Data Mining – Project – Due: March 23, 11:55 PM
(Forming a group is recommended. Groups of up to 4 are allowed.)

This is an exploratory project. You are encouraged to collect interesting data sets for an application domain that interests you. The more data you collect the better it is for finding interesting patterns. Your project should consist of the following three phases:

Data Collection (for the domain you like/are interested in)
Data Preprocessing (data cleaning and transformation into a useful form)
Data Mining (using algorithms you have seen so far)

At least one of these phases should be not trivial. For example the data collection and preprocessing phase could be non-trivial (e.g. the sites you use have some specific APIs that you need to use and/or the data needs special preprocessing).
Or the data mining process could be non-trivial. E.g. you collect a lot of data and then some algorithms such as those in Weka, being main memory algorithms, have a hard time to mine your data. In such a case, you should modify or recode those algorithms, or research available algorithms that can be more efficient for large amounts of data.

You should submit a report describing your work as well as give a short presentation (5 min) outlining the main points of your project.

You are expected to write the report in a research paper format. See
http://infolab.stanford.edu/~widom/paper-writing.html
on how a paper can be structured. The length of the report should approximately 10 pages.

Some interesting data/articles references are:

http://www.kaggle.com/c/titanic-gettingStarted
http://www.discoverycorpsinc.com/dancing-with-the-stats/
http://archive.ics.uci.edu/ml/datasets/YouTube+Comedy+Slam+Preference+Data#
http://www.sciencedirect.com/science/article/pii/S1877050916309036
http://archive.ics.uci.edu/ml/datasets/Farm+Ads
http://archive.ics.uci.edu/ml/datasets/YearPredictionMSD
http://labrosa.ee.columbia.edu/millionsong/pages/tasks-demos
http://archive.ics.uci.edu/ml/datasets/URL+Reputation
http://cseweb.ucsd.edu/~voelker/pubs/mal-url-icml09.pdf
http://www.yelp.ca/academic_dataset
https://grouplens.org/datasets/movielens/
http://2013.msrconf.org/challenge.php
http://2015.msrconf.org/challenge.php
http://2017.msrconf.org/#/challenge
https://dnc1994.com/2016/05/rank-10-percent-in-first-kaggle-competition-en/

Also, the practical book:
Programming Collective Intelligence by Toby Segaran
contains a wealth of ideas on how to find data and how to mine them.