Finger Excercise #1 (10 pts)

Tiempo Estimado: 4hs

Descripción:

El objetivo de este ejercicio es probar el lenguaje de alto nivel Pig para realizar consultas sobre archivos distribuidos.

Para ello vamos a trabajar sobre un Virtual Machine que ya tiene instalado Hadoop y Pig junto con un entorno gráfico muy simple de usar desde el browser de forma tal de solo tener que concentrarnos en los datos y su procesamiento.

Instrucciones:

Bajar el Oracle VirtualBox: https://www.virtualbox.org/
Bajar el HortonWorks Sandbox 2.1 en su versión para VirtualBox: http://hortonworks.com/products/hortonworks-sandbox/#install

Abrir el VirtualBox bajado de Horton, automáticamente se abre el VirtualBox de Oracle para correrlo.
Si todo sale bien deberia bootear el VirtualMachine y deberiamos poder acceder a la interfaz desde el browser accediendo al http://127.0.0.1:8000/.

Enunciado:

El set de datos que utilizaremos es el mismo que utilizaran para realizar el TP, y puede obtenerse de Kaggle. Puntualmente estaremos utilizando el archivo labeledTrainData, de donde tomaremos los reviews de peliculas, y nos apoyaremos del archivo auxiliar AFINN-111.txt, de donde podremos obtener una puntuación para cada palabra..

- **labeledTrainData.tsv** - Los campos se encuentran separados por tab, y contiene 25,000 reviews de película descriptos por los campos: id, sentiment, y el texto del review.

Ejemplo de labeledTrainData:

```
"2381_9"1       "\"The Classic War of the Worlds\" by Timothy Hines is a very entertaining film that obviously goes to great effort and lengths to faithfully recreate H. G. Wells' classic book. Mr. Hines succeeds in doing so. I, and those who watched his film with me, appreciated the fact that it was not the standard, predictable Hollywood fare that comes out every year, e.g. the Spielberg version with Tom Cruise that had only the slightest resemblance to the book. Obviously, everyone looks for different things in a movie. Those who envision themselves as amateur \"critics\" look only to criticize everything they can. Others rate a movie on more important bases,like being entertained, which is why most people never agree with the \"critics\". We enjoyed the effort Mr. Hines put into being faithful to H.G. Wells' classic novel, and we found it to be very entertaining. This made it easy to overlook what the \"critics\" perceive to be its shortcomings."
```

- **AFINN-111.txt** - Los campos se encuentran separados por tab, y contiene palabra, y una puntuación (que puede ser positiva o negativa), indicando el sentimiento asociado con esa palabra.

Ejemplo de AFINN-111:

```
abductions      -2
abhor   -3
abhorred-3
ability 2
aboard  1
absentee-1
absolve 2
```

Escribir un programa en Pig que permita, utilizando los archivos anteriormente descriptos, obtener cuales son los 5 reviews mas positivos, y cuáles los 5 mas negativos. Hay que tener en cuenta que los reviews tienen distinta cantidad de palabras, por lo que hacer la suma va a favorecer (y penalizar) los reviews mas largos. Por eso es necesario calcular el promedio.

Publicar los resultados en Facebook en el grupo de la materia, el primero con los resultados correctos se lleva los puntos.

Para que resulte mas facil, les dejamos las primeras líneas de programa, que incluyen la apertura y parseo de los datos. A partir de ahí puede cada uno seguir con la resolución (y si quieren hacer rehacer todo tambien pueden, no estan obligados a usarlo!)

```
Palabras = load 'AFINN-111.txt' using PigStorage ('\t') AS (pal:chararray, puntaje:int);

Calif = load 'labeledTrainData.tsv' using PigStorage ('\t') AS (nro_calif:chararray,
pts_calif:int, desc_calif:chararray);

TopCalif_Pals = foreach Calif generate nro_calif, pts_calif, TOKENIZE(desc_calif) AS
pals_calif_pts;

Pals_Calif = foreach TopCalif_Pals {
   pal = foreach pals_calif_pts generate token;
   generate nro_calif, pal.token;
   };

Pals_Calif_FL = foreach Pals_Calif generate $0, FLATTEN($1);

/* YOUR CODE HERE */
```