INTRODUCTION TO DATA SCIENCE PROJECT

DATA ANALYSIS OF TED-TALKS

A data analysis of a TedTalks dataset containing information about all audio - video recordings of TedTalks uploaded to the official TED.com

by
Pranavi : PES1201800368
Poojasree D : PES1201802032
Swanuja : PES1201800369

IDS Report
Data Analysis of TedTalks

CONTENT

Our TedTalks dataset contains information about all audio-video recordings of TED Talks uploaded to the official TED.com website until September 21st, 2017. The numerical columns in our dataset are id, the number of comments and the number of views. Our categorical columns are description, film date, published date, related tags, related themes, related videos, speaker, Ted event, title, URL and transcripts.

PROCESS

The data analysis of TedTalks is done by undertaking four main processes :

1. Data cleaning
2. Normalization
3. Graph visualization
4. Hypothesis testing
5. Correlation coefficient

DATA CLEANING

Our uncleaned dataset contained empty cells and there were three methods undertaken to clean the dataset-

1. Hotdeck
2. Filling missing values with the previous value or with the mean/median/mode
3. Taking related data from other related rows of the same column

Missing 'views' and 'comments' values were filled with the mean value. Missing 'dates' and 'ted_event' values were filled with ffill (previous value). Missing 'id' value were filled with the incremented previous value. Missing 'tags', 'themes', 'titles', 'speaker' and other related categorical data was filled with values taken from other columns. For example - if the speaker was missing, it was filled with the first two words of the title as the speaker's name always appears in the title first, and so on. Similar methods were coded to fill every missing value to its most approximate yet best value.

NORMALIZATION

The goal of normalization is to change the values of numeric columns in trh dataset to a common scale, without distorting differences in the range of values. It is a good technique to use when you do not know the distribution of your data or when you know the distribution is not Gaussian (a bell curve). Normalization is useful when your data has varying scales and the algorithm you are using does not make assumptions about the distribution of your data, such as k-nearest neighbors and artificial neural networks.
In our dataset, we have normalized the number of views and the number of comments such that all the values lie between 0 and 1. The formula used for this was -
df["views"]=((df["views"]-df["views"].min())/(df["views"].max()-df["views"].min()))*1
The normal distribution was plotted importing seaborn and using the function ‘distplot’.

GRAPH VISUALIZATION

Graph visualization is an important part of data analysis since it conveys tons of information pictorially, through a single graph. Our graph visualization consists of distribution plots, bar graphs, scatter plots, frequency distributions and wordclouds.

There are two graphs for distribution plots - distribution of number of views and number of comments.
The average number of views on TED Talks is 5.9L and the median number of views is 4.05L. This suggests a very high average level of popularity of TED Talks. We also notice that the majority of talks have views less than 12L.

On average, there are 97.49 comments on every TED Talk. Assuming the comments are constructive criticism, we can conclude that the TED Online Community is highly involved in discussions revolving TED Talks. There is a huge standard deviation associated with the comments. In fact, it is even larger than the mean suggesting that the measures may be sensitive to outliers. We shall plot this to check the nature of the distribution. The minimum number of comments on a talk is 7 and the maximum is 2284. The range is 2187. The minimum number, though, may be as a result of the talk being posted extremely recently.

There are two bar graphs we show the top 10 most viewed videos and the top 10 commented videos.

Our scatter plot shows the number of views vs number of comments and we can see a very distinct linear correlation between the two.

We have a frequency distribution of the most used words in a talk.

We have two wordcluds which shows the most used tags and the most used words in a talk.

HYPOTHESIS TESTING

H0 : Sum of videos posted in the first quarter of 2009 = Sum of videos posted in the first quarter of 2011
H0 : Sum of videos posted in the second quarter of 2009 = Sum of videos posted in the first quarter of 2011
H0 : Sum of videos posted in the third quarter of 2009 = Sum of videos posted in the first quarter of 2011
H0 : Sum of videos posted in the fourth quarter of 2009 = Sum of videos posted in the first quarter of 2011

H1 : At least one of the H0 is false

Degrees of freedom = 3

Chi2 = 7.5177759

P = 0.0571

Taking a confidence level of 95%, the critical value is found to be 9.348.

Since the critical value is greater than P, we reject the null hypothesis.

CORRELATION COEFFICIENT

We have tried to find the correlation coefficient between the number of views vs the length of the transcript. We have plotted graphs for two parts - one with the outliers and one without the outliers.