CSC3730 Project

Schedule:

Form a team of 3-4 students

Final presentation ppt and project report (the week before final exam)

What to do:

The task for the project is to implement a pipeline of data analytics techniques including data collection, data cleaning and knowledge discovery for a suitable (real world or prototype) application domain.

Choosing an application:

Pick topics of interests, related to the course, of suitable difficulty.

Dataset

http://webscope.sandbox.yahoo.com/ Yahoo WebScope

http://www.data.gov/ U.S. Government's open data

http://www.freebase.com/ Freebase

http://www.yelp.com/developers/documentation/ Yelp

https://developers.google.com/apis-explorer/#p/ Numerous APIs from Google (e.g., Maps, Freebase, YouTube, etc.)

http://developer.trulia.com Trulia
http://www.zillow.com/howto/api/APIOverview.htm Zillow real estate listing sites

Numerous graph datasets (large and small):

https://snap.stanford.edu SNAP

http://konect.uni-koblenz.de/networks/ Konect

List of lists of datasets for recommendations

https://gist.github.com/entaroadun/1653794/

http://the.echonest.com/press-release/the-echo-nest-and-columbia-university-announce-million-song-dataset/ Million song dataset by Echo Nest. It contains not only the basic information of songs (artist, genre, year, length etc), but also some musical features(like tempo, pitch, key, brightness)

Movies data:

http://developer.rottentomatoes.com Rotten Tomatoes
http://stackoverflow.com/questions/1966503/does-imdb-provide-an-api/ IMDB

http://grouplens.org/datasets/movielens/

http://archive.ics.uci.edu/ml/datasets.html/ UCI also has a collection of links to various datasets sorted for various tasks (Classification, Regression, etc)

http://aws.amazon.com/datasets/ Amazon AWS Public Data Sets

http://www.sigkdd.org/kddcup/index.php/ KDD Cup: annual competition in data mining, like Kaggle

Final presentation [about 10 minutes per team]:

1. Goal and Motivation
    a. Brief overview of the problem
    b. Data: where you got it? Size and storage? Format?
2. Your approaches: Data Analysis Algorithm
3. Your experimental results and Conclusions: What you learn?
4. Presentation delivery

Final project report:

1. **Writeup**: Describe in depth the problem you would like to solve, your approach and your discoveries/insights/experiments, etc.

2. **Software**: packaging, documentation, and portability. The goal is to provide enough material, so that other people can use it.

3. **Grading scheme & Submission instructions**

- Writeup
    - Introduction – Motivation
    - Problem definition
    - Description of your approaches: data analysis algorithms
    - Experiments/Conclusions

- Software: create a zip file of the code and a short **README.txt** file. This file should describe the package in a few paragraphs, how to install it, how to use it, and how to run a demo (if any).

Some examples:

1. Handwritten digit recognition

http://yann.lecun.com/exdb/lenet/

Goal: Handwritten digit recognition.

Motivation: automatic zipcode recognition.

Survey: Yann's paper: Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition

The model is multi-layer neural network. Describe what it is, advantages/disadvantages. What is the classification accuracy of the model?

2. Text Categorization

Goal: Classify online news stories to different topics, Economics, Sports, etc.

Survey: David Paper: David Lewis et.al. RCV1: A New Benchmark Collection for Text Categorization Research

Models: compare different classifiers including knn, logistic regression, SVM etc.

Data:

https://archive.ics.uci.edu/ml/datasets/Reuters+RCV1+RCV2+Multilingual,+Multiview+Text+Categorization+Test+collection