# Final Project Proposal

by Kate Bäumli, Corey Hulse, Amanraj Mahal, John Sigmon

April 9, 2018

## 1 Problem Statement

The project we are proposing is an exploration of two data sets based on TED Talk[1] transcripts. The first data set is from a Kaggle competition and comes with transcripts and various meta-data. The second data set is located on a personal GitHub repo and contains only transcripts, but they are cleaned. The two data sets almost completely overlap. The competition is also exploratory in nature, however we plan to answer different questions than the ones posed on the Kaggle page. Our specific goals are listed below.

## 2 Problem Approach

Our first task is to combine the data sets. We consider this a preliminary step. Afterwards, we plan to use TextBlob and tf-idf[2] vectors to do sentiment analysis on the transcripts and analyze them for patterns and correlations. Next, we will divide the data set based on sex and analyze the transcripts for differences in speech patterns. We will also use these analyses to see if it is feasible to predict how popular a talk will be. Predicting popularity may allow suggestive improvement of future talks based on correlation between features and talk popularity. Two potential ways of quantifying popularity are via number of video likes or via supplied meta-data tags such as 'inspiring'. Finally, we will attempt to develop a model that can generate speaker biographies or entire talks, depending on time constraints. It is our belief that there is a relatively substantial and worthwhile story to be found by exploring these ideas.

## 3 Related Work

There exist several well prepared kernels on the Kaggle page for the data set. As far as we know, there is no overlap between our proposed project and the kernels that have been posted here.

Some of the things that have already been explored are:

- Correlation between length of talk and number of views
- Correlation between talking speed and number of views
- Correlations between talk 'tags' and number of views
- Occupation of the speaker and number of views
- Correlation between day of the week of the talk and number of views

We feel that the topics we plan to explore will largely diverge from this work, and that the results will be worthwhile.

[1] by the way, this stands for Technology, Entertainment, and Design

[2] Term Frequency-Inverse Document Frequency is a numerical statistic used to reflect the importance of a word in a collection of words