CIS 600 – SOCIAL MEDIA AND DATA MINING

SPRING 2020 PROJECT

TED TALK ANALYSER

INDEX

Table of Contents

GROUP MEMBERS

- Bharat Matai
- Debopriyo Battacharya
- Parth Shah
- Samarth Shah

BACKGROUND AND MOTIVATION

In the present universe of expanding use of online networking and systems, an assessment, truth or a straightforward story has an intensified impact on the population. Web based or social media life has set become a reason for manipulating an individual's convictions and conclusions. A captivating report by New York Times Consumer Insight Group uncovered the inspirations that members referred to for sharing data via web-based networking media. These incorporate a longing to uncover significant and engaging substance to other people; to characterize themselves; to develop and sustain connections and to get the word out about brands and causes they like or backing. These variables have made interpersonal organizations advance from being a helpful methods for staying in contact with loved ones to being utilized in manners that really affect society. Web based life is being utilized fit as a fiddle governmental issues, business, world culture, instruction, professions, advancement, and the sky is the limit from there

WHY PEOPLE SHARE

The Psychology of Social Sharing

49% **ENTERTAINMENT**
49% say sharing allows them to inform others of products they care about and potentially change opinions or encourage action.

68% **DEFINE OURSELVES**
68% share to give people a better sense of who they are and what they care about.

78% **RELATIONSHIPS**
78% share information online because it lets them stay connected to people they may not otherwise stay in touch with.

69% **SELF-FULFILLMENT**
69% share information because it allows them to feel more involved in the world.

84% **SUPPORT A CAUSE**
84% share because it is a way to support causes or issues they care about.

What People Like To Share On Social Networks

Pictures	43%
Opinions	26%
Status update of what and how they are doing	26%
Links to articles	26%
Personal recommendations of things they like	25%
News items	22%
Links to other websites	21%
Links to other people's post	21%
Status update of what they are feeling	19%
Video clips	17%
Plans for future activities, trips and plans	9%

4

TED talks which began in 1990 with a motto of "Ideas worth spreading" held as a conference each year was one such stage where individuals at their apex of accomplishment affected numerous individuals through their accounts and encounters. From that point forward, TED talks have become overall turning into a voice for any individual who has a compelling contribution. What began as just talks in the field of software engineering has since widened its point of view to remember talks for some logical, social, political, and scholarly subjects.

TED Talks are turning out to be increasingly more famous with expanding number of themes and recordings. Individuals are discovering TED Talks increasingly intriguing and an incredible wellspring of data for a wide range of data. TED Talks have end up being an incredible stage for different organizations, demonstrating your plan to the world and to address worldwide issues. Late achievement and prevalence of the TED Talks can't be restricted to the assorted variety of points or the nature of the speakers and their substance. There are different elements that should be represented before nailing down the achievement of TED Talks to a solitary factor.

MOTIVATION AND ABSTRACT

We accept that these discussions have a great deal of impact over structure an individual's sentiment about a theme, changing the individual's feeling about the subject, engaging an individual in a related field or in any event, identifying with individuals with comparable foundations. There has consistently been a few points in the past which have been exceptionally persuasive and some of the time even the discussion of the century. We will probably distinguish such subjects and make a pattern of what has been the individuals' most loved and what may hold for what's to come. Because of the worldwide idea of these discussions, there are a high number of talks that will happen later on. There have been around 2600 TED talks in the year 2018 alone with 5 to 7 TED talks for every week. We likewise distinguish the speakers and attempt to think of an approach to assist them with understanding their crowd better by giving a reproduced consequence of the discussion dependent on specific parameters of the discourse being introduced. Thought behind structure TED Talk Analyzer is to experience these different factors and think of a conceivable decision about what makes any Talk a hit or a failure. TED Talk Analyzer examinations dataset of a wide range of TED Talks and finds what makes a subject or TED Talks more well known than the other. The application recognizes the most widely recognized characteristics in the current famous TED Talks and make a rundown of parameters which acts like necessities that must be mulled over before conveying a discussion. A higher parameter match would speak to a higher likelihood of being remembered for the rundown of well known TED Talks. In light of the suggestion, the speaker can design his substance as indicated by have better reaction from the crowd. Application recommended which subject were well known as of late and what number of Talks were conveyed on them. The application additionally suggests a rundown of comparative TED Talks dependent on the KNN calculation. Proposals are likewise sifted with the necessities enter by the client so it is as exact and exact as could reasonably be expected. To give better understanding, charts have likewise been utilized in the application.

Our group intends to broaden the venture in future by joining the Sentimental Analysis on the suggested TED Talks and apply different AI methods to give every one of these suggested Talk a SCORE dependent on the client inputs like span, number of dialects, number of remarks and perspectives.

