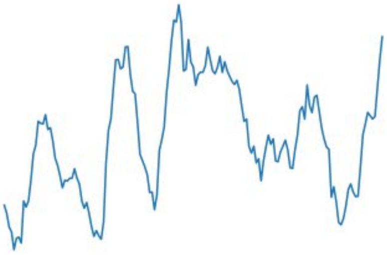
German University in Cairo
Media Engineering and Technology
Lecturer: Mervat AbuElkheir
TAs: Mohamed Abdelfattah

CSEN1076 Natural Language Processing and Information Retrieval

Spring term 2022

Project – Exploring Sentiment across movie scripts

Your project will be to build emotional arcs for movie-scripts using sentiment analysis. The task is to build a graph (similar to the one shown below)



that will determine the main arc of the movie plot based on the script analysis. There are 6 main arcs for any story. To do this, you will need to

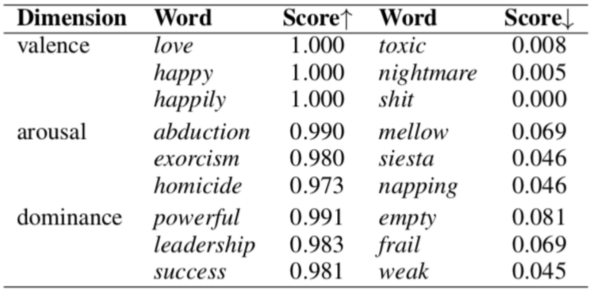
A) Collect and download all the movie scripts from imsdb.

B) Preprocess each movie script in order to remove extra spaces.

C) Extract a vector list of words that represent this particular movie script.

D) Using the NRC-VAD, you are going to map this vector into it's corresponding representation of valence, arousal and dominance vectors.

Entries with Highest and Lowest Scores in the VAD Lexicon



Dimension	Word	Score↑	Word	Score↓
valence	*love*	1.000	*toxic*	0.008
	happy	1.000	*nightmare*	0.005
	happily	1.000	*shit*	0.000
arousal	*abduction*	0.990	*mellow*	0.069
	exorcism	0.980	*siesta*	0.046
	homicide	0.973	*napping*	0.046
dominance	*powerful*	0.991	*empty*	0.081
	leadership	0.983	*frail*	0.069
	success	0.981	*weak*	0.045

E) Plot all 3 vectors onto the same figure (using different colors) and save to "movie_name.jpg"

A) Data Collection

You can use beautifulsoup to write a script to download all movie scripts from imsdb or alternatively, download it from this google-drive link here.

B) Data Preprocessing

In this part, you will be cleaning the text files before extracting a list of words from them. Make sure your preprocessing pipeline includes at least 3 steps e.g. (removing spaces, removing stopwords, removing punctuation)

C) Feature Extraction

In this part, you will make sure that each movie script has now been converted into a vector of filtered words. Please note that the number of words in each movie script is going to be different depending on your preprocessing technique.

D) VAD Vectorization

Converting each movie script's list of words into valence, arousal and dominance could be done manually using **map()** function or could be done using **emotion()** function using labMT's builtin method. If you're going to use the latter, make sure you replace the *happsList* and *labmt* variables with your own dictionary that you have read from NRC-VAD-Lexicon.txt. Since some of the words in the scripts do not have a corresponding value in the VAD dictionary, you can replace them with 0s. Finally, you now have **3 very large vectors**, that consist of 0s and other values that were replaced from the VAD dictionary. You need to strip down all the 0s from the vectors and average them using windows of size 500. So every 500 (non-zero) values will be replaced with a single value that represents the average.

E) Output

Now that you have 3 vectors for every movie script. Plot all 3 vectors onto the same figure (using any 3 different colors) and save that figure to a jpg file with the same name as the movie script. For instance, the corresponding figure for the movie "17 Again" would be "17 Again.jpg"

Submission

You may work in groups of 3-4 members. Deadline is on Tuesday, 7th of June at 23:59. You should submit a Jupyter notebook containing all your code and a folder with ALL OF THE .JPG files. Please use this link for submission.

Best of Luck!