición, preguntar si no pertenece al o los conjuntos de exclusión.

Para contar los caracteres representables deben tenerse en cuenta los que separen palabras, como espacios o signos de puntuación, pero no el de fin de línea ('\n').

Para determinar la importancia de una palabra, debe calcularse el cociente entre la frecuencia de la palabra en el texto y la máxima frecuencia que se observe de cualquier palabra en el documento. Se debe decidir y fundamentar con información de pruebas un cociente a partir del cual considerar que una palabra adquiere importancia.

Orientación

Se recomienda tener en cuenta las siguientes operaciones:

- *línea=arch.readline()* (o también la lectura implícita *for línea in arch: …*) determina que *línea* contenga una línea completa del archivo de texto abierto para lectura *arch*, incluyendo al final de la misma el carácter de control (no representable) de salto de línea '\n'
- *str.split()* devuelve una lista con las palabras de la cadena *str* sin considerar espacios en blanco ni el fin de línea (para separar palabras de una línea, pero que pueden tener caracteres de interrogación, exclamación, destaque -comillas- o aclaración -paréntesis- al comienzo o al final, o caracteres de puntuación al final)
- *str.lower()* devuelve una cadena sustituyendo las mayúsculas de la cadena *str* por las correspondientes minúsculas (para pasar todas las palabras a minúsculas de manera de no repetir palabras del vocabulario que comiencen con mayúscula con la misma que comience en minúscula).
- *str.lstrip(cad)* devuelve una cadena eliminando cualquier secuencia de caracteres que aparezcan en *cad* que estén al comienzo de la cadena *str* (para eliminar signos del comienzo de una palabra que no formen parte de ella, por ejemplo con *str.lstrip('\'¡-("¿')* -como la comilla simple se está usando

como delimitador de la cadena, para incluirla dentro de una cadena explícita debe especificarse con la barra \ como prefijo para distinguirla de su uso como delimitador)

- *str.rstrip(cad)* devuleve una cadena eliminando cualquier secuencia de caracteres que aparezcan en *cad* que estén al final de la cadena *str* (para eliminar cualquier signo del final de una palabra)
- *sorted(lista_desord)* devuelve una lista con los elementos de *lista_desord* ordenados de menor a mayor (para ordenar alfabéticamente la lista de palabras distintas o vocabulario del archivo)
- *set(lista)* devuelve los elementos de una lista representados en un conjunto, eliminando elementos duplicados en caso de que los hubiera

Se debe definir funciones propias del programa para descomponer el problema, por ejemplo, para cargar del archivo exclusiones.txt uno o más conjuntos de palabras a excluir de vocabularios, obtener conteos parciales de cada línea, obtener una lista con las palabras de una línea en minúsculas, incorporar una palabra al diccionario de palabras del archivo o aumentar su frecuencia, obtener la lista de palabras más importantes a partir del diccionario de palabras del archivo con sus frecuencias, etc.

Se proveen tres archivos de texto, exclusiones.txt, texto1.txt y texto2.txt, el primero con conjuntos de palabras que no deben incluirse en el vocabulario y los dos últimos para la prueba del programa, que se debe tener en cuenta que están codificados en UTF-8, por lo que para abrirlos se debe especificar *open(nomarch, 'r', encoding='UTF-8')*.