DATA SOURCES

- https://www.kaggle.com: The dataset utilized for examination intention was accumulated from Kaggle. The dataset comprises of different components like the name of the TED Talk, its depiction and URL. The information with respect to the dialects wherein TED Talk is interpreted and number of labels to the TED Talk is additionally put away. The quantity of remarks will likewise be mulled over as all the above parameters are truly acceptable marker of the prominence of the TED Talk.

- https://twitter.com: The Twitter APIs are used to display the current trending tweets mentioning TED Talks.

APPROACH

- Gathering, preprocessing and stacking TED Talks dataset.
- Examining dataset for speaker data of a Talk.
- Digging information for getting most famous labels as of late.
- Sifting information to show it in graphical structure dependent on factors like distributed year, evaluations and labels.
- Making K-Nearest Neighbor Algorithm dependent on the dataset to discover is the TED Talk will give positive input or negative criticism.
- Utilizing Twitter APIs to bring drifting tweets identified with TED Talks.
- Suggesting the TED Talks dependent on KNN calculation made and matching it with client inputs.

LIBRARIES

- Pandas (Data frame analysis and manipulation)
- Numpy (Preprocessing)
- Matplotlib (Graphs Visualization)
- tkinter (GUI)
- Seaborn (Statistical data visualization)
- Sklearn (Machine learning)
- Twitter (Twitter Data)

COLLECTION AND PREPROCESSING OF DATA

The dataset in regards to the TED Talk was brought from CSV file to information outlines with assistance of Pandas library. The dataset contained Published Date and Filmed Data of TED Talks on exceptional arrangement. Accessible sections from the information were remarks: The quantity of first level remarks made on the discussion portrayal: An ad spot of what the discussion is about length: The term of the discussion in seconds occasion: The TED/TEDx occasion where the discussion occurred film date: The Unix timestamp of the shooting dialects: The quantity of dialects wherein the discussion is accessible main speaker: The first named speaker of the discussion name: The official name of the TED Talk. Incorporates the title and the speaker. num_speaker: The quantity of speakers in the discussion published date: The Unix timestamp for the production of the discussion on TED.com appraisals: A stringified word reference of the different evaluations given to the discussion (motivating, captivating, stunning, and so on.) related talks : A rundown of word references of prescribed converses with watch next speaker_occupation: The control of the fundamental speaker labels : The subjects related with the discussion title : The title of the discussion url: The URL of the discussion sees : The quantity of perspectives on the discussionTo make the mining and analysis of data, the dataset was preprocessed, and it was converted to DD-MM-YY format. For any video to be popular, one of the basic indicators is the number of views it has achieved. We sorted the dataset in decreasing order of the views of TED Talk. Once this was done, we went through dataset about selecting all the parameters of interest that could potentially be important for deriving crucial information from the dataset.
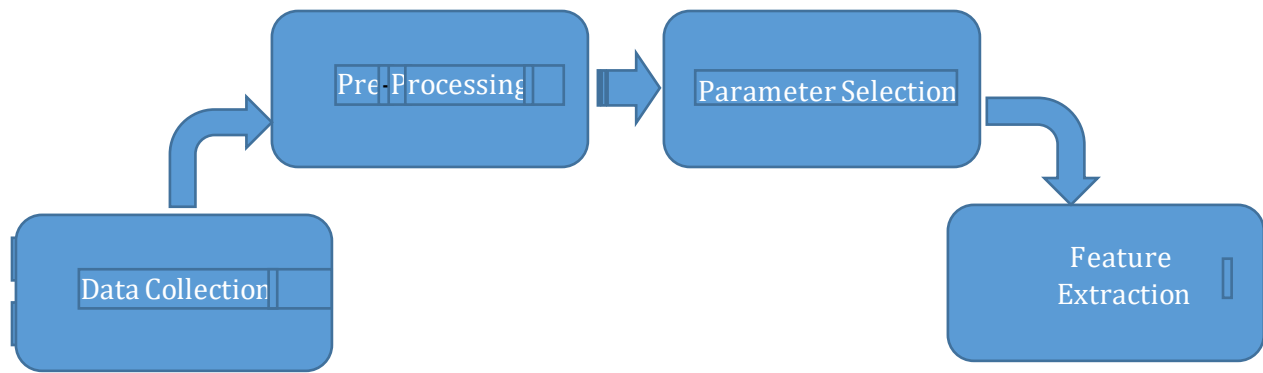
DATASET ANALYSIS

In order to make data meaningful, following steps were taken:

- Stacking the dataset and preparing it.
- Bringing in and stacking bundles important for AI and perception.
- Making capacities to change over the dataset into the necessary structure.
- Changing over the segments with the assistance of capacities made before.

