BU.330.740. Large Scale Computing with Hadoop

Assignment 2: Spam Filtering Using Spark MLlib

Learning Goal: using Spark MLlib, EMR cluster and Jupyter notebook to implement spam filtering following example in lecture 4 notes page 38-41.

Input Data:
spam.txt which contains some spam email samples, and
ham.txt which contains some normal email samples

Implementation:
Follow lecture 4 notes page 38-41, implement spam filtering using PySpark on AWS. Use Jupyter Notebook running on Hadoop cluster. Refer to lab 4 which implements wordcount on lecture 4 notes page 36-37.

Extension:
Collect some spam text samples, and some non-spam text samples (one potential source is your own email), and some test samples (spam and non-spam). Train and test the model on your own samples and see whether it works.

Submission:
After implementation, download .ipynb file. Zip the file and submit on Blackboard.

Reference:

https://github.com/databricks/learning-spark/blob/master/src/python/MLlib.py

https://www.kaggle.com/uciml/sms-spam-collection-dataset