STAT250 Final Project
—Video Recommendation

Yu Pei
SID 999438479

1 Background

The ability to access data through the web is an essential skill needed in job market. And Natural Language Processing(NLP) has a wide range of applications given that we have a huge amount of text data on the web. Like using wikipedia or IMDB's information to build video recommendation. After actually get the data, we can perform many statistical techniques to analyse the informations, like using TFIDF or word2vec(based on a paper Efcient Estimation of Word Representations in Vector Space). With this project, I will be able to get familiar with scrapping data from the internet, processing data in different formats, word analysis and some new statistical learning methods.

The outlined task is : we have a json file with data on 474 youtube videos in the television and movie category. What I wanted to do is the following:

- Assign tags as appropriate for famous actors (e.g. Jennifer Lawrence, Kate Walsh), movies (e.g. "The Hunger Games"), and television shows (e.g. "Game of Thrones") to each youtube video.
- Create your own related video algorithm: Using the information given, as well as any tags you might assign, create an algorithm than takes on video as input and generates three related videos from the set. Using your algorithm, write a file that shows the related videos for every video in the set.

2 Plan

I am really not too worry about the statistical analysis part, since some people have already figure out some useful methods to do it. So my main focus is getting the data and be able to walk through this procedure.

- this is a very small data set and purely agnostic methods are unlikely to work, so I need to use an external source of data to identify possibly important actors, films and television shows and then assign them. One source that could be useful is, dbpedia, or use the youtube API. I want to automate the fetching of this data.
- Create my own related video algorithm: Using the information given, create an algorithm than takes on video as input and generates three related videos from the data. Using my algorithm, write a file that shows the related videos for every video in the set.

Then finally I will review how well this algorithm do in terms of the relavence of the recommended video, although it might be subjective. Hopefully with the guidence of Instructor Duncan, I can do this within a month.