PARAMETER IDENTIFICATION AND FEATURE EXTRACTION

▢ PARAMETER IDENTIFICATION

When the information was changed over into the ideal organization, it was important to distinguish the parameters that will be helpful to investigate what makes a TED Talk all the more engaging and popular contrasted with others. With all extraordinary sort of parameters related with the TED Talk, the one parameter that stands apart is the speaker of the Talk. Any discourse or TED Talk conveyed will be all the more persuading and significant if the speaker conveying it has abundant measure of information on the given field.

Another significant parameter that will assist a client with identifying on the off chance that he/she is keen on a TED talk is the labels. Every video is related with different labels. A client can simply look through the recordings by the intrigued labels and channel out the undesirable TED Talks.

▢ FEATURE EXTRACTION

Utilizing these two parameters present in the dataset, we broke down the dataset and indicated how the patterns in TED Talk are molding out contingent upon the speaker occupation and the different labels identified with that TED Talk.

SPEAKER'S OCCUPATION

We utilized the 'speaker_occupation' segment present in the dataset. Utilizing this, main 10 moderator's occupation was recognized.

```
Writer          45
Designer        34
Artist          34
Journalist      33
Entrepreneur    31
Architect       30
Inventor        27
Psychologist    26
Photographer    25
Filmmaker       21
Name: speaker_occupation, dtype: int64
```

TOP 10 Presenter's Occupation Count

When the control of most well known speaker was recognized, we utilized this information to match it with the quantity of perspectives a TED Talk had gotten. This data was helpful in demonstrating how often a TED Talk introduced by an 'Author' or 'Craftsman' was seen and how well it was acknowledged by the clients.

```
Occupation:  Views
Writer:  2967762
Designer:  1273358
Artist:  1036726
Jornalist:  1450326
Entrepreneur 1992387
```

TOP 5 SPEAKER'S OCCUPATION: NUMBER OF VIEWS

TAGS

Another significant section in the dataset is 'labels'. Labels are another method for discovering which sort of TED Talk is slanting and is mainstream lately. We utilized labels related to every TED Talk to distinguish the most well-known labels in past three years: 2017,2016 and 2015.

```
Most popular Tags for year 2015
==============================
 technology        39
 science           34
 health            25
 global issues     24
 design            23
TEDx               22
 women             21
TED Fellows        20
 social change     19
 innovation        19
Name: tags, dtype: int64
```

Popular Tags in 2015

```
Most popular Tags for year 2016
==============================
 society           83
 innovation        76
 technology        74
 future            68
 science           60
 communication     59
 humanity          58
 social change     55
 identity          55
 community         52
Name: tags, dtype: int64
```

Popular Tags in 2016

```
Most popular Tags for year 2017
==============================
 society           86
 humanity          70
 social change     60
 communication     54
 technology        49
 science           43
 innovation        43
 community         39
 future            36
 personal growth   36
Name: tags, dtype: int64
```

Popular Tags in 2017

DATA ANALYSIS: STATISTICS AND GRAPHS

Graphs are essential in indicating statistics and making exhausting information all the more engaging and straightforward. How the TED Talks have developed alongside the time is appeared with the assistance of the segment 'published_date'. This segment was adjusted while pre-processing and the year wherein a specific Talk was conveyed is extricated from this segment. The quantity of Talks conveyed in a given year is determined utilizing inbuilt python work. Utilizing the include of TED Talks conveyed in every year is shown in visual chart group utilizing Matplotlib. The chart shows the numbner of talks distributed each year from 2006 to 2017. We can see that number of talks expanded steadily till the year 2012 and afterward it began to diminish marginally towards 2017. The explanation could be on the grounds that the

quantity of watchers in the year 2012-2013 were more and in this manner there were more talks distributed in 2012 because of the interest.

TALKS VS PUBLISHED_YEAR
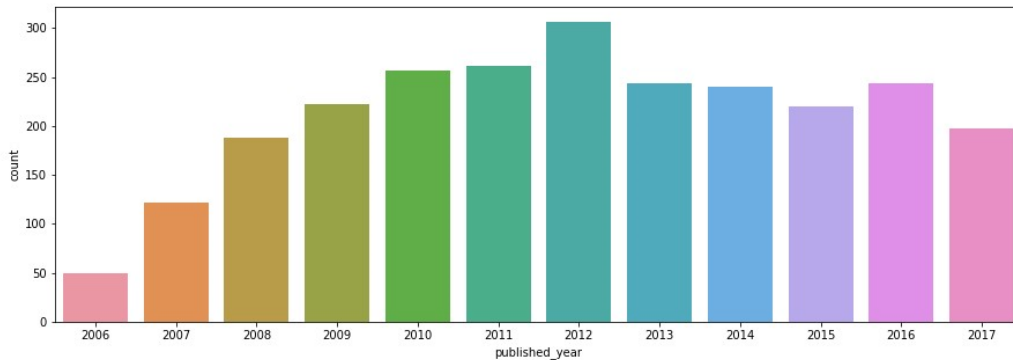
Also, every TED Talk is given evaluations. The section 'evaluations' contains different watchwords which can be utilized to rate how the crowd has reacted to the discourse. A portion of the catchphrases incorporate 'Enlightening', 'Confounding', 'Wonderful', 'Interesting, etc. These are really acceptable markers about how well the TED Talk was gotten by people in general. The quantity of TED Talks containing every catchphrase was determined and insights were shown in the diagram position utilizing Matplotlib. We at that point separated each evaluating type into positive and negative. The diagram shows the absolute tallies of each appraising. The ones in red are negative appraisals though the staying ones are The blue ones are certain ones. We can see that the individuals for the most part give lesser negative rating than positive. Consequently, when we do examination on evaluations for suggestions we have to have higher proportion set for negative appraisals.

