TED TALKS

Background,Problem and context:

TED, which operates under the slogan 'Ideas worth spreading' is a nonprofit devoted to spreading ideas, usually in the form of short, powerful talks (18 minutes or less). TED began in 1984 as a conference where Technology, Entertainment and Design converged, and today covers almost all topics from science to business to global issues in more than 100 languages. The first TED, concieved by Richard Saul Wurman,who co-founded with Harry Marks, included a demo of the compact disc, the e-book and cutting-edge 3D graphics from Lucasfilm, while mathematician Benoit Mandelbrot demonstrated how to map coastlines using his developing theory of fractal geometry.But despite a stellar lineup, the event lost money, and it was six years before Wurman and Marks tried again.

The TED Conference became an annual event in Monterey, California, attracting a growing and influential audience from many different disciplines united by their curiosity and open-mindedness and also by their shared discovery of an exciting secret. It has been held since 1990.Back then, TED was an invitation only event. It is now you are welcome and encouraged to apply to attend.TED Talks are giving so much information and knowledge about fields that are alien to the audience and amaze them in the form of outstanding short stories,breathtaking visuals and subtle humor. We will explore datasets containing transcripts available in www.kaggle.com of all audio-video recordings of TED Talks uploaded to the official TED.com website until September 21st, 2017.The problem may be articulated as

1. Similarity between most viewed and most favorited Talks of all time?
2. Summarization of most viewed and most favorited Talks?
3. Sentiment Analysis on comments of TED Talks?(What type of rating the Talks got as positive,negative and neutral?)
4. Based on the similarity,building a recommendation system to decide which talk they might watch?

Scope:

For this project my focus will be on TED Talks to find insights about the world of TED and will explore the following analysis to get meaningful derivations and inferences.

1. Which themes are most popular among TEDsters?
2. Which are the most viewed and most favorited Talks of all time?
3. What kind of topics attract the maximum discussion and debate (in the form of comments)?
4. The area which has more TED Talks being released?
5. Which months are most popular among TED and TEDx chapters?
6. TED Talk which has more influence on the crowd taking the number of views into consideration?

7. What is making the TED talks popular, considering the duration of the talk?
8. Number of languages TED talk is released in, number of views, number of responses received over them?
9. Has the Most Popular Ted Talks been released in more languages?
10. Occupation of the speaker and how it is related to TED talks.
11. Is the TED talk on global issues more popular and speakers from which fields are more leaned towards it?
12. Taking the gender of the speaker into account, to find out which gender has more popular and powerful speakers?

Target Clients

These meaningful derivation, inference and results can be used by several kinds of people with different aims in mind.

- A TED talk presenter can use it to know the common topic in which people are more interested to listen and thus he can plan the topic of his talk accordingly.
- Also, the patterns from rating data for a TED talk can help speakers to know whether their talk has been informative or confusing or ok and improve accordingly.
- Audience who wants to know the summary of the talk from a specific area without further going through the whole talk.
- Another audience of our result can be a person looking to gain information from any TED talks from a specific area. Our result would help such people to know which talk and speaker has the most popularity in that area as a suggestive guide. It is highly likely that a popular talk will have some informative message and learnings thus the listener can be benefited most.

Data

The source of the dataset contains information about all audio-video recordings of TED Talks uploaded to the official TED.com website until September 21st, 2017. The TED main dataset contains information about all talks including number of views, number of comments, descriptions, speakers and titles. The TED transcripts dataset contains the transcripts for all talks available on TED.com.

https://www.kaggle.com/rounakbanik/ted-talks

This dataset contains two csv files:

ted_main.csv - Contains data on actual TED Talk metadata and TED Talk speakers.
transcripts.csv - Contains transcript and URL information for TED Talks.

Approach:

To examine the specific problem, we will apply a full Data Science life cycle composed of the following steps:

1. Data Wrangling to audit the quality of the data and perform all the necessary actions to clean the dataset.
2. Data Exploration for understanding the variables and creating intuition on the data.
3. Feature Engineering to create additional variables from the existing.
4. Data Normalization and Data Transformation for preparing the dataset for the learning algorithms (if needed).
5. Training / Testing data creation to evaluate the performance of our models and fine-tune their hyperparameters.
6. Model selection and evaluation. This will be the final model for our goal.