Introduction

Two possible projects:

- Shortest paths in a graph. On a social graph, shortest paths from a node gives you the “degrees of separation” from that node to the rest of the network. The longest shortest path (over all nodes) gives us the "degrees of separation" for the whole graph. This is also known as the diameter of the graph.

- Movie data mining. Starting with a collection of movie ratings from users, come up with interesting trends.

Shortest paths in a graph

Write a map-reduce solution to finding the length of the shortest path from a specified source vertex (given as a command line argument) in a graph. We are specifically interested in the length of the longest shortest path from the specified vertex. See Chapter 5 in the first reference for a map-reduce algorithm for shortest paths.

Computing the “degrees of separation” (aka the diameter of the graph) is extra credit for undergraduates and required for graduate students.

For input data you have two choices:

- Generate random graphs. This can be done using map-reduce. See the Last.fm example from class on how to generate data using map-reduce.

- Use real Facebook data. An anonymized sample data set is available (for 957k users!) from a research group in UC Irvine. See reference section for a link to the website.

References


2. Facebook data set from research group at University of California, Irvine.  
   http://odysseas.calit2.uci.edu/doku.php/public:online_social_networks
3. Review article on Facebook experiment on degrees of separation.  

4. Large collection of social network data sets from Stanford.  
http://snap.stanford.edu/data/

Movie data mining

Starting with a collection of movie ratings from users, come up with at least two interesting trends (for undergraduates) and three for graduate students.

Data

GroupLens Research has collected and made available rating data sets from the MovieLens web site (http://movielens.umn.edu). The data sets were collected over various periods of time, depending on the size of the set.

- MovieLens 100k - Consists of 100,000 ratings from 1000 users on 1700 movies.
- MovieLens 1M - Consists of 1 million ratings from 6000 users on 4000 movies.
- MovieLens 10M - Consists of 10 million ratings and 100,000 tag applications applied to 10,000 movies by 72,000 users.

See link in the reference section for accessing the data. Please make sure to read licensing information before using the data!

IMDB (The Internet Movie Database) allows the usage of its entire data set under some restrictions. Please make sure to read their license before downloading/using their data!

Insights

For example: what are $k$ most popular movies of all time, $k$ most popular movies from a particular year, $k$ most popular movies for a certain age group etc.

You can also attempt the movie rating problem but it is more complicated. See references below for more details.
References

1. Movie Lens data sets.
2. IMDB movie data
3. Wikipedia entry on the Netflix competition

Submission

• Submit the Eclipse project for your MapReduce program along with a ready to run jar file at the top level of your submission. Please do not submit your Hadoop install folder!

• Please don’t submit big data sets 🙁. However, you are required to submit a small data set (input and output) to illustrate your results 😊. The total submission size should be under 10 MB.

• Make sure to include a README with your experiences. This can be in text or PDF format.

Change directory to your assignment directory and execute the following command (on onyx) to submit the assignment.

submit amit cs430 p4

or

submit amit cs530 p4

This command will pick up all files in the current directory (as well as any subdirectories recursively) and time-stamp them before transferring the combined files to the instructor’s account.