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 has transcripts and various meta-data. The second data set is located on GitHub and contains only transcripts, but they are cleaned (i.e. no punctuation.) The posted competition is also exploratory in nature, although we plan to explore different questions than the ones posed on the Kaggle page. Our specific goals are spelled out below.

2 Problem Approach

Our first 'thing on the agenda' is to combine the two data sets. We plan to use TextBlob to do sentiment analysis on the transcripts and analyze them for patterns and correlations with the provided meta data. We plan to divide the data set based on sex and analyze the transcripts for patterns in speech such as most frequently used words. We will also use these analyses to see if there is a way to predict how popular a talk will be, either via number of likes, or via supplied meta-data tags like "inspiring". Depending on our time constraints, we may attempt to develop a model that can generate speaker biographies or entire talks.

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. — summarize the kernels.

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