TALKS VS RATING
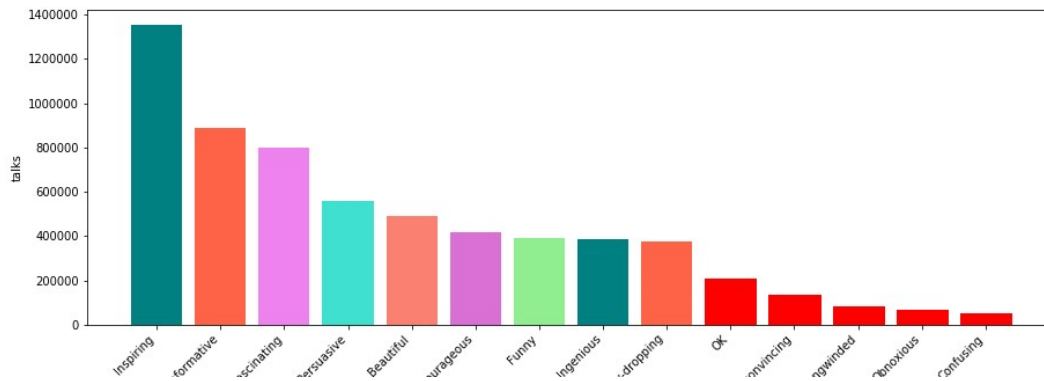
We had discussed earlier the importance of tags and how it can be used to filter the required TED Talks and narrow down the data as per our needs. Hence, we displayed the TED Talks against the most popular tags in the dataset.

TALKS VS TAGS
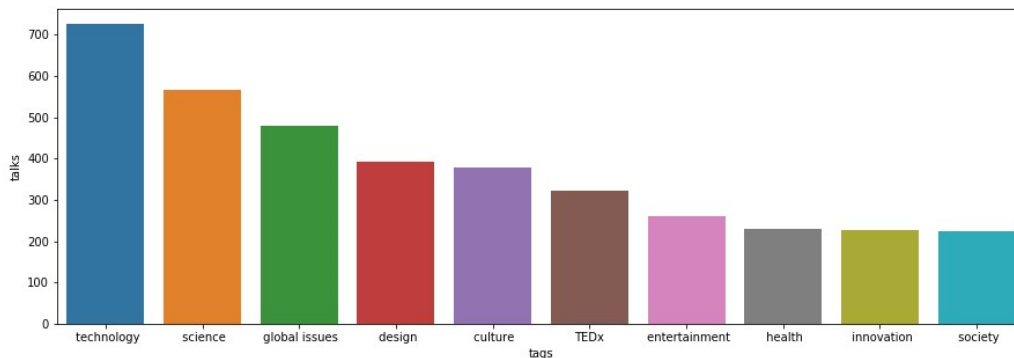
These graphs help us analyze the trending topics and how it can change in time. We can see that technology has been the most talked about topic and we believe it will continue to do so. Soon after, it is followed by science making technology and science the most talked about topic since the inception of TED talks.

