Springboard Capstone 1 : TED Talk Text Analysis

Pavan Poosarla; pavanpoosarla01@gmail.com ; Sep, 2019-Feb, 2020

1 INTRODUCTION

The goal of this capstone project is to automatically assign tags to a TED talk based on its transcript. To do this, we train our model on the existing TED talk dataset. TED talk dataset is taken from the Kaggle repositories.

Kaggle TED talk transcripts : https://www.kaggle.com/rounakbanik/ted-talks/version/3

This data has been scraped from the TED talk official website and is available now by a creative commons license. The database consists of talks scraped till Sep 21st, 2017. It has two files, containing information from about 2550 talks from >350 individual TED and TEDx events across the world. This gives a excellent medium to develop text analysis and NLP techniques to automatically predict sentiments from verbal speech

ted_main.csv : This file contains the metadata associated with each talk. The columns included in this file are as follows

Column	Description
comments	Number of first-level comments on the talk
description	A short note about the talk, usually 1-2 sentences
duration	In seconds
event	Event the talk was presented in
film_date	Unix timestamp of the filming
languages	Number of languages talk is available in
main_speaker	Main Speaker
name	Official name of TED talk, including speaker and title
num_speaker	Number of speakers
published	Unix timestamp of when talk appeared on ted.com
ratings	A stringified dictionary of the various ratings given to the talk (inspiring, fascinating, jaw dropping, etc.)
related_tags	A list of dictionaries of recommended talks to watch next
speaker_occupation	The occupation of the main speaker
tags	The themes associated with the talk
title	The title of the talk
url	The URL of the talk. This is common column with transcripts file, used for merging the two files on
views	The number of views on the talk

transcripts.csv : It consists of a file consisting of the text of the transcript along with the corresponding url of the talk. The columns on this file are

Column	Description
transcript	This is the transcript of the talk. Text is not divided into paragraphs. It includes comments and audience reactions in parenthesis. For example *(Applause), (Laughter), etc.*
url	URL link of the talk. This is the common column with the *ted_main.csv* used to merge both files on

2 OBJECTIVE

The overall objective of the project, apart from that of applying machine learning methods to real world dataset and validating the predictions, is to test the efficacy of NLP and text analysis methods for spoken language.

This dataset allows for testing features specific to the spoken word, like words/ minute, audience interactions, etc. In addition, we can also derive metrics to evaluate quality of speech by measuring audience engagement

3 DATA WRANGLING

The first step in creating a usable dataset to extract statistics from and test machine learning methods is to do data wrangling and output a cleaned dataset. We develop a cleaned dataset, 'df_clean'. To do this, we do the following steps

1. **Join** : The two individual files, 'ted_main.csv' and 'transcripts.csv' are merged on the column url with a inner join. This outputs a dataframe containing information on 2467 talks, each having 18 data columns.
2. **Missing Data** : Looking for missing values in all the columns, we find the dataset to be relatively clean. Only column having missing values is the 'speaker_occupation' column with 6 values missing. We replace missing data with the string 'Unknown'
3. **Remove Outliers :** We also drop the talks with more than 1 speaker. As we have a very few talks with more than 1 speaker, we do not want to contaminate the training data with these talks.
4. **Drop Extra Columns :** Next, we drop the columns which are either redundant or unnecessary for out analysis. The columns dropped and the justification is shown in the table on the next page. Note that some of these columns will provide insights into the data storytelling , but may not be relevant to the machine learning aspects (Categorical with too many categories, etc.)

Format Changes : Another change we do is to convert the date/time fields , i.e, 'film_datestamp' and 'pub_datestamp' into python datetime format. This makes it easier to perform datetime manipulations of these datefields in the future. It also makes these column data human readable.

Dropped Column	**Reasoning/ Justification**
event	Not relevant for learning. Perhaps relevant to data storytelling
languages	Not relevant as we only consider english transcripts. Can be relevant to data storytelling, where it is included.
name	Redundant data. Has title and speaker name
num_speaker	All are 1. Redundant data
related_talks	Dropped. Not relevant
speaker_occupation	Not relevant to machine learning. Relevant to data storytelling
url	Dropped. Not relevant
views	Dropped. Ma be too generic a metric to predict or use for training.

Once the above operations are completed, we also extract extra features based on analysis of the transcript text. The features extracted from the text are

1. *sentence_count* : Counts the number of sentences in the transcript.
2. *word_count* : Counts the number of words in the transcript