

PONTIFICIA UNIVERSIDAD CATÓLICA DE CHILE ESCUELA DE INGENIERÍA DEPARTAMENTO DE CIENCIA DE LA COMPUTACIÓN

IIC2233 - Programación Avanzada 2° semestre 2015

Actividad 21

Manejo de Strings

Instrucciones

La agencia de publicidad "Soluciones Super" quiere analizar los tweets realizados durante el partido Chile-Perú de la Copa América y así evaluar el impacto de las campañas publicitarias.

Tienen pocos recursos para invertir en una solución real, por lo que le han pedido a usted, alumno del fantástico curso de Programación Avanzada, que implemente un prototipo de análisis de palabras frecuentes. Específicamente quieren saber cuáles son las palabras más utilizadas durante ciertos eventos del partido. Además quieren saber cuántas veces se menciona a los jugadores, ya sea por un apodo, nombre, apellido o twitter.

Ellos le han entregado tres archivos:

- eventos.csv que tiene los minutos del partido en los cuáles ocurrió un evento (tarjeta, cambio, gol, ...)
- jugadores.json que tiene los nombres, apellidos, usuarios de twitters y apodos de los jugadores de ambos equipos
- tweets.csv que tiene la lista de tweets, la hora en que fueron realizados y el usuario

Se han definido las siguientes reglas:

- Una palabra es una cadena de caracteres de largo 4 o más
- Una palabra es considerada frecuente si se encuentra 5 o más veces
- Para cada evento solo debe analizar los tweets de 3 minutos antes y 5 minutos después del evento. Solo estos tweets deben ser considerados en las palabras frecuentes.

Requerimientos

- Obtener las palabras frecuentes para cada evento
- Obtener la frecuencia de la aparición de un jugador en los eventos a partir de su nombre, twitter y apodos
- El output del análisis es:

Eventos.txt: para cada evento registre:
Minuto1: [Palabra1: frecuencia1, Palabra2: frecuencia2...]
Minuto2: [Palabra1: frecuencia1, Palabra2: frecuencia2...]

• Jugadores.json: que para cada jugador indique la aparición en cada evento de la siguiente forma: Nombre Apellido 1: [Evento1: frecuencia1, Evento2: frecuencia2...], Nombre Apellido 2: [Evento1: frecuencia1, Evento2: frecuencia2...]...

Notas

- El método transform_event_time(minute) recibe un minuto del evento y entrega una tupla (hora, minuto) de acuerdo a la hora UTC (la zona horaria de la hora de los tweets).
- El método frequent_words genera un archivo EventoX.txt, con X el minuto del evento, que puede subir al sitio http://www.wordclouds.com/ para revisar su resultado! Para el evento del minuto 93 debería ver algo así (solo importa la jerarquía de tamaño, el resto es random).



To - DO

- (1.00 pts) Obtener los tweets de la ventana para cada evento
- (1.50 pts) Obtener las palabras frecuentes para una ventana
- (1.50 pts) Obtener la frecuencia de cada jugador para cada evento
- (1.00 pts) Filtrar las palabras menos frecuentes
- (1.00 pts) Guardar los archivos con la información correspondiente

Tips

- Una estructura que le puede servir es defaultdict.
- Un tweet puede tener saltos líneas, por lo que el método readlines no le servirá.
- No es necesario tener una clase por tweeet o jugador.