Despite this, there have been certain years where this trend has changed to other topics like society, humanity, communication among others. This has been observed recently from 2017 with the previous years being technology at the first position

```python
# -*- coding: utf-8 -*-
"""
Created on Fri Apr 17 22:32:28 2020

@author: Parth Shah, Bharat Matai
"""

#Function which creates a graph for Number of Talks Vs Ratings
import numpy as np
import matplotlib.pyplot as plt
def ratingsPlot(dataset,counter,neg_descriptors):
    dataset['ratings'] = dataset['ratings'].apply(lambda x: eval(str(x)))
    for i in range(len(dataset['ratings'])):
        x,y=0,0
        for j in range(len(dataset['ratings'][i])):
            counter[dataset['ratings'][i][j]['name']] += dataset['ratings'][i][j]['count']
            if dataset['ratings'][i][j]['name'] in neg_descriptors:
                y+=dataset['ratings'][i][j]['count']
            else:
                x+=dataset['ratings'][i][j]['count']

    frequencies = list(counter.values())

    descriptors = [x for _,x in sorted(zip(frequencies,counter.keys()), reverse=True)]
    neg_indices  = [x for x in range (len(descriptors)) if descriptors[x] in neg_descriptors]
    frequencies.sort(reverse=True)

    indices = np.arange(len(descriptors))
    bar = plt.bar(indices, frequencies, 0.8,color=['teal', 'tomato', 'violet', 'turquoise','salmon' ,'orchid','lightgreen'])
    [bar[i].set_color('r') for i in neg_indices]
    plt.xticks(indices, descriptors, rotation=45, ha="right")
    plt.savefig("rating.png")


# -*- coding: utf-8 -*-
"""
Created on Fri Apr 17 21:08:03 2020

@author: Debopriya Bhattacharya, Bharat Matai
"""
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from tkinter import messagebox


#Function - Creates graph for Number of Talks Vs Tags

def tagsTalksPlot(dataset):
    tags = []
    for i in range(len(dataset.loc[:,'tags'])):
        ls = list(dataset.loc[:,'tags'])[i][2:-2].split(',')
        for c in range(len(ls)):
            value= list(dataset.loc[:,'tags'])[i][2:-2].split(',')[c]
            tags.append(value.replace("'",""))
    tags = pd.DataFrame(tags,columns=["tags"])
    tags.iloc[:,0].value_counts().head(10)
    tags = pd.DataFrame(tags.iloc[:,0].value_counts()).reset_index()
    plt.figure(figsize=(15,5))
    sns.barplot(x=tags["index"].head(10),y=tags["tags"].head(10))
    plt.xlabel("tags")
    plt.ylabel("talks")
    plt.savefig("tagTalk.png")
```

```
#Function -  counts Number of Tags per year
def tagsCountYearly(dataset,year,listTags):
    tags = []
    dataset = dataset[dataset["published_year"]==year]
    for i in range(len(dataset.loc[:,'tags'])):
        ls = list(dataset.loc[:,'tags'])[i][2:-2].split(',')
        for c in range(len(ls)):
            value= list(dataset.loc[:,'tags'])[i][2:-2].split(',')[c]
            tags.append(value.replace("'",""))
    tags=pd.DataFrame(tags,columns=["tags"])

    print(tags["tags"].value_counts().head(10))
```

K-NEAREST NEIGHBOR ALGORITHM

One of the most simplest and learning algorithm is KNN algorithm. For every data point set, we look at the K nearest training data points. After which we take the class that is occurring most frequently and then assign that class to the test data. Therefore, it can be said that K represents the training data points that lie near the test data points which will be later used to find the class.

All cases are stored by KNN and classifies new cases as per the similarity measures. A new object i.e, talk in our case, is classified by a majority votes for its neighbor classes. The object is always assigned to the most common class amongst its k

Consider the following image to understand the KNN better: https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm

K-NN CLASSIFICATION
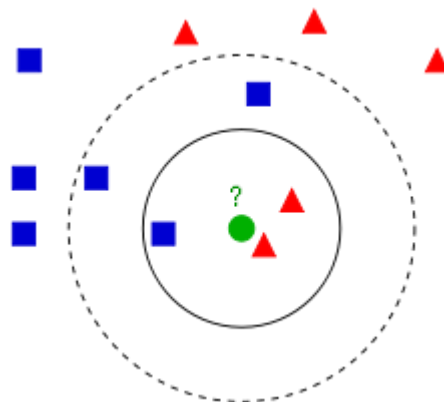
In the example above, the green circle represents the test data and that we need to identify which class i.e., blue square or red triangle should be assigned to it.

For example if k =3, then red triangle should be assigned as the number of red triangle is more compared to blue squares i.e., 2>1.

Now For example, if k=5, then the 'blue square' class should be assigned as the number of blue square is greater than the number of red triangles i.e., 3>2.

A new column 'r_feedback' is generated before implementing the algorithm to detect whether a talk gets a negative feedback or a positive feedback. That was implemented by setting a threshold value for 3 attributes i.e., number of comments, language, conversions and views.

However, in our dataset, we used columns, 'comments', views' and 'languages' to measure the feedback of any TED talk. It is said that with increase in the number of views, the popularity of the TED Talk is also said to be increased. Same is the case with comments. If the speech is popular than there are high chances that the comments is directly proportional. And also if the TED talk is translated in more number of different languages than there's higher chance of it being more popular and therefore, higher chance of success.

