Lab 11
Big Data Spring 2016

In this lab we will use the Spark's MLlib library to perform frequent itemset analysis on a small sample of market basket data.

Setup

1. Start an EMR cluster. Under "Applications" select the "Spark" radio button. Use your public/private key pair. While your cluster is starting, we will go over the algorithms used in Spark's FPGrowth module.

2. SSH into the master node and get the files for the lab:
   ```
   hadoop fs -get s3://bigdataclassecc/Lab11/freqitems.py
   hadoop fs -get s3://bigdataclassecc/Lab11/groceries.csv
   hadoop fs -copyFromLocal groceries.csv
   ```

Run the Sample Program

1. Type `cat freqitems.py` to view the program

2. To run the job, use the command
   ```
   spark-submit freqitems.py groceries.csv > freqitemsoutput.txt
   ```

3. Type `cat freqitemsoutput.txt` to view the output file

Deliverable

Due date: Monday, May 9, 12:00pm (noon).
Suppose you are deciding which items to place next to each other at the grocery store, so you only care about frequent itemsets of size 2 or greater which appear in at least 5% of the transactions.

1. Modify the freqitems.py file to meet these constraints. (This involves both setting an appropriate minSupport and modifying the code to prune itemsets of size 1).

2. Run the job with your modified freqitems.py file using spark-submit, saving the output to the file modifiedoutput.txt

3. Submit the modifiedoutput.txt file to NYU Classes. (Note you can move the output to your S3 bucket with the command `hadoop fs -put modifiedoutput.txt s3://yourbucket`).