Exploring Textual Data through Interactive Visualizations

Group Members:

M Talal Qureshi i212697

Abdur Rehman i21137

Dataset : Cornell Movie Dialogs Dataset

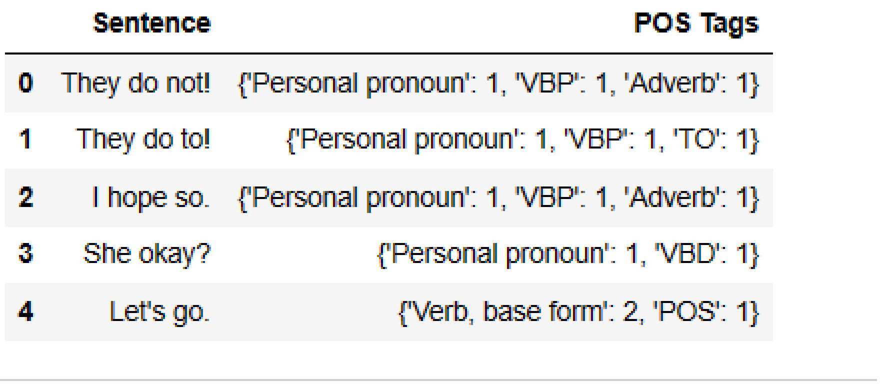
Natural Language Processing Analysis

1. Sentiment Analysis

We have performed a Sentiment Analysis by using SentimentIntensityAnalyzer from NLTK library. There is threshold value of 0.1 for the sentiment to be positive or negative, And then aggregate the result for each character and create a count of it, and draw an analysis that the which character has how much positive or negative in there speech.

2. NER

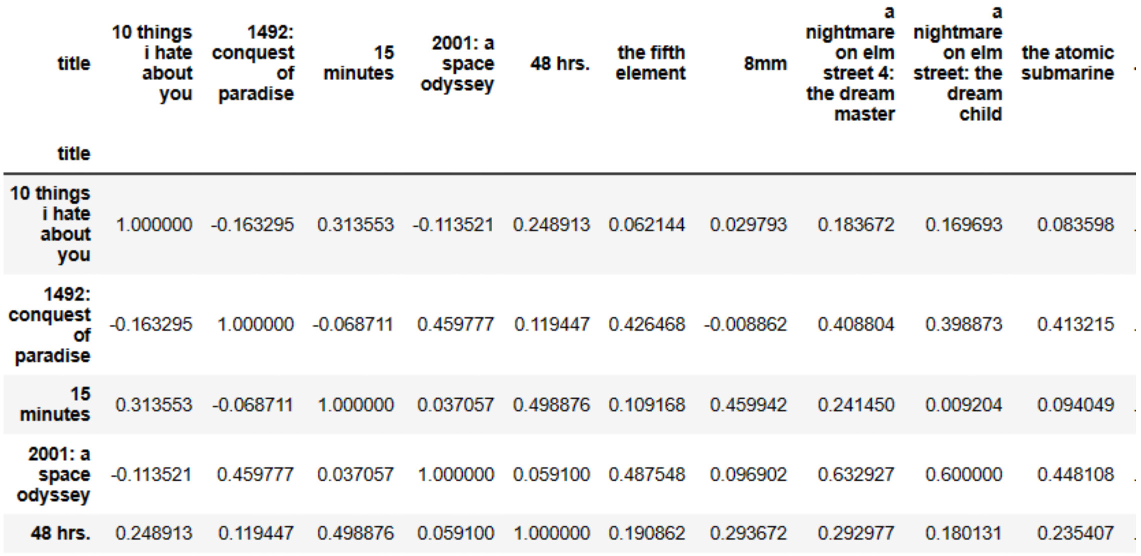
We have performed a Named Entity Recognition to extract Part of Speech from each Sentence and combine it with sentiment analysis to understand how different parts of speech contribute to the overall sentiment of the dialogues



	Sentence	POS Tags
0	They do not!	{'Personal pronoun': 1, 'VBP': 1, 'Adverb': 1}
1	They do to!	{'Personal pronoun': 1, 'VBP': 1, 'TO': 1}
2	I hope so.	{'Personal pronoun': 1, 'VBP': 1, 'Adverb': 1}
3	She okay?	{'Personal pronoun': 1, 'VBD': 1}
4	Let's go.	{'Verb, base form': 2, 'POS': 1}