Considering our algorithm, there were 2 classes for classification of a TED Talk. They were represented as 0 and 1 where 'positive feedback' was indicated by 1 and 'negative feedback' by 0

To decide whether a TED talk would have a positive feedback or a negative feedback there were indicators that were taken into considerations:

Comments:

If the number of comments were greater than 20000, then the TED talk could have a positive feedback else it might have a negative feedback.

Views:

If the number of views were greater than 300000 then the TED talk could have a positive feedback else it might have a negative feedback.

Number of Translated Languages:

If the number of languages the TED Talk translated was more than 30, then the TED Talk could have a positive feedback else it may have a negative feedback.

If the TED talk was satisfied by all the 3 identifier than it was assigned as a 'positive' if either of the identifier was not satisfied, then the TED talk in data set is set to 'negative'.

K-neighbor classifier with k=10 is applied in the training set of the dataset. Once the classifier is set, we use the classifier to predict the feedback of the TED talk.

A future application of mining tweets relevant to a specific TED talk to analyze the sentiments of the users based on the talk was created. After the dataset containing the speech in words was present, we can use the language processing module to analyze the usage of words to understand the overall sentiment of the speech. That can later be compared against the response sentiment to provide a better prediction to the user for his/her next speech.

```python
# -*- coding: utf-8 -*-
"""
Created on Fri Apr 17 23:09:48 2020

@author: Debopriya Bhattacharya, Bharat matai, Samarth Shah
"""
from sklearn.model_selection import train_test_split
from sklearn.neighbors import KNeighborsClassifier
import xlsxwriter

def knn(dataset,listTags,duration,languageEntry):
    #splitting of dataset
    r_fb=[]
    dataset['ratings'] = dataset['ratings'].apply(lambda x: eval(str(x)))
    for i in range(len(dataset['ratings'])):
        if dataset['comments'][i]<10000 and dataset['views'][i]<500000 and dataset['languages'][i]<30:
            r_fb.append(0)
        else:
            r_fb.append(1)

    dataset['r_feedback']=r_fb
    X=dataset[['comments','views','languages']]
    Y=dataset['r_feedback']

    # Splitting the data into training sets and testing sets.
    X_train,X_test,y_train,y_test = train_test_split(X,Y,random_state=0,test_size=0.5)

    #With 5 neighbours, instatiating the model
    classifier=KNeighborsClassifier(algorithm='brute',n_neighbors=10)

    #Fit the model on the training data
    classifier.fit(X_train,y_train)

    #Checking if the model is performing properly on test data.
    print("\n\nAccuracy Score of the proposed Algorithm is:",classifier.score(X_test,y_test)*100)
```

KNN Code

A python library named sklearn was used for training and testing the data into the algorithm. The value of k was 10. And we set the test size to 0.5 which means we used half the dataset for training and remaining half for testing. The accuracy of the KNN algorithm for the dataset turned out to be 93.41%.

RELATED TED TALKS
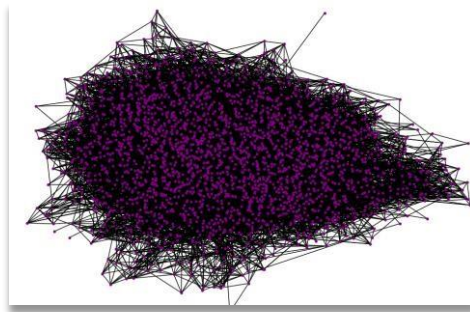
Every node in the above image is connected to a minimum of 3 other nodes. This means how each TED talk is relevant to each other even after the topics being so different.

This is the reason why because of this relation with other nodes that gives us recommendation of related TED talks after watching a particular TED talk..
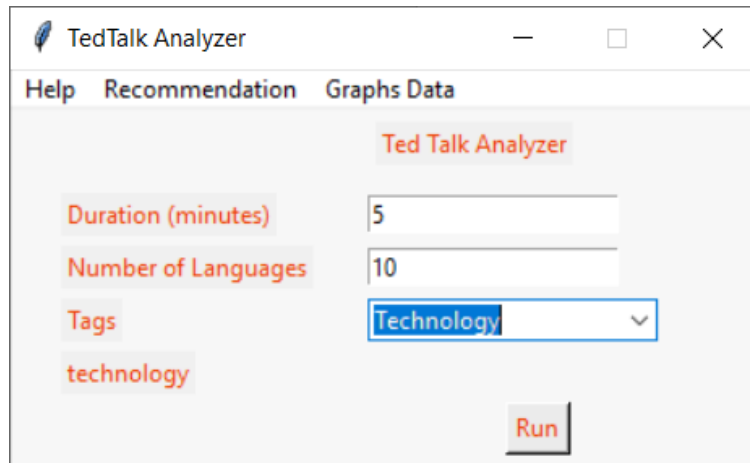
USER PARAMETER (INPUT)

