

Práctica-Formatos de datos para la Web(I)

Ejercicio 1[6 puntos]

Considerar el archivo cvs PitchingPost que contiene información sobre Baseball, se pide realizar funciones en Python que permitan realizar las siguientes acciones:

1. Crea un nuevo archivo cvs denominado AcumAnnos.cvs el cual debe contener la frecuencia de los años. Tendrá la estructura:
Años,Frecuencia
1980,2
1981,6
....
2. Crea un nuevo archivo cvs denominado AcumJugadores.cvs el cual debe contener la frecuencia de los jugadores. Tendrá la estructura:
Jugador,Frecuencia
bystrma01,2
carltst01,5
....
3. Crea un nuevo archivo cvs denominado Ordenado.cvs el cual tenga la información ordenada por el nombre del jugador.

Ejercicio 2[4 puntos]

Considerar el archivo tweets.txt que contiene tweets en formato json. Se pide realizar funciones en Python que permitan:

- a) Encontrar la frecuencia de un término, la cual puede calcularse como el cociente entre el número de ocurrencias de un término en los tweets y el número de ocurrencias de todos los términos en todos los tweets. Como resultado se debe generar un archivo que tenga en cada línea el nombre de un término seguido de la frecuencia del término. Además no puede aparecer líneas con términos repetidos. Hay que tener en cuenta lo siguiente:
 - Los datos recuperados de twitter que se encuentran en el archivo proporcionado deben convertirse en una lista en python. Para ello se leerá línea a línea el archivo recuperado, de manera que los elementos de la lista serán los tweets que se han recuperado.
 - En cada tweet, el contenido del tweet se encuentra en el campo "text", es por ello que se accede de forma indexada a ese campo.
- b) Encontrar los 10 hashtags más utilizados en los tweets recuperados. Observar que no hace falta obtener el hashtag de un tweet pues twitter ya lo ha recuperado. Los hashtags se encuentran en un campo denominado "hashtags" que se encuentra indexado a su vez por un campo denominado "entities". Como resultado se debe generar un archivo que tenga en cada línea el nombre de un hashtag seguido del número de veces que aparece.

Normas de entrega

- Fecha tope de entrega: 20/10/2015
- La entrega se realizará subiendo un archivo zip con las soluciones realizadas en Python. El archivo tendrá como nombre FormatosWeb(I)_GrupoX donde X será el número de grupo correspondiente.