3. Semantic Similarity

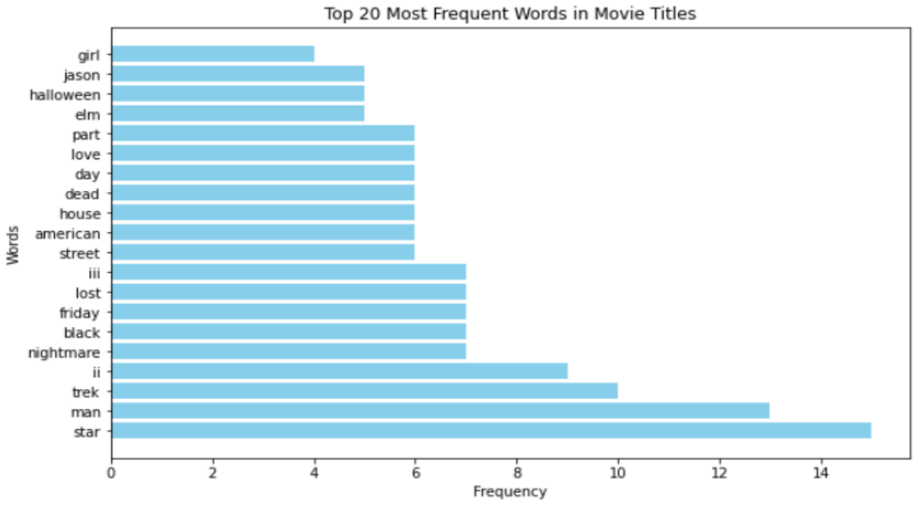
We have performed Semantic Similarity to Compute the semantic similarity between movie titles or genres using word embedding,to measure how closely related different movies are in terms of their titles or genres , for this we have used spacy model of en_core_web_md and then performed web embedding ,and then use cosine similarity to find similarity in genre of movies.



title	10 things i hate about you	1492: conquest of paradise	15 minutes	2001: a space odyssey	48 hrs.	the fifth element	8mm	a nightmare on elm street 4: the dream master	a nightmare on elm street: the dream child	the atomic submarine	.
title											
10 things i hate about you	1.000000	-0.163295	0.313553	-0.113521	0.248913	0.062144	0.029793	0.183672	0.169693	0.083598	.
1492: conquest of paradise	-0.163295	1.000000	-0.068711	0.459777	0.119447	0.426468	-0.008862	0.408804	0.398873	0.413215	.
15 minutes	0.313553	-0.068711	1.000000	0.037057	0.498876	0.109168	0.459942	0.241450	0.009204	0.094049	.
2001: a space odyssey	-0.113521	0.459777	0.037057	1.000000	0.059100	0.487548	0.096902	0.632927	0.600000	0.448108	.
48 hrs.	0.248913	0.119447	0.498876	0.059100	1.000000	0.190862	0.293672	0.292977	0.180131	0.235407	.