A package named tkinter was used to create a Graphical User Interface to analyze the TED Talks. Tkinter fetches input from the users. 3 input fields were used to filter the TED talks depending on the user's liking. They are 'Duration', 'Languages', and 'Tags'.
It also allowed to see the following graphs that were generated and stored in the system:
i.   Number of talks vs Published year.
ii.   Number of talks vs Ratings.
iii.   Number of talks vs Tags.
On studying the graph there can be several analysis to obtain the user information like strength of relations, topic which is generally relevant and least generally relevant, etc. Video Streaming module can take advantage of this graph to provide recommendation to its users for related videos.
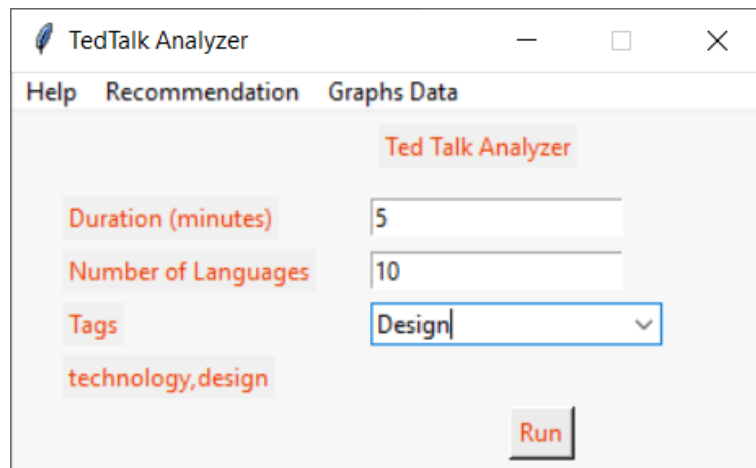
GRAPHICAL USER INTERFACE

TED TALK RECOMMENDATION

The KNN algorithm that is being used along with the user parameter to recommend TED talk depending on the user's liking that have higher possibility of getting positive feedback. The recommendation process can be split into following steps:

1. From the dataset the KNN algorithm is applied to a TED Talk. Depending on it, a class will be assigned by the classifier. If the class predicted is 0 then it will be discarded from the possibility of being a positive TED talk.
2. If the class assigned is 1 then it will be passed further to match other parameters.
3. After knowing that the prediction for the TED talk is positive feedback, we check whether it matches the user's requirement.
4. If the requirements fail we don't consider it for recommendation. Otherwise it will be added in the list of the recommended TED Talks and displayed to the user.

Consider, the following example:
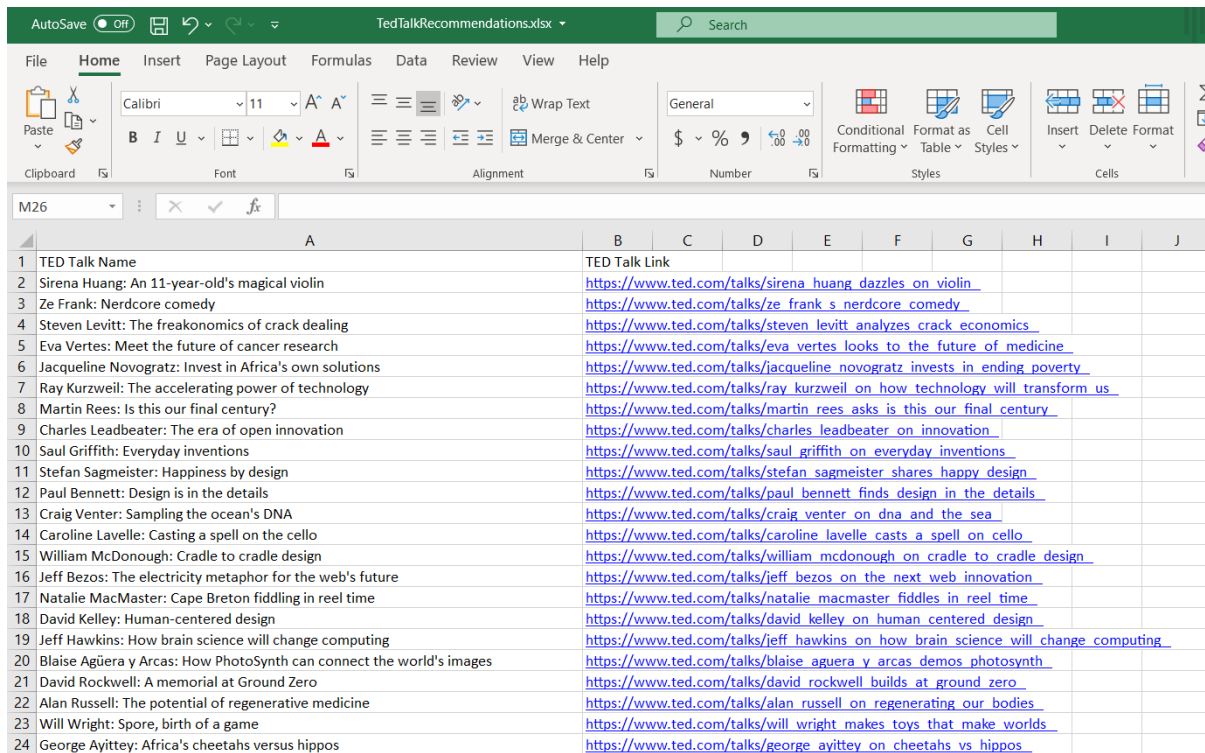
Considering this example, a user wants to know if the TED talks have higher possibility of getting a positive response from the audience if the TED talk is atleast 'x' mins long and has to be translated in minimum of 'y' different languages. And that each of those TED talks needs to be either relevant to 'Technology' or 'Culture'.

