Predict the Popularity of a TED Talk

TED Ideas worth spreading

Liu, Shijie | Dr. Liu, Xinlian | Hood College of Frederick Maryland

Abstract

TED (Technology, Entertainment, and Design) Talks are being posted by TED Conference LLC for free on their website and YouTube Channel under the slogan of “ideas worth spreading”. TED Talk offers a wide range of topics within the research and practice of science and culture, and often through storytelling. It has a variety of presenters as well, such as Writer, Researcher, and Scientist. TED also gives people the right to raise their own local TEDx Event. How could TED filter and determine which talk shall be published? What an organizer could do to make his/her talk more popular? In this project, I would like to apply Exploratory Data Analysis on all the available attributes and help people to have a better understanding of what are the variables that could affect the popularity of a TED Talk. In addition, applying Natural Language Processing technique on the TED Talk transcript, and then train a classification model with algorithms that could predict the popularity of TED Talk.

KEYWORDS: Exploratory Data Analysis, Visualization, Classification, Natural Language Processing

Techniques

Classification Algorithms:

- Logistic Regression
- Random Forest Classifier
- Decision Tree Classifier
- Multinomial Naïve Bayes

Classification Model Evaluation Metrics:

- Prediction Accuracy
- Confusion Matrix
- Classification Report
- Cross Validation

Sample Data

TED Main Dataset- Sample Data:

comments	description	duration	event	film_date	languages	main_speaker	name	num_speaker	published_date	ratings	related_talks	speaker_occupation	tags	title	url	views
4553	Sir Ken Robinson makes an entertaining and pro...	1164	TED2006	1140825600	60	Ken Robinson	Ken Robinson: Do schools kill creativity?	1	1151367060	[{'id': 7, 'name': 'Funny', 'count': 19645}, {...	[{'id': 865, 'hero': 'https://pe.tedcdn.com/im...	Author/educator	['children', 'creativity', 'culture', 'dance',...	Do schools kill creativity?	https://www.ted.com/talks/ken_robinson_says_sc...	47227110

TED Transcript Dataset- Sample Data:

transcript	url
Good morning. How are you?(Laughter)It's been great, hasn't it? I've been blown away by the whole thing. In fact, I'm leaving.(Laughter)There have been three themes running through the conference which are relevant to what I want to talk about. One is the extraordinary evidence of human creativity in all of the presentations that we've had and in all of the people here. Just the variety of it and the range of it. The second is that it's put us in a place where we have no idea what's going to happen, in terms of the future. No idea how this may play out.I have an interest in education. Actually, what I find is everybody has an interest in education. Don't you? I find this very interesting. If you're at a dinner party, and you say you work in education — Actually, you're not often at dinner parties, frankly,(Laughter)If you work in education, you're not asked.(Laughter)And you're never asked back, curiously. That's strange to me. But if you are, and you say to somebody, you know, they say, "What do you do?" and you say you work in education, you can see the blood run from their face. They're like, "Oh my God," you know, "Why me?"(Laughter)"My one night out all week."(Laughter)But if you ask about their education, they pin you to the wall. Because it's one of those things that ….	https://www.ted.com/talks/ken_robinson_says_schools_kill_creativity

Data Preprocessing

```
1 tedMain.ratings[0]
```

```
"[{'id': 7, 'name': 'Funny', 'count': 19645}, {'id': 1, 'name': 'Beautiful', 'count': 4573}, {'id': 9, 'name': 'Ingenious', 'co
unt': 6073}, {'id': 3, 'name': 'Courageous', 'count': 3253}, {'id': 11, 'name': 'Longwinded', 'count': 387}, {'id': 2, 'name':
'Confusing', 'count': 242}, {'id': 8, 'name': 'Informative', 'count': 7346}, {'id': 22, 'name': 'Fascinating', 'count': 10581},
{'id': 21, 'name': 'Unconvincing', 'count': 300}, {'id': 24, 'name': 'Persuasive', 'count': 10704}, {'id': 23, 'name': 'Jaw-dro
pping', 'count': 4439}, {'id': 25, 'name': 'OK', 'count': 1174}, {'id': 26, 'name': 'Obnoxious', 'count': 209}, {'id': 10, 'nam
e': 'Inspiring', 'count': 24924}]"
```

ratings	Informative	Unconvincing	Jaw-dropping	Longwinded	Funny	Fascinating	Inspiring	Obnoxious	Beautiful	Confusing	OK	Ingenious	Persuasive	Courageous
93850	7346	300	4439	387	19645	10581	24924	209	4573	242	1174	6073	10704	3253

Figure 1: Convert ratings from string object to a matrix table

```
1 tedMain.loc[0:5,['film_date','published_date']]
```

	film_date	published_date
0	1140825600	1151367060
1	1140825600	1151367060
2	1140739200	1151367060
3	1140912000	1151367060
4	1140566400	1151440680
5	1138838400	1151440680

	film_date	published_date
0	2006-02-24	2006-06-26
1	2006-02-24	2006-06-26
2	2006-02-23	2006-06-26
3	2006-02-25	2006-06-26
4	2006-02-21	2006-06-27
5	2006-02-01	2006-06-27

Figure 2: Convert UNIX Timestamp to Human Date

Figure 3: Adding a Event Class Feature based on *event*

Dataset Description

Feature Name	Data Type	Description
comments	int64	The number of first level comments made on the talk.
description	object	A blurb of what the talk is about.
duration	int64	The duration of the talk in seconds.
event	object	The TED/TEDx event where the talk took place.
film_date	int64	The Unix timestamp of the filming.
languages	int64	The number of languages in which the talk is available.
main_speaker	object	The first named speaker of the talk.
name	object	The official name of the TED Talk. Includes the title and the speaker.
num_speaker	int64	The number of speakers in the talk.
published_date	int64	The Unix timestamp for the publication of the talk on TED.com.
ratings	object	A stringified dictionary of the various ratings given to the talk.
related_talks	object	A list of dictionaries of recommended talks to watch next.
speaker_occupation	object	The occupation of the main speaker.
tags	object	The themes associated with the talk.
title	object	The title of the talk.
views	int64	The number of views on the talk.
url	object	The URL of the talk.
transcript	object	The official English transcript of the talk.

