CSE427S – Final Project

M. Neumann

Due Dates: see below (**no extensions!**)

Getting Started

Update your SVN repository, you will find additional materials for the *final project* in the folder `final_project`. Use the file `partners.txt` to indicate group work.

You have to choose one out of three possible projects and you can work in groups of 3 students. The workload of the projects is designed for group work! If you worked on your own on the homework assignments so far, I highly recommend to find a partner or join a group. Your learning experience and the probability to successfully finish the project will increase tremendously when working in a group!

Grading

Grades for all three projects will be assessed as follows:

- **milestone 1**: project and team selection, problem understanding, implementation concept, and routing (**10%**)
- **milestone 2**: data pre-processing and analysis (**15%**)
- **final submission**:
  - implementation (**30%**)
  - results (**20%**)
  - documentation of approach/implementation (**25%**)

  Approach and implementation, as well as the obtained results should be documented in the project report. This report should be readable for an informed outsider and it should not require the reader to look at or run any code. The grading of results (20%) and documentation (25%) also account for cleanliness, readability, and presentation of the report!

Due Dates and submission

All deadlines for this project are **HARD** deadlines (no automatic or non-automatic extensions!!).

- **milestone 1** (**10%**)
  - due: 04/21/2016 1pm

    - submission: short oral presentation (to instructor or TA)

- **milestone 2** (**15%**)
    - due: 04/26/2016 1pm
    - submission: SVN repository commit

- **final submission** (**75%**)
    - due: 05/05/2016 1pm
    - submission: SVN repository commit

Getting Help

First, please note that all three projects are BIG data analysis projects. That means they are a lot of implementation and debugging work and additionally, the mere execution time (once your programs are running) will also be BIG! So, keep this in mind when you plan your time management. Also, note that we cannot give an extension for the deadlines as the projects have to be graded before the final grading deadline!
If you need help for your project you can best get it from the **responsible TA(s) or instructor**:

- Project 1:
    - Kunyao (office hours: MON 6:30-8:30pm in Urbauer 114)
    - Marion (office hours: THU 11:30am-12:30pm in Jolley 403)

- Project 2:
    - Grace (office hours: WED 1-3pm in Bryan 405)
    - Israel (office hours: MON 1-3pm in Bryan 405)
    - Suo (office hours: SUN 1-3pm in Urbauer 216)

- Project 3:
    - Xuting (office hours: TUE 3-5pm in Urbauer 216)
    - Marion (office hours: THU 11:30am-12:30pm in Jolley 403)

You can also use **Piazza** for project related discussions. Make sure to use the tag for your project, so that the responsible TAs and other groups working on the same project can easily find your posts.

Project 1: MapReduce Approach to Collaborative Filtering for the Netflix Challenge

In this project your group will predict 100,000 movie ratings for users in a subset of the original NETFLIX data issued for the NETFLIX Prize. This challenge aimed at substantially improving the accuracy of predictions about how much someone is going to enjoy a movie based on their movie preferences. It was issued by the Netflix company and on September 21, 2009 a $1mio Grand Prize was awarded to the winning team.[1]

Goal: Analyze the NETFLIX data using PIG (or MAPREDUCE) and, based on the outcomes of this analysis, develop a feasible and efficient implementation of the collaborative filtering algorithm in MAPREDUCE. After computing the predicted ratings, evaluate those ratings by comparing them to the true ratings (gold standard). Note that MAPREDUCE is only required to find the k-most similar users or items. You do not need to use MAPREDUCE for the predictions and the evaluation. **This is a competition!** Part of the grades for the results (10%) will be assigned according to a ranking of the number and quality of all teams' predictions!

Project 2: Large-scale Text Processing and Sentiment Analysis in HIVE

Sentiment Analysis tries to assess the *emotion* or *mood* in a text document. Essentially it can be computed by comparing the set of words in each document to an existing dictionary of positive words, negative words, and neutral words and by analyzing the most frequent positive or negative expressions. Throughout the course we collected your **emotional** descriptions for each homework assignment. Now, you will be able to use this data for sentiment analysis!

Goal: Explore and use the text (pre-)processing features in HIVE to assess whether the homework reviews homework review data gathered throughout the semester were positive or negative. You will also investigate how the homeworks on the different parts of the course were perceived. Finally, you will predict the emotions of your own reviews and compare them to what your true emotion (when writing the review) was. The average scores on this task will be graded as a **competition** among all groups doing this project (10% of the grade will depend on this).

Project 3: k-means for Geo-location Clustering in SPARK

In this project you and your group will interactively get to know SPARK and use it to implement an iterative algorithm that solves the **clustering problem** in a parallel fashion. *Clustering* is the process of grouping a set of objects (or data points) into a set of k clusters of similar objects. Thus, objects that are similar should be in the same cluster and objects that are dissimilar should be in different clusters.
Clustering has many useful **applications** such as finding a group of consumers with common preferences, grouping documents based on the similarity of their contents, or finding spatial clusters of customers to improve logistics. More specific use cases are

- *Marketing*: given a large set of customer transactions, find customers with similar purchasing behaviors.
- *Document classification*: cluster web log data to discover groups of similar access patterns.
- *Logistics*: find the best locations for warehouses or shipping centers to minimize shipping times.

[1] You can read more about it here: `http://www.netflixprize.com/`.

We will approach the clustering problem by implementing the k-**means algorithm**. k-means is a distance-based method that iteratively updates the location of k cluster *centroids* until convergence. The main user-defined ingredients of the k-means algorithm are the distance function (often Euclidean distance) and the number of clusters k. This parameter needs to be set according to the application or problem domain. (There is no magic formula to set k.) In a nutshell, k-means groups the data by minimizing the sum of squared distances between the data points and their respective closest centroid.

Goal: Implement k-means in SPARK and use it for geo-location clustering on various datasets of spatial locations.