Brandeis University
Department of Computer Science
COSI 129a - Introduction to Big Data Analysis
Fall 2016

Assignment 3: Using Mahout on the Cluster

1 Introduction

This assignment is the continuation of the second assignment, taking an index created in that assignment and running a classifier over it. Its sole purpose is for you to successfully use the Mahout software package on the cluster. This should be much less involved than the previous two assignments. You will again do this in your hadoopXX group.

2 Resources

The following resources are provided on the cs129a account:

1. /home/o/class/cs129a/assignment3/professions.txt
2. /home/o/class/cs129a/assignment3/wiki-big-lemma-index.tbz2

The file *professions.txt* lists the professions of 674K persons. Here are the first five lines:

```
Ken Bennett (Australian rules footballer) : rules footballer
Philippe-Jacques van Bree : painter
Stefano Protonotaro da Messina : poet
David Evans (rugby player) : rugby union player
Mattias Mete : footballer
Daniel A. Farber : historian, legal scholar, social scientist
```

One person can be tagged with multiple professions in the professions file. The semantics of this multiplicity could be different from person to person. A person could have had multiple jobs in their life, and is therefore recognized as having worked in different professions. Another type of multiplicity could result from the hierarchical system of profession names in Wikipedia. For example, a *zoologist* could also be called a *biologist*. You may have observed that the total number of people in *professions.txt* is not the same as the total number of people in *people.txt* from assignment 2. People having uncommon jobs or people whose job couldn't be extracted are excluded from data set.

The *wiki-big-lemma-index.tbz2* file is also on HDFS on the clusters at */shared*. This file is given in case you do not have your own version of the ARTICLE_LEMMA_INDEX that you created in assignment 2.

3 The Problem

You need to classify the articles that you processed in assignment 2 given the list of professions in *professions.txt* using a Naive Bayes classifier in Mahout. There are several parts to this:

- Create the statistical model. Here you select the elements from ARTICLE_LEMMA_INDEX for which you know what the professions are. You need to find the relevant Mahout command and figure out what the correct input to it is (and you will probably need to do some processing over ARTICLE_LEMMA_INDEX and *professions.txt*).
- Run the model over the rest of the data and classify each person. For each article, you are allowed to report at most three most likely professions resulting from your classifier run. The format of your output should be the same as in *professions.txt*. For example:

  Albert Einstein : *physicist*, *cosmologist*

- Evaluate how well your classifier works. For this part you should split *professions.txt* into a training set and a test set, typically you try to make the training set as big as you can, in this case a test set of several tens of thousands names seems reasonable. You should report what percentage of the articles in the test set you get right, you may consider your answer correct if one of the professions you find (at most three) is listed in the test set for the article.

See `http://mahout.apache.org/users/classification/bayesian.html` for pointers on how to use Mahout for creating a model and running a classifier.

4 Submission

You are required to submit the following materials for grading:

- Your source code, which should be well-documented.
- Your classification result in the specified format.
- A short report (about 2 pages) of your project, including (1) a short description of what you did, (2) clear pointers on how to run your code, and (3) your evaluation results.

5 Grading

Your grade will depend on:

- The quality of your code (does it run? is it well-structured? is it easy to understand?) (60%)
- The quality of your report. (40%)

6 Extra Credit

You can earn extra credit if you explore a couple of other machine learning algorithms from Mahout and compare them with the Naive Bayes classifier.