Information Retrieval
CS F469
First Semester, 2019-2020

Assignment 1

Music Lyrics Search Engine

Documentation

Team #4

Ekanshi Agrawal	**Sandesh Thakar**	**Vrutik Halani**	**Vivek Soni**
2017A7PS0233H	2017A7PS0181H	2017A7PS1732H	2017A7PS0173H

Functions:

```
CosineSimilarity(processed_query,word_by_id,id_by_word,song_tf_idf_by_songid)
```

:The function is used to get the similarity score between the string array (processed_query) and the songs that exist in the dataset to retrieve the top fifteen songs relevant to the query. The function also takes into account the number of words that are common in the query and the song apart to retrieve more accurate results.This function returns the top 15 songs that match the query.

```
autocorrect(str1,str2,n,m)
```

:The function calculates and returns the edit distance between the given strings(str1 and str2)

```
dataset1(filename)
```

:This function reads the dataset of words and lyrics and indexes all the words by ids and collects songs from dataset and return them

```
tf_idf_calculation(songs)
```

:This function iterates through all the songs that have been retrieved from the dataset and calculates the tf-idf score of each word of the given songs

```
get_song_details(filename)
```

:This function reads the dataset and gathers song details and indexes the song details by song_id and returns it

```
process_query(query,id_by_word,words)
```

: This function pre-processes the given query by replacing short forms like “I’ve” to “I have”, checking for spelling corrections replacing words by their ID.

```
fetch_query()
```

: This function fetches the query from tkinter Text field and gets the processed query for searching in the dataset.

```
write_in_text(relevant_songs)
```

: This function writes the top 15 songs retrieved in the Text Area

Libraries used:

```
import Tkinter
```

: For designing the GUI of the search engine

```
import math
```

: To use the log function for calculating the tf-idf.

```
from stemming.porter2 import stem
```

: To stem the fetched query. We stemmed the query since the given dataset is also stemmed.

About :

The algorithm used in this search engine to fetch the nearest answer is Vector Space Model which calculates tf-idf score of each word in each song and also takes into account the stop words since they are an important part of the lyrics. The extra function we used in order to modify the Cosine Score to fetch appropriate results corresponding to the queries is to divide the cosine scores of every song by a term :

POWER(10,DIFF) :

DIFF = DIFFERENCE OF NUMBER OF WORDS IN QUERY
AND NUMBER OF QUERY WORDS IN SONG

This factor increases the accuracy of the search engine by a very greater factor. The main aim of using this factor was :
Query = “In the side streets something’s moving “
The song which has something around 20 times jums first in the list but actual song doesn’t come but after adding the factor we did find better results.