Scalable Data Analytics **Homework 3**
Spring 2020 **03/15/2020**

Deadlines Homework 3 is due on Mar.15th 23:55pm. No submission is accepted after the due.

How to submit: Please submit a zip file to the *Assignment/Homework_3* folder in the iCollege. The zip file name should be 'Yourname-Pantherid.zip'. In the zipped folder it should contain two python files '1-length.py' and '2-wordranking.py' for the first and second problems, respectively, and one report "HW2.pdf" illustrating your experiments.

In the report, you need to explain how you design the PySpark programme for each problem. You should include following sections:
1) The design of the programme.
2) Experimental results, 2.1) Screenshots of the output, 2.2) Description of the results.

You may add comments to the source code to help better understanding of the code for the graders. The report should be in a PDF file.

Data Set: The instructor has prepared a collection of Amazon reviews (2038 review comments from 88 products) in the file Amazon_Comments.csv. The data set is in the iCollege under the folder Homeworks. The columns are named and organized in the following manner: ProductID, ReviewID, ReviewTitle, ReviewTime, Verified, ReviewContent, ReviewRating seperated by "∧".

---

1. (7.5 points) Length of the comments

   There are fake comments created by the computers in the Amazon review system. Prof. Michael Luca from Harvard Business School argues [1] that there's been some evidence that fake reviews are sloppier in general: "Short, vague reviews are a pretty good marker, [along with] poor punctuation and grammar."

   Here are some examples of probably fake comments (e.g., "GREAT") and their corresponding ratings (e.g., 5 Star) in our data set:

   ```
   6^220^Five Stars^2016-01-09^false^ Quality product.^5.00
   6^221^Five Stars^2016-01-09^false^ Great quality.^5.00
   6^222^Five Stars^2015-11-25^false^ Excellent^5.00
   6^223^Five Stars^2016-01-14^false^ GREAT^5.00
   ```

   It looks like that these fake reviews tend to be more common in the 5 star ratings than 1 star ratings. Let's examine the average length (number of the words) of the comments for each rating and see if it really holds.

   Please design and implement a PySpark programme to examine the average length of comments (column: ReviewContent) in each rating (column: ReviewRating). We have

[1] Six Clues That an Online Review Might Be Fake

5 levels of rating here where 1 star rating represents the worst experience and the 5 star rating represents the best experience. Hint: you can remove punctuation in each comment with the following code:

```
import re
re.sub('\W+', ' ', mystring).
```

'\W+' is a regular expression that matches any non-alphanumeric characters.

What expected:

You should turn in an one python file which prints out the average length of the comments for each star rating:

```
$ spark-submit 1-length.py

1 star rating: average length of comments __
2 star rating: average length of comments __
3 star rating: average length of comments __
4 star rating: average length of comments __
5 star rating: average length of comments __
```

2. (7.5 points) Top words

Please design and implement a PySpark programme to pick up the top 10 words for each rating. Some words such as "great", "good" are common in the 5 star rating comments, and others such as "bad", "worst" are common in the 1 star rating comments.

Please remove the stop words such as "the", "an", "of", etc. in each comment before obtaining the results.

What expected:

Your Python code should print out the top 10 common words for each star rating:

```
$ spark-submit 2-wordranking.py

top 10 common words
1 star rating : __ __ __ ...
2 star rating : __ __ __ ...
3 star rating : __ __ __ ...
4 star rating : __ __ __ ...
5 star rating : __ __ __ ...
```