CS7265-Homework 3

In HW3, we are going to practice Machine Learning algorithm on Apache Spark. We will classify spam message using SMS Spam collection data set at UCI machine learning repository: https://archive.ics.uci.edu/ml/datasets/SMS+Spam+Collection. The dataset includes text messages of SPAM and HAM. Please follow the procedure:

1. Upload the data file to the server.
2. Generate TF-IDF data file. You can obtain structured data of TF-IDF (matrix) from the unstructured data of text data. Check the page: https://spark.apache.org/docs/2.1.0/ml-features.html#tf-idf
3. Compute accuracy with 5-fold CV by using Naïve Bayes classifier on spark. Check at: https://spark.apache.org/docs/2.2.0/mllib-naive-bayes.html
4. You have to show accuracy of each experiment.

Submission:

You have to submit the followings to D2L:

1. MS word file
   - Describe what you did for the homework assignment.
   - Five accuracy of 5-fold CV

2. Source code file(s)
   - Must be well organized (comments, indentation, …)
   - Please upload its PDF version as well as the original file.

DO NOT upload a zip file.

Deadline:

You have to submit HW4 by **Monday, November 5, 2018**. Late assignments will be accepted up to 24 hours after the due date for 50% credit. Assignments submitted more than 24 hours late will not be accepted for credit.