Lab Course: distributed data analytics
Exercise Sheet 8

Nghia Duong-Trung, Mohsan Jameel
Information Systems and Machine Learning Lab
University of Hildesheim
Submission deadline: Friday 17th, at 23:59PM (on LearnWeb, course code: 3117)

Instructions

Please following these instructions for solving and submitting the exercise sheet.

1. You should submit a zip or a tar file containing two things a) python scripts and b) a pdf document.
2. In the pdf document you will explain your approach (i.e. how you solved a given problem), and present your results in form of graphs and tables.
3. The submission should be made before the deadline, only through learnweb.
4. if you are M.Sc. Data Analytics winter 2016 intake student, you should submit to "Second term students" link on LearnWeb.
5. If you are not M.Sc. Data Analytics student, you can submit to anyone of the links above.
6. You have to provide performance analysis even if it is not explicitly mentioned in the question.

This lab is only for Winter 2016 intake students.
Comment on Marking: Lab will be graded based on the submitted report and code. A small subsmaple of students might be asked to present and explain their solution.

Exercise 1: Find an image histogram using aggregation (5 points)

First of all, please revisit exercise sheet 3 to refresh the concepts of an image histogram and OpenCV. [Note: This time you have to read your input from HDFS and run your application using YARN, and also you are not allowed to use calcHist function from OpenCV]. In this exercise, you have to find an image histogram using aggregation in Spark. Along with your code you also have to explain following:

1. Initialize your spark context for gray values.
2. How you design your sequence and combination operation functions.
3. Implement only for gray scale histogram.

Remember to include your image example in the submission.

Exercise 2: Using Apache Spark Mllib (5 points)

In this exercise you are going to develop a Sentiment Analysis tool using Apache Spark transformation and action functions. You can use dataset mentioned in
`http://thinknook.com/twitter-sentiment-analysis-training-corpus-dataset-2012-09-22/`.

The Twitter Sentiment Analysis Dataset `http://thinknook.com/wp-content/uploads/2012/09/Sentiment-Analysis-Dataset.zip` contains 1,578,627 classified tweets, each row is marked as 1 for positive sentiment and 0 for negative sentiment. You have to develop a complete pipeline for processing the text, i.e. only bag of words approach will not be sufficient, think about other options. Use Naive Bayes as base model, which according to the link gives performance of around 70% to 80% (classification accuracy).

1. Explain your pipeline by following standard machine learning approach.
2. Explain your preprocessing steps and how much each added step improves accuracy. You can use a table to list your results for each technique.
3. Develop a pipeline for textual data pre-processing and Naive Bayes model.
4. Report evaluation on prediction classification accuracy.
5. Effect of varying number of executors on the classification accuracy.

[Hint: You can use libraries provided by Apache Spark to accomplish your task. But if you want to add some new techniques you have to implement them in Spark framework. Also note your marks will also depend on the accuracy you achieve as compare to other students submission] [mllib reference `https://spark.apache.org/docs/latest/ml-guide.html`]

Exercise 3: Matrix Factorization with Coordinate Descent using Apache Spark (10 points)

0.1 Problem statement

You are familiar with solving recommender systems problem using a matrix factorization. In this task you have to parallelize coordinate descent method for matrix factorization using Apache Spark. Sample parallel coordinate descent algorithms are explained in resources mentioned in "Related reading material" section. The objective function optimizes a square loss.
Your document should have answers to the following questions (but not limited to).

1. Data division strategy?
2. How you parallelize your algorithm? Explain transformation and action functions
3. Any other optimization technique you used to optimized your task i.e. caching. How did it effect the performance ?
4. Present performance analysis i.e. speedup graph, [hint: have a look at the "Related reading material" how to calculate speedup].
5. Measure the affect of using varying number of workers on the convergence].

0.2 Dataset

For this exercise you will use movielens10m or movielens20m dataset available at `http://files.grouplens.org/datasets/movielens`. The movielens dataset is a rating prediction dataset with five star ratings (on a scale of 1-5). [Hint: you can start with a smaller movielens dataset i.e. movielens1m or movielens100k. But your final solution should be based on 10m or 20m dataset]

0.3 Related reading material

1. Parallel Coordinate Descent `http://www.caam.rice.edu/~optimization/L1/optseminar/Parallel%20BCD.pdf`
2. Matrix factorization with coordinate descent `http://www.cs.utexas.edu/~cjhsieh/icdm-pmf.pdf`
3. Parallel Speedup analysis `https://portal.tacc.utexas.edu/c/document_library/get_file?uuid=e05d457a-0fbf-424b-87ce-c96fc0077099&groupId=13601`
4. Spark launching configuration `https://spark.apache.org/docs/latest/submitting-applications.html`