DOKUZ EYLUL UNIVERSITY
ENGINEERING FACULTY
DEPARTMENT OF COMPUTER ENGINEERING

CME 1203-INTRODUCTION TO COMPUTER ENGINEERING

Assignment

Film Script Analysis and Representation

by
Dırahşan Çağrı İrdemez

Lecturers
Prof.Dr. Yalçın Çebi
Res.Asst. İlker Kalaycı

IZMIR
21.10.2014

CHAPTER ONE

PROGRESS DESCRIPTION

This project is about making a program which can analyze movie scripts and find the repeating frequency of words. User has options of providing a script as a ‘txt’ file or as an url to (Imsdb) a website. After the script input is taken, the program can compare two scripts and sort the first twenty common words, or just find word frequencies based on one file.

The aim of the project is people who has a relation with theatre culture, including writers, critics, actors and actresses and hobbyst. The analization of a script can give the user foresight about the movie. Is it a horror movie which probably has words like die, kill, hell and some swear words? Or is it a comedy movie which has a lot of ‘laughs’ in paranthesis. Also, by comparing two files user can predict how much similar are these films.

In order to complete this project, a high level information about python data structures was necessary but most significant data type was dictionaries. Since amount of the repetation of the words should have been kept, by assigning name word as a key and its value to its repetation, keeping frequency of words was possible. Another structure was lists and sets and sometimes making convertion on these structures was necessary.

CHAPTER TWO

TASK SUMMARY

2.1 Completed Tasks

Removing stopwords from scripts and finding the frequency of the words for one script or finding the frequency of the common words for two scripts has been added. Also, loading from a file or from a website (imsdb) is implemented. Generating bar charts and tag cloud for visualization of the data is programmed.

CHAPTER THREE

EXPLANATION OF ALGORITHMS

3.1 Functions

Tabulate function: This function allows us to create proper looking tables in python terminal screen. It takes argumants as lists, sets or dicts and a string list for headers and returns a table.

Pyplot.bar function: This is a function of matplotlib library which lets you draw bar chart using your data. In this project, it was used for showing the frequencies of the words.

Pyplot.imshow: This is a function of matplot library which lets you show an image in a window. In this project, it was used for showing the tag cloud.

3.2 Algorithms and Solution Strategies

In order to get the script data from a website (imsdb), a python library called urllib is used. This library has functions to send a request to a webpage and get the data of it. After that, since the data take from urllib has html tags, words with or and <pre> or </pre> tags are removed. When the html tags are clean, the second part was cleaning punctuation signs. Afterwards, the text is splitted to a list of words by using built-in string.split() function.

The second part process in order to get the desired data is removing the stop words. In order to achieve that. I have imported two libraries with stop word lists; stop_words and nltk and extended them with each other to create a rich stopword pool. Actually, nltk has a function to find the frequencies of words but, we were asked to implement it without using nltk's that function.

CHAPTER FIVE

CONCLUSION

In this assigment, I have created a program which can be used by artists and normal theatre viewers. I have learnt a lot of things about matplotlib library and have a good fundamantal of data structures of python.