Machine Problem 5

Due: 5/6

Apache Mahout:

This assignment is more open ended than in previous ones. This time you are being asked to find a dataset analyze it using Apache Mahout. Apache Mahout is a machine learning library built to run on Hadoop to use the map/reduce paradigm. Mahout can be used to perform things such as collaborative filtering, user and item based recommendations, clustering, and classification. You should set up Hadoop and Mahout on your own computer or your own virtual machine.

First, choose a couple of datasets two different datasets.

Second, convert these datasets to a SequenceFile for Mahout use.

Third, run certain machine learning algorithms on these datasets.

Fourth, compare the results from the these datasets.

1. (20 pts) Setting up Hadoop and Mahout on your local machine (suggest using the psuedo-distributed version):
   http://hortonworks.com/hadoop/mahout/

2. (20 pts) Complete the Naive Bayes Wikipedia Example
   http://mahout.apache.org/users/classification/wikipedia-bayes-example.html

3. (60 pts) Find fun with Mahout
   1. Choose 1 Mahout tool
   2. Use this tool on any dataset (dataset suggestions listed below)
   3. Collect the results
   4. Write up what you have found from this

For your deliverables, you should turn in a 3-5 page report in pdf format. The report should include answers to the following:

1) Describe the process of setting up Hadoop and Mahout on your local machine

2) Describe what you learned from the naive bayes wikipedia example

3) Describe what you did with Mahout
   a. Description of the datasets
   b. Description of the tools
   c. Description of the process (what you did in mahout and how you did it, converting data to mahout compatible format, use of the mahout api, use of scripts etc)
   d. List results from analyzing the data (this can include graphs, tables, etc)

4) Any comments about using mahout (any problems, how can mahout be enhanced for more skillsets, etc)

5) Any screenshots you deemed necessary

You should turn this report into compass2g along with any output files that you may generate in one zip file named after your team number. Source code is not necessary to turn in. Please cite the extra credit work that you did.

Extra Credit (20 points):

There are multiple ways to receive extra credit points, here are some possibilities. If you have another idea on how to receive extra points please post it on piazza.

1. Inputs coming from HBASE, Storm, etc (10 pts)

2. Show the output with some form of visualization (graph, chart, etc) (10 pts)

Describe extra credit in write up!!!
Misc:

Example on converting files for Mahout use:

Location of usable datasets:
https://aws.amazon.com/datasets/2759?_encoding=UTF8&fromSearch=1&queryArg=searchQuery&searchPath=datasets&searchQuery=weather&x=0&y=0

http://archive.ics.uci.edu/ml/datasets/Abalone
http://grouplens.org/datasets/movielens/
http://www.kdnuggets.com/datasets/index.html
http://www.scaleunlimited.com/datasets/public-datasets/
http://www.occamslab.com/petricek/data/