Ejercicio 1

Objetivo: Manipulación de cadenas y uso de expresiones regulares para construir un tokenizador para el español con una serie de restricciones.

Referencias:

Tutorial

http://docs.python.org/tutorial/index.html

String Processing:

http://docs.python.org/library/string.html

http://docs.python.org/tutorial/introduction.html#strings

Regular Expressions:

http://docs.python.org/library/re.html

http://docs.python.org/howto/regex.html#regex-howto

Construir un tokenizador para el español, que, dado un fichero de texto de entrada (*entrada_tokenizador_2023.txt*), separe en tokens, y los muestre en un fichero de salida en el formato que se muestra en (*ejemplo_salida_tokenizador_2023.txt*). Por lo menos el tokenizador deberá funcionar correctamente para el ejemplo.

El tokenizador debe cumplir las siguientes restricciones:

1) Los símbolos que hay que separar de cada palabra son: ( ) . , ‘ “ ? ¿ ! ¡ … ; :
2) No se deben separar los números decimales, ejemplo: 44,45 45.60
3) No se deben separar fechas 12/12/22, 12-03-23ni horas, 9:30
4) Las fechas en formato 12 de febrero de 2023, 12 de enero, … hay que mantenerlas como un token
5) No se deben separar direcciones web http://www.colorin.com ni correos electrónicos xx@cdit.com
6) Hay que mantener menciones a usuarios (@user) y hashtags (#hashtag) como se utilizan en Twitter.
7) Hay que mantener acrónimos, por ejemplo: EE.UU., S.L., CC.OO., S.A., U.R.S.S, …
8) Respetar los emoticonos: 💕👫👯😊
9) Se deben conservar los tratamientos: Sr., Sra., Dr., Dra., D., Dª, …
10) Se deben agrupar los nombres propios, asumiendo que un nombre propio es un Nombre en Mayúscula inicial y dos apellidos con mayúscula inicial: Juan Pérez Oliva