**Data Analytics - Capstone Project**

**R OR Python**

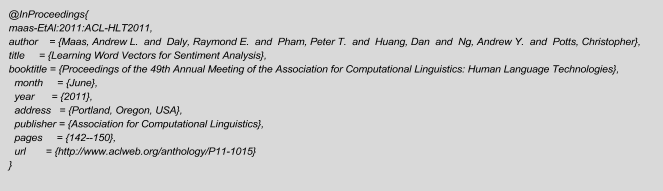
What is Capstone Project?

The Capstone project is the culminating assignment that will allow you to have an integrated experience of the program. The approach to this project is to think, define, design, code, test and tune your solution, in such a way that you apply all aspects of the data analytics process.

**Dataset:** IMDB Dataset

Link: <http://ai.stanford.edu/~amaas/data/sentiment/aclImdb_v1.tar.gz>

**Citation:**

**

**Description**:

This dataset contains 25,000 highly polar movie reviews for training, and 25,000 for testing. There is additional unlabeled data for use as well.

Real world is filled with text data and is usually messy hence cleaning and handling text is an important step towards making smarter Machine Learning algorithms.

IMDB dataset is one of such usual messy dataset which hides a lot of information under the hood which is awaiting to be discovered. IMDB holds sentiments of bunch of people which can give away a lot of insights about the movie itself, understanding such sentiments can boost any production house income. The derived information can be very useful in building customized recommendations for users based on their reviews. And these models can be used to find unhappy customers, the reason for dissatisfaction and so on. Sentiment analysis is great tool for the same.

The dataset contains positive and negative reviews (Binary Classification) to train the sentiment analysis model and predict the outcome for unseen/test data.

Ironically despite the abundance of data, most of the real-world data is not labelled, or at least not labelled to exactly suit the requirements. Active learning can be used to divide this data into clusters, analyses each cluster and provide suitable labels to each cluster once a good cluster model is achieved. The clustered data can then be used as labelled data for further analysis.

So here we have come with an interesting problem for movie buffs. The following project consists of IMdB data set where there are reviews of each movie. Here, we need to classify the reviews into positive and negative sentiments thus classifying the movies having the most good reviews. You should also try to find out the keywords in reviews which are defining the overall reviews of the movie. This will help in knowing the movies or genre of movies performing best and later on you can check out those movies if you have not. Such kind of problems are generally handled by Data Scientists working in IMDB, movie directors, movie review writers, and Bookmyshow to explore the people perspective on different movies.

Project Approach

Phase 1 Cleaning (Score 30 + 10 Extra credits):

**Each student must solve**

1. Read the labelled data from respective folders (pos & neg) and store in data-frames (eg: train\_df & test\_df) with suitable structure

Hint: columns = [review, label]

1. Remove stop words from the data
2. Apply Stemming and Lemmatization

Extra Credits

1. Apply feature selection to select most important words/features and drop others (Bonus)

Phase 2 Exploration (Score 30 + 20 Extra credits):

**Each student must solve**

1. For labelled data, find the most common words associated with each category (positive & negative)
2. Discover the lowest frequency and highest frequency words
3. Classify each words along with its frequency with the review it is connected to and also merging the same with that particular movie.

Extra Credits

1. Read unlabeled data from respective folder (unsup) and store in unsup\_df
2. Create a cluster to separate positive and negative words (bonus) using k-means algorithm

Phase 3 Visualization (Score 30 + 10 Extra credits):

**Each student must solve**

1. Create a wordcloud with positive and negative words after cleansing
2. Visualise the positive and negative words distribution (Hint: Histogram)
3. Visualise the movies with the high frequency keywords (Hint: Scatter plot)

Extra Credits

1. Repeat visualization step 1 & 2 after feature selection and note the impact (Bonus)

Phase 4 Hypothesis creation and testing (Score 10):

**Each student must solve**

1. Create Hypothesis involving relationships between dependent and independent variables using parametric/non parametric tests for various machine learning algorithms such as k-means clustering,  classification algorithms .

Phase 5 Model Building (Score 80 + 180 Extra credits):

**Each student must solve**

1. Supervised Learning: Build a sentiment analysis model to predict positive and negative classes **(Score 40)**
2. Unsupervised Learning: Build a clustering model consisting of 2 clusters based on positive and negative reviews **(Score 40)**

Extra Credits

1. Supervised Learning: Compare the performance of different machine learning models, at least 2 **(Score 40)**
2. Unsupervised Learning: Compare the performance of different models, at least 2 **(Score 40)**
3. Divide the data into 4 clusters to enable finding more classes. Analyse each cluster and try to find correct label for the new cluster. Repeat clustering until 4 new labels can be found, other than the original labels (positive and negative) **(Score 50)**
4. Active Learning: Cluster the training dataset and try and find the genre. Manually annotate the cluster and then try to find the labels in the new testing dataset. **(Score 50)**