Scalable Data Analytics **Homework 2**
Spring 2018 **March 26, 2018**

Deadlines Homework 2 is due on April.9th 04:00pm. 50% late penalty will be applied within the first week of due date and no submission is accepted thereafter.

How to submit: Please submit a zip file to the *Assignment/Homework*_2 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 problem respectively, AND one report "HW2.pdf" illustrating your experiments.

In the report, you need to explain your basic idea about how to design the PySpark program for each problem. You should include the answers to the following questions.
1) Explanation of the code design
2) Experimental Results
2.1) Screenshots of the output. Please take a screenshot of the output in the terminal.
2.2) Explain your results.

You may add comments to the source code such that the source code can be read and understood by the graders. The report should be a PDF file. Please use a text editor, such as Microsoft Word, to write a report. Please transfer the file into a PDF file and then submit it.

Submission Materials: a) Your report b) Source code (.py 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 "$\wedge$".

---

1. (7.5 points) Length of 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
```

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

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 program to examine the average length of comments (column: ReviewContent) in each rating (column: ReviewRating). We have 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 from the Python string with the following code: s.translate(None, string.punctuation).

You will turn in an one python file and print out the average length of the comments in each rating like follows:

```
$ 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 for each rating

Please design and implement a PySpark program to pick up the top 10 words for each rating. Some of the words such as ”great”, ”good” are more common in 5 star rating comments than the 1 star rating comments.

Hint: you can use (RATING, WORD) pair as keys and count the frequency of such pairs.

Your Python code should print out the top 10 common words for each rating comments like follows:

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

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