**Project Proposal for Predicting the popularity of TED Talk with Data Mining**

Shijie Liu

Computer Science Department

Hood College, Frederick, MD 21701, USA

sl23@hood.edu

**ABSTRACT**

TED Talk is a media program that has videos posted on their website and YouTube channel, as well as the Podcast station. In this project, I would like to present the relationship between talk topic and its popularity with visualization technique. I am planning to build a classification model that could predict the popularity of a future TED Talk by applying Natural Language Processing algorithms.

**Keywords**

Visualization technique, Classification, popularity, Natural Language Processing

# INTRODUCTION

TED Talks are being posted online under the slogan “ideas worth spreading”, it includes talks on scientific, cultural, and academic topics, and the speakers also widely spread with different roles, such as, scientists, education researchers, businessmen, artists, etc.[1] Until October 2018, there are approximately 2,900 TED Talks freely available on the TED website [2]. Like Charlie said, TED Talks has become one of the Internet’s most powerful platform because they are spreading ideas through the stories of remarkable people and they could be supported world widely with different languages for transcripts [3].

Today, the speed of data growth is extremely fast. According to IBM, there is 2.5 quintillion bytes of data created every day [4]. Everything is in a format of data, as well as TED Talks. There is a dataset being found on Kaggle that related with TED Talk, it includes a number of attributes and the plain transcript for each talk. Why TED Talk is so popular and what things people like to hear from TED Talk? Getting all those questions answered is not a trivial task. Especially, when the data is large and multidimensional, it is practically impossible to get a potentially interesting and actionable insight without the help of suitably designed machine learning algorithms [5]. More important, machine learning is everywhere in today’s Natural Language Processing, the goal of deep learning is to explore how computers can take advantage of data to develop features and representations appropriate for complex interpretation tasks [6].

The purpose of this project is to apply data mining techniques on the TED Talk dataset to get a better understanding of what makes a popular TED Talk, what are the most popular topics people would like to discuss about. And building a prediction model that could predict the popularity of a TED talk when transcripts is given.

# BACKGROUND

# TED Talk is my favorite program because of its diverse contents that cross different fields and creative ideas. One day, when I visit the website of TED Talk, there was a question pop-up automatically asking me “what interests you”? (Technology, Science, Innovation, and Humanity, etc.) Right after, another question came up asking me “what you’re looking for”? (Professional Growth, Inspiration or motivation, and Smart entertainment, etc.)

# After selecting my interested topic and idea, there is one recommended talk pop-up to me along with a sentence saying, “The idea offers ‘professional growth’ and matches your interest in ‘innovation’”. Frankly, I was so surprised that I can just let the website know my interest and then the website will give me a recommended video that matches my interest. This recommend feature reminds me the concept that we can extract information with algorithms from large dataset and present it in a comprehensible way. Consequently, I strongly want to apply visualization technique to a dataset to visualize important information, as well as train a classification model that could learn from the dataset and predict an output based on input.

# OBJECTIVES

Here are three objectives I want to achieve with TED Talk Dataset:

* Visualize what makes a TED Talk popular
* Visualize what topics people like to discuss about
* Visualize what are the most popular words being used by speakers
* Predict the popularity of a TED Talk when transcripts are given.

# RELATED WORK

## Data Preparation:

### Data Selection:

### The TED Talk datasets were downloaded from Kaggle, according to the uploader, these datasets contain information about all audio-video recordings of TED Talks uploaded to the official TED.com website until September 21st, 2017. The TED Talk main dataset contains 17 columns that including number of views, number of comments, descriptions, speakers and titles, etc. The TED Talk transcripts dataset contains 2 columns that including the URL and the available transcripts [7].

### Data Construction:

While checking the Data frame shape, it was noticed that both dataset has column ‘url’, planning to join two datasets into one data frame by using ‘url’ as a key.

