Scalable Data Analytics **Homework 4**
Spring 2023

Deadlines Homework 4 is due on April.10th at 12:30 pm. 50% late penalty will be applied within the first week of the due date and no submission is accepted thereafter.

How to submit: Please submit a zip file to the *Assignment/Homework_4* folder in the iCollege. The zip file name should be 'Yourname-Pantherid.zip'. In the zipped folder it should contain three python files '1-k-means.py', '2-lr.py' and '3-length-DataFrame.py' for the first, second and third problems respectively.

---

1. (4 points) Machine Learning, K-Means algorithm

   In the homework folder in the iCollege you will find a file "kmeans-data.txt". Please write a PySpark programme (using spark.ml library) to cluster the data into two clusters.

   You should turn in an one python file which prints out centers for each cluster. Print the centers:

   ```
   $ spark-submit 1-k-means.py

   Cluster centers:
   [0.1 0.1]
   [9.1 9.1]
   ```

2. (4.5 points) Machine Learning, Logistic Regression

   Download the Yelp review data set. Use the logistic regression to predict the rating of first 10 comments. You can train your model with the first 1000 comments.

   You should turn in an one python file which prints out the comment and the prediction results:

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

   Comment 1 --> prediction = 2.000000
   Comment 2 --> prediction = 4.000000
   Comment 3 --> prediction = 3.000000
   ......
   Comment 10 --> prediction = 5.000000
   ```

3. (4 points) Spark DataFrame

   Rewrite your code for the HW3 Question 1 using Spark DataFrame.

```
$ spark-submit 3-length-DataFrame.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 __
```