If the above 3 parameters does not match with the TED talks in the database, then it isn't recommended. And if they are matched then it is added to the recommendation list and returned to the user.

LIST OF RECOMMENDED TED TALKS

```
List of Recommended TED Talks:
===============================================

1 ) Sirena Huang: An 11-year-old's magical violin

2 ) Ze Frank: Nerdcore comedy

3 ) Steven Levitt: The freakonomics of crack dealing

4 ) Eva Vertes: Meet the future of cancer research

5 ) Jacqueline Novogratz: Invest in Africa's own solutions

6 ) Ray Kurzweil: The accelerating power of technology

7 ) Martin Rees: Is this our final century?

8 ) Charles Leadbeater: The era of open innovation

9 ) Saul Griffith: Everyday inventions

10 ) Stefan Sagmeister: Happiness by design

11 ) Paul Bennett: Design is in the details

12 ) Craig Venter: Sampling the ocean's DNA
```

TED talks recommendations are stored in excel files. And therefore we have used python and Xlsxwriter module. The file is stored in a current directory named 'TedTalkRecommendations.xlxs'. It contains the name of all the recommended TED talk videos along with the link to that video.

TED TALK RECOMMENDATIONS EXCEL FILE

SOCIAL MEDIA AND TED TALKS

TED Talk focuses on providing information to the user on social media. The social media taken into focus for our application purpose is Twitter. It is emerged to be the biggest platform to discuss regarding any trending topics and almost anything. We plan to use this to show users what people are discussing regarding TED talks. Twitter has proven to be the best place to get to know the behaviour of people and judge the success of any event. TED talk is no exception and more accurate judgement and precise suggestion is achieved. TED talk Analyzer uses Twitter APIs to fetch the data for a particular hashtag.

All the recent tweets containing '#TedTalk' are requested from the Twitter server. The application fetches the most recent tweets containing '#Tedtalk'.

```
=====================================
Tweet:  Thread by @RevealingDaniel: #Helena Thread #Blackout #TedTalk #QAnon So I just happened across this #TedTalk. Th
e t… https://t.co/OSwG059FT7
=====================================


=====================================
Tweet:  There's this lovely #TedTalk by @grahamshawdraw that you draw along with and I cant recommend it enough. I turne
d m… https://t.co/l8SsvNVaYg
=====================================


=====================================
Tweet:  RT @marshawright: Tweet #positive #quotes today

USE&gt;#ThinkBIGSundayWithMarsha&lt;

for a RT to 550K #Klout70 #TEDTalk #startup #marketing htt…
=====================================


=====================================
Tweet:  RT @marshawright: Tweet #positive #quotes today

USE&gt;#ThinkBIGSundayWithMarsha&lt;

for a RT to 550K #Klout70 #TEDTalk #startup #marketing htt…
=====================================
```

TWITTER SEARCH RESULT

CONCLUSION

Our analysis has given an insight on nature and popularity of TED talks that have been collected. The dataset from Kaggle was used in our project containing statistical data on the talks. We generated results on different parameters. At first we tried to understand the nature of our data and identified the key parameters to use as features for classification and prediction model. Using this we have analyzed how the topic popularity changed over time, the rate at which users have grown, etc. We took into consideration parameters, like viewer count, rating, speaker and his/her occupation etc. With the ability to combine all this we were able to build a classification model for sentiment analysis of the talk.

In future we would like to implement algorithms like k means clustering on our dataset and Naïve-Bayes on our dataset. With the availability of user comments, we will be able to implement an improved version of the sentiment analysis. Also with the availability of geo tagging and locations, we will be able to suggest possible locations for a TED talk.