### Data Cleaning:

1. Remove missing values: The shape of TED main dataset is 2550 \* 17, while the shape of TED transcripts dataset is 2467 \* 2. After joining two datasets, the shape of combined data frame became 2553 \* 18. Then, it was decided to use pandas to check which column has missing value, and it turns out column ‘transcript’ has 86 missing values.

2. Remove useless columns: While analyzing the columns and its values, it seems like the column ‘name’ is a combination of column ‘main\_speakear’ and ‘title’.

3. Remove unrelated rows: While analyzing the value for column ‘event’, it was noticed that there are a lot of event not belongs to TED Talk Category. For example, ‘AORN Congress’, ‘BBC TV’, ‘DICE Summit 2010’, etc.

### Data formatting:

While analyzing the data types for each column, it was noticed that the column ‘film\_date’ is int64, which is being represented as UNIX Epoch Time. In order to make more sense, conversion shall be applied to this column, from int64 to date time format.

## Data Modeling:

While dealing with the transcripts of each talk, words representations are a critical component of many natural language processing systems. A word representation is a mathematical object associated with each word, often a vector [9]. It is common to capture the rich relational structure of the lexicon. Vector-space models do much better in this regard because they encode continuous similarities between words as distance or angle between word vectors in a high-dimensional space [8]. In other words, Vector-space models represent word meanings with vectors that capture semantic and syntactic information of words [10].

### Model Building Techniques:

* Data Tuning with CountVectorizer:
* Term Frequency-Inverse Document Frequency (TF-IDF)
* Multinomial Naive Bayes

### Model Evaluation Metrics:

* Confusion matrix
* Accuracy score

# PROJECT TIMELINE

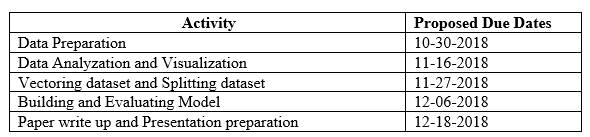


Figure 1. Proposed Due Date for break-down tasks

.

# REFERENCES

1. Wikipedia. 2018. WikipediA: TED (conference). Retrieved from https://www.wikipedia.org.
2. TED. 2018. TED Talks. Retrieved from https://www.ted.com/talks.
3. Charlie Rose. 2015. CBS: TED Talks. (April 2015). Retrieved October 20, 2018 from https://www.cbsnews.com/news/ted-talks-60-minutes-charlie-rose/
4. Ralph Jacobson. 2013. IBM. (April 2013) Retrieved October 20, 2018 from https://www.ibm.com/blogs/insights-on-business/consumer-products/2-5-quintillion-bytes-of-data-created-every-day-how-does-cpg-retail-manage-it/
5. Vasant Dhar. 2013. Data science and prediction. Commun. ACM 56, 12 (December 2013), 64-73. DOI: https://doi.org/10.1145/2500499
6. Richard Socher, Yoshua Bengio, and Christopher D. Manning. 2012. Deep learning for NLP (without magic). In Tutorial Abstracts of ACL 2012 (ACL '12). Association for Computational Linguistics, Stroudsburg, PA, USA, 5-5.
7. Rounak Banik. 2017. TED Talk Dataset. (2017) Retrieved October 20, 2018 from https://www.kaggle.com/rounakbanik/ted-talks/home
8. Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1 (HLT '11), Vol. 1. Association for Computational Linguistics, Stroudsburg, PA, USA, 142-150.
9. Joseph Turian, Lev Ratinov, and Yoshua Bengio. 2010. Word representations: a simple and general method for semi-supervised learning. In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL '10). Association for Computational Linguistics, Stroudsburg, PA, USA, 384-394.
10. Eric H. Huang, Richard Socher, Christopher D. Manning, and Andrew Y. Ng. 2012. Improving word representations via global context and multiple word prototypes. In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1 (ACL '12), Vol. 1. Association for Computational Linguistics, Stroudsburg, PA, USA, 873-882.