Implementation

Data Collection
The Datasets being used for this project is being downloaded from Kaggle.

Data Understanding
Using Exploratory data analysis(EDA) Technique to get a better understanding on the characteristics for existing attributes.

Data Pre-Processing
Handle missing values if there is any, normalize attributes upon needed, and add new attributes upon necessary.

Model Building
Build various Classification Models using different Algorithms and train the model with training dataset.

Model Selection
Compare the performance and accuracy of different models and select the best Classification Model.

Exploratory Data Analysis

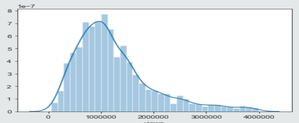

Figure 4: Distribution of number of views for each talk

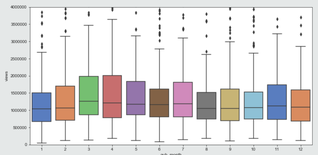

Figure 5: Boxplot for Publish Month and Views

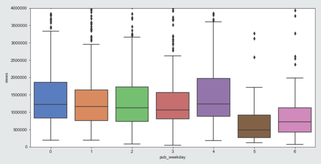

Figure 6: Boxplot for Publish Weekday and Views

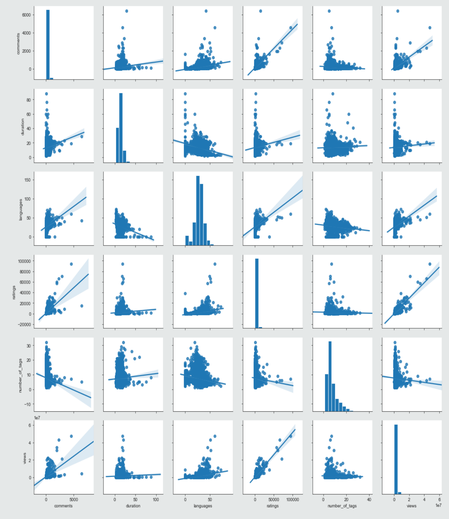

Figure 4: Regression Visualization

Result

- **What makes a popular TED Talk?**

Important features:

- *Ratings*
- *Comments*
- *Languages*
- *Duration*
- *Number of tags*
- *Published weekday*

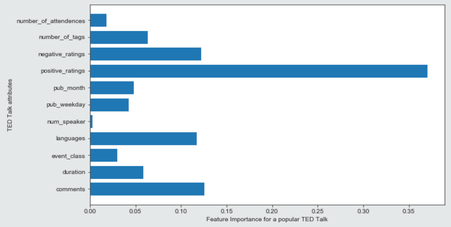

Figure 7: Feature Importance of Random Forest Classifier

- **What are the most popular topics in TED Talk?**

Figure 8: Most Popular Topics via Word Cloud

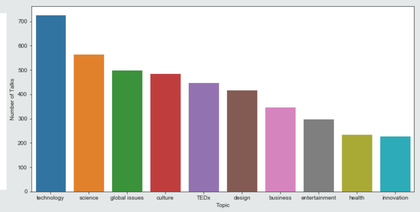

Figure 9: TOP 10 popular topics

- **Predict the popularity of a TED Talk based on its transcript ?**
- **Answer:** Prediction Model based on Multinomial Naïve Bayes [2] algorithm is selected as a better model that has Prediction Accuracy of **68%**.

Future Work

- Although this project has met its original objectives that it could identify those key features that contribute to the popularity of a TED talk and it could also predict the popularity based on transcript, but the features being used to train the classification model is limited and the prediction accuracy is lower than expected.
- Future work will be carried out to understand why the prediction accuracy is lower and do further pre-tuning and post-tuning to enhance the classification model. Further, understanding the reality of all features and take full advantage of all available features while training the classification model, such as, title, description, and tags are not being used for this project, but those features are actually important contributors for TED Talk popularity. Consequently, future work would also be carried out to apply Topic Modeling algorithms to classify the topic class to a categorical value, the proposed algorithm is *Latent Dirichlet allocation* (LDA). [3]

Acknowledgement & Reference

- This project is a course project for Data Mining at Hood College of Frederick, Maryland, during the Fall Semester of 2018. Firstly, I would like to thank Dr. Liu, Xianlian for giving us the fantastic opportunity that we can select our project topic by ourselves and the great guidance given through the entire semester. In addition, I would like to give thanks to my classmates and friends for providing suggestions on this project. Finally, I would like to express my deepest thanks and sincere appreciation to my family, girlfriend, and colleagues for their love, understanding, and support.

[1] Wikipedia. 2018. Wikipedia: TED (conference). Retrieved from https://www.wikipedia.org.

[2] L. Jiang, Z. Cai & D. Wang. 2010. Improving Naive Bayes for Classification, International Journal of Computers and Applications, 32:3, 328-332, DOI: 10.2316/Journal.202.2010.3.202-2747

[3] Machine Learning Plus. 2018. Topic Modeling with Gensim (Python). Retrieved December 16, 2018 from https://www.machinelearningplus.com/nlp/topic-modeling-gensim-python/