**Team E Progress Report 1**

**Achievements**:

Based on the feedback on the project proposal, we found our initial proposal (house price prediction) lacked innovation and interest from other well-studied projects and literature reviews and also we found ourselves lacked motivation upon the original topic. Therefore, we decided to change our topic to a more motivated one, and the following is our new project proposal:

Motivation:

Youtube as one of the biggest video-sharing website has provided many business and advertising opportunities. Over 400 hours of videos are uploaded on Youtube in 1 minute. There are many traditional media platforms as well as individual online celebrities joining the game. However, only a few of the videos became a hit. Besides having good content, there are many factors that contribute to large views on YouTube. Even many tutorials can be found online to teach people how to make a trending video. However, many of them are made based on personal experience without statistical analysis. We want to analyze this problem using data science technology to reveal things that really make your video a hit.

Goal:

1. Analyze factors that highly affect Youtube video popularity.
2. Find out the top ten rules and attributes that contribute to high views on YouTube.
3. Help individual YouTubers as well as traditional social media to most effectively convey their opinions and deliver their values.

Dataset:

The data comes from the Kaggle dataset: Trending Youtube Video Statistics collected using Youtube API, which is a record of several months of daily top trending videos on Youtube (US region). The size of the dataset is 40,000, containing 16 attributes including both numerical and categorical features. Sample numerical features are views counts, likes counts, dislikes counts, publish time, comment counts. Sample categorical features are video title, category, channel title, tag words, video descriptions, thumbnails.

Technical Approaches:

1. PCA Feature Selection
2. Text Analytics: Naive Bayes, Parameterization with Bag-of-words, Brown Clustering, Word Embeddings, Association Rule
3. Visual Recognition on thumbnail pictures of video

**Challenges:**

* Problem: Difficulties in feature selection and new feature creation.
* Solution: Try and test for many times, find some relevant research.
* Problem: When we are trying to find the relationship between popularity and tags, there will be too many different tags.
* Solution: Merge the same meaning, only use the data with the tag that appear frequently
* Problem: The data may be unbalanced for some of the “big channel” such as AMC may have inborn more popularity than the regular channel and all of the data are mixed.
* Solution: Using the data only in the regular channel(such as select the data with the channel only named by people’s “name”)
* Problem: No experience in NLP and image recognition
* Solution: Use existed coding package, find some relevant research