# Predicting TED Talk Categories

Laura Borton

## Project Design:

TED Talks are 18-minute talks about “ideas worth spreading.” The themes have grown from Technology, Entertainment, and Design (TED), to include Business, Science, and Global Issues.

The TED.com website has titles, summaries, and transcripts for each talk. Each talk usually falls into one or more of the six general categories above, and also into more specific categories. These categories are listed as Tags on each talk’s webpage. Could one of the six general categories be predicted by using Natural Language Processing (NLP) on its summary?

## Tools and Data:

An Excel spreadsheet with the summary and tags for over 2200 talks from 2006 - 2018 was used as the database. If a talk was not tagged with Technology, Entertainment, Design, Business, Science, and Global Issues, it was not used in the prediction data set. This was the case for about 250 talks. If a talk had more than one tag of a general topic, the first one listed on the spreadsheet was used, as the tags were most likely ranked in order of importance.

Python was used for loading, preprocessing, modeling, and visualizing the data.

## Preprocessing:

After the data was loaded into Python, common words, or stop words, were removed from each talk summary. Gensim, a popular Python library for NLP, has a preprocessing feature which lowercases all the words and tokenizes them. Tokenization is a technique that breaks a sentence or a stream of text into words. Next, bigrams and trigrams, combinations of every word with its preceding and/or following word(s), were added to the list of tokens. Lastly, all the tokens were lemmatized, keeping only verbs, adverbs, nouns, and adjectives.

Next, the features were converted to a matrix of token counts with CountVectorizer or a matrix of Term Frequency - Inverse Document Frequency (TF-IDF) features. TF-IDF gave better results through the entire workflow. The final matrix shape was 1957 x 10355. There were too many features for the number of observations so the features were reduced using Latent Semantic Analysis (LSA). To keep a ration of about 100 rows to 1 feature, the dimensions were reduced to 20.

## Predictive Modeling:

The reduced matrix was split into test/train sets with a 70/30 ratio. The target variable was one of the six general categories. Numerous classification models were run, specifically: Linear SVM, SVM - linear kernel, SVM - rbf kernel, random forest, and gradient boosting. The same algorithms, and Naive Bayes - Gaussian, were run with the CountVectorizer input, but the accuracy results were lower. Linear SVM with the Tf-IDF input produced the highest accuracy of 55%. Although 55% accuracy is low, it is better than a random guess of 1 out of 6, or 17%.

## Clustering:

Looking at the tags for some of the TED talks, it seemed that some could have been

categorized differently. The clusters could encompass more than just the six general categories, but could be more general than the 200 specific topics on the TED website. Plotting the Sum of Squared Errors (SSE) vs the K-Means number of clusters showed that 25 clusters would be the optimal number of groups to have. After looking at the summaries that fell into each group, the results were mixed. Some of the talks were categorized appropriately--for example, business/finance and activism. Other categories had only a common word, like 'world', as their common theme.

Latent Dirichlet Allocation (LDA) was also tried as a topic modeling technique. Beginning with 25 clusters as before, the talks were grouped. Many of the 25 were had overlapping topics so the number of clusters was reduced to 6, in hopes of clustering the talks by the general topics originally selected. These also had many overlapping topics so the number of clusters was reduced as to have no overlap. With three clusters, the topics seemed to be incoherent.

Summary:

Predicting TED talk categories based on a short one- or two-sentence summary gives mediocre results, at best. Clustering the talks based on the summaries does worse. However, if the entire transcript were used instead, the results would probably be better. As an example, a Jamie Oliver talk on obesity was categorized under both Global Issues and Business. This seemed surprising based on his summary:

*"Sharing powerful stories from his anti-obesity project in Huntington, West Virginia -- and a shocking image of the sugar we eat -- TED Prize winner Jamie Oliver makes the case for an all-out assault on our ignorance of food."*

However, when reading the transcript, he mentions, "Mexico, Australia, Germany, India, China, all have massive problems of obesity and bad health." This is why this is a Global Issue talk. In addition, large food corporations and their influence are discussed, explaining the Business category.

There is an opportunity to improve the prediction accuracy and the topic clustering by using the transcripts instead of the short summaries.