W4249 Spring 2016 Applied Data Science

Project 4 Mining Relational data

In this project, we will explore relational data where the data were collected to have information on a set of individuals and on their relations. Part of the big data hype has been fueled by the availability of the huge amount of online human generated information on social networks, individual rating/preferences on products, opinions and views shared via social media (such as facebook, twitter, etc). Topics in this area include

- Collective filtering
- Clustering/community detection
- Network analysis
- Identity resolution

Challenge: mining Amazon movie reviews

For this project, you will carry out data mining on a large set of Amazon moive reviews. You can do one of the following for this project

- Collective filering: movie recommendation engine for users.
- Community detection for movies based on movie reviews.
- Other association mining

We are interested in what insights you can derive about movies and movie reviews from this data set.

For this project, we will give tutorials on the following topics:

- Data processing of this movie review data set.
- Collective filtering.
- Network analysis including community detection.

For presentation, the team should present their findings from this project such as,

- how to make better movie recommendations
- interesting insights about movies and movie reviews

Repositary requirement

Each team should organize the project repo on GitHub according to the structure of the starter codes.

```
proj/
├──doc/
├──figs/
├──lib/
├──output/
├── README
```

- The data is too big to be hosted on GitHub.
- The `doc` folder should have documentations for this project, presentation files and other supporting materials.
- The `figs` folder contains figure files produced during the project and running of the codes.
- The `lib` folder contain computation codes for your data analysis. Make sure your README.md is informative about what are the programs found in this folder.
- The `output` folder is the holding place for intermediate and final computational results.

The root README.md should contain an abstract of your findings and you should include a contribution statement as well.

Project learning support

We will give a tutorial in class and having live discussion and brainstorm sessions. The instruction team will join team discussions during class and online.

- week 1: data processing and collective filtering
- week 2: network analysis

Suggested team workflow

This is a relatively short project. We only have about two weeks of working time.

1. [wk1] Week 1 is the **data processing** week. Read data description, **project requirement**, and browse data and the starter codes and brainstorm about what to do.
2. [wk1] As a team, download the data, discuss data management need of this project. Due to the size of this data set, you would need to have codes coming to the data instead constantly sharing processed data among yourselves. Try adapt the starter codes to a *subset* of the data set to get a sense of computational burden of this project.
3. [wk2] Based on outcomes from week 1 brainstorm sessions, Carry out the data analysis.
4. [wk2] Week 2 is the **data mining** week. The data mining is likely to take a lot of time. Start early.

Working together

- Setup a GitHub project folder with everyone listed as contributor. Everyone clones the project locally and create a local branch.
- The data is too big to be stored on GitHub. You can fork the repo to a local folder for this class or all your data science projects and have the data stored in "../data/" from root. Your project folder may look like

```
local proj/
├──data/
├──GitHub/
├── README
```

- The team can work with subgroups of 2-3 work together more frequently than the entire team. However, everyone should check in regularly on group discussion online and changes in the GitHub folder.