**Spark Technical Test**

**You will be given:**

Two files:

* tweets.zip: containing a sample data set consisting of tweets captured during the World Cup Brazil 2014.
* phrases.txt: the set of phrases used to capture these tweets from the Twitter streaming API.

**Task:**

Write code that is using the latest release of Apache Spark and the Apache Spark Core Scala API (rdd transformations and actions) to answer these questions:

* What is the total number of tweets in the data set?
* What are the ids of all users that tweeted?  
  Output - a text file containing a unique list of user ids (one per line).
* What are the counts of all the words in the text field of all tweets?  
  Output - a csv file containing the word and the associated count on each line.
* What are the top N phrases in the dataset?  
  Input - The phrases are given to you in a separate file (phrases.txt), you need to match them against the tweet text first and only count the phrases that matched.
* N has to be configurable.  
  Output - a text file containing the top N phrases and the associated count on each line.
* What are the N-grams of the text of the tweets and the associated count  
  Input - N has to be configurable.  
  Output - a csv file containing the N-gram and the associated count on each line.  
    
  The code for each of the questions above needs to be contained in its own class.

The classes have to be parametrized accordingly.  
  
The answers need to contain the code for each class and instructions how they should be run as well as the outputs (files) when running the classes on the given data set