BUDT 758X Project Proposal

See-through the talk

GROUP 28: Yao Xiao, Kavya Purushothaman, Pooja Datre

Introduction

We will be working with data regarding TED Talks, our dataset consists of information regarding all the audio-video recordings of TED until the 21st of September 2018. TED is a nonpartisan nonprofit devoted to spreading ideas, usually in the form of short, powerful talks. It 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 110 languages. Meanwhile, independently run TEDx events help share ideas in communities around the world. The dataset contains information about all talks, the various types of information it consists is comments, description of the talk, published dates, ratings, related talks, duration of the talk, speaker occupation, title, event year, the number of languages the talk has been published in, URL to the talk, the theme of the talk(For eg: funny, tech-savvy, business etc.).

The reason we chose this data was that TED talks are growing and turning out to be a point of discussion among the youth ball around the world, getting to know the insights about the speaker, the titles, themes, etc. would give us some insights on these talks and help make them even more successful than they're right now. TED Talks always concentrate on some very cool ideas and spreading them among numerous people around and create an impact in some or the other way. We can know what people focus on, what they like and what people think about through analyzing data about TED Talks. It will give us some fantastic details regarding this talk.

Questions of Interest

- How popular are the talks of one genre? (We can show it using graphs)
- What kind of TED Talk topics are the most populated or the most viewed ones?
- Does every talker focus on their own domain or some of them focus on other domains as well?
- What kind of topics attract the maximum discussion and debate (in the form of comments)?
- What is the relationship between the number of comments and the number of views on audio-video recordings?
- Are the topics which are mainly preferred by the speakers also preferred by the people?

- What topics or speakers have ratings which consist of some particular negative words such as unconvincing or confusing?
- What occupation do most of the speakers have? Does any particular occupation have the majority of the speakers from in there?
- Does the number of languages the audio-video recordings has been published in a particular year relate to the number of views?

Data Processing and Analysis

Dataset Description

There are 17 columns in this data. They are comments, description, duration, event, film_date, languages, main_speaker, name, num_speaker, published_date, ratings, related_talks, speaker_occupation, tags, title, URL, views.

Data Processing Tasks

- Download and scrape data from web pages
- Indexing, selection, and filtering
- Data cleaning(Managing missing data and making sure each column has values of the same data type)
- Data transformation
- Data Processing (Processing text data and numerical data)
- Data Visualization(Showing the insights through graphs)

Data Analysis

We will be making use of python to analyze the data. We can analyze various columns in this dataset in order to get information with regard to the questions we intend on answering. We can know what topics are the most preferred one among the viewers, are these topics also the ones which are more frequently spoken about by the speakers?. We will find out what occupation do most of the speakers hold, are speakers evenly from various different occupations or are the majority of the speakers having one occupation?. We will also create charts(visualize the data) to observe the relation between the number of comments and the number of views and know about the most viewed talk. We will find out how the number of languages that the recording is published in affects the number of views that particular recording has. We also will try to figure out how many ratings have words such as unconvincing or negative words to know what kind of recordings receive negative views. This will give us some great insights and hence help us make this better in the future.

Expected Findings

The findings we intend to have are as follows:

- The most preferred topic among the viewers
- The occupation most of the speakers belong to
- The number of languages an audio-video recording should be published in
- Topics creating a huge amount of discussion among viewers
- Are the number of views and number of comments interconnected
- What topics have negative ratings

Project Timeline

Task	**Task Lead**	**Due Date**
Data set collection (Scrape data from web pages)	Yao Xiao	October 15
Data cleansing	Kavya Purushothaman	October 23
Analyze the data and visualize the result of the analysis	Pooja Datre	November 11
Project completing	Yao Xiao, Kavya Purushothaman and Pooja Datre	November 30
Project showcase	Yao Xiao, Kavya Purushothaman and Pooja Datre	December 5