Student Notes

# Necessary Libraries and Packages

1. Python 2.7
2. pandas - http://pandas.pydata.org/
3. numpy - http://www.numpy.org/
4. scikit-learn - <http://scikit-learn.org/stable/>
5. scipy - <https://www.scipy.org/>

# Demos provided:

## Sentiment analysis

1. Task: given a review, predict it is a positive one or negative one?
2. Dataset: IMDB movie review: <http://ai.stanford.edu/~amaas/data/sentiment/>

## Salary prediction

1. Task: given the details about an employee, can you predict the salary?
2. Dataset: employee dataset: http://archive.ics.uci.edu/ml/datasets/Adult

## Business Reviews Prediction

1. Task: given the details about business reviews and user data, predict the ratings for businesses by different users?
2. Dataset: Yelp Reviews dataset: https://www.yelp.com/dataset\_challenge/dataset

# References:

## Features

1. N-gram language model: <https://en.wikipedia.org/wiki/N-gram>
2. Bag-of-words model: <https://en.wikipedia.org/wiki/Bag-of-words_model>
3. tf-idf value: <https://en.wikipedia.org/wiki/Tf%E2%80%93idf>
4. Advanced topic:
   1. word embeddings – word2vec: <https://arxiv.org/pdf/1301.3781.pdf>
   2. word embeddings – glove: <http://nlp.stanford.edu/pubs/glove.pdf>
   3. document representations - NAACL 2015 workshop: <http://nlp.stanford.edu/~manning/talks/NAACL2015-VSM-Compositional-Deep-Learning.pdf>
5. Feature engineering: http://www.cs.princeton.edu/courses/archive/spring10/cos424/slides/18-feat.pdf

## Algorithms

1. Linear Regression: <http://people.duke.edu/~rnau/notes_on_linear_regression_analysis--robert_nau.pdf>
2. K-Means Clustering: <http://cs229.stanford.edu/notes/cs229-notes7a.pdf>
3. Logistic regression: <https://en.wikipedia.org/wiki/Logistic_regression>
4. Support vector machine: <https://en.wikipedia.org/wiki/Support_vector_machine>
5. Random forest: <https://en.wikipedia.org/wiki/Random_forest>