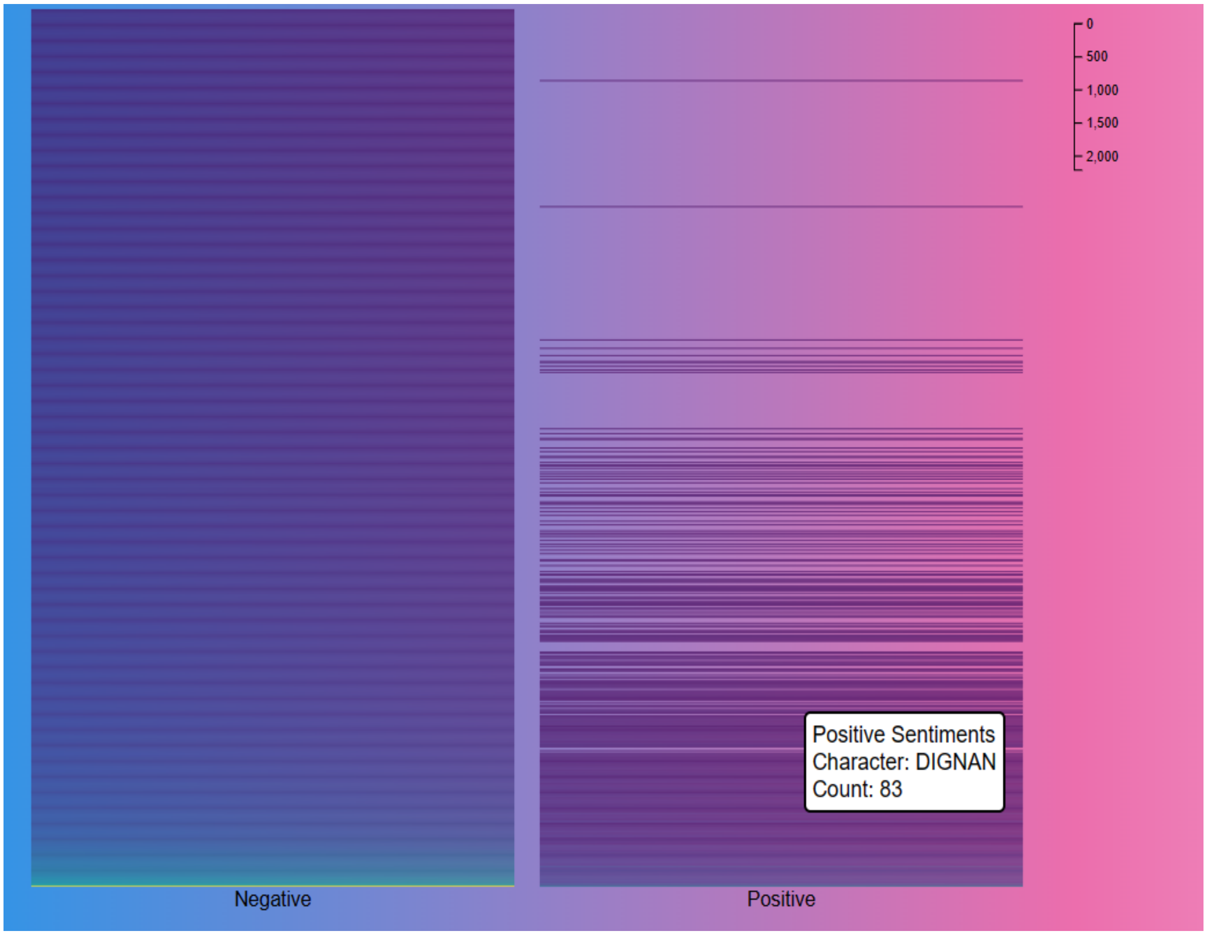
4. Title Analysis

We have performed Tite Analysis to compute analyses on the titles of the movies and count which frequency of word in Titles to find the frequent words in the movies titles.

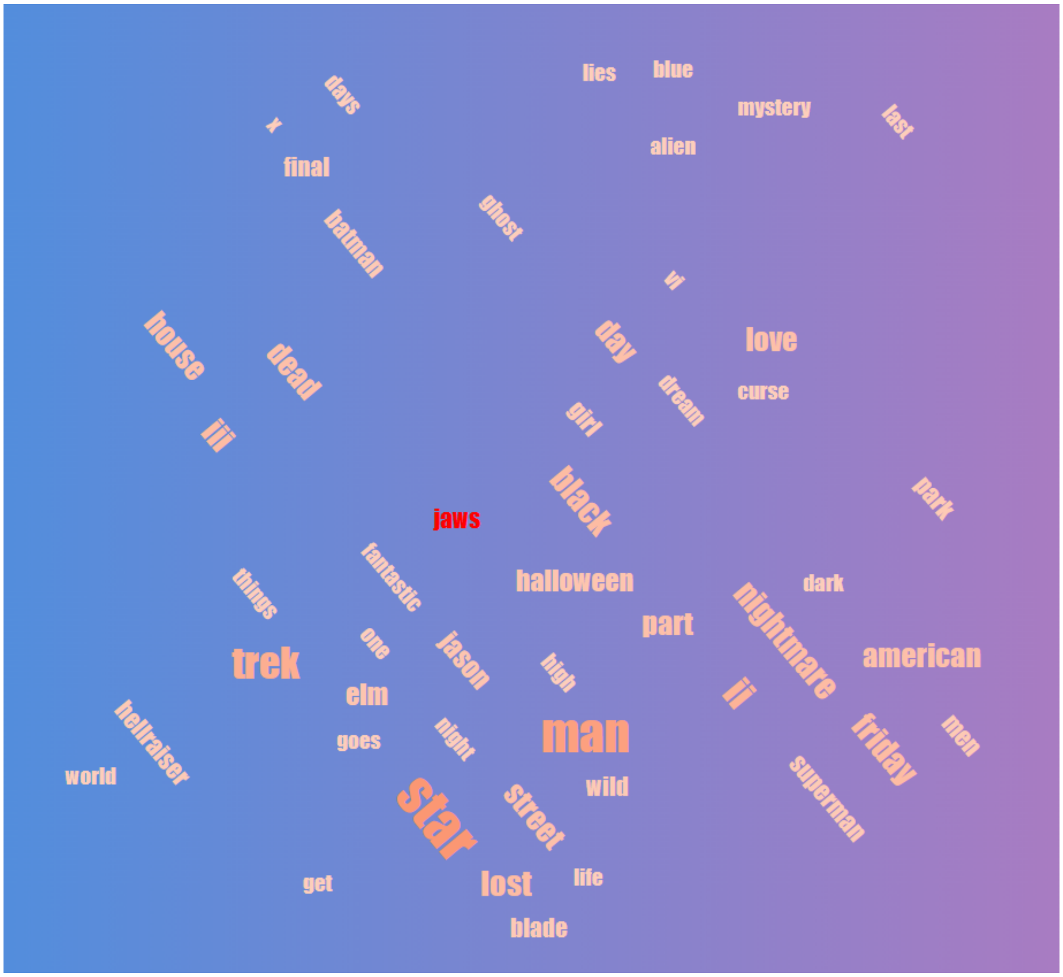


Visualizations:

1. Heat Map for Sentiment Analysis



2. Word Cloud :


lies
blue
days
mystery
last
x
alien
final
ghost
batman
vi
love
house
dead
day
curse
dream
girl
iii
black
park
jaws
fantastic
halloween
dark
things
nightmare
part
trek
one
jason
american
high
elm
ii
men
hellraiser
goes
night
man
friday
world
wild
superman
star
street
get
lost
life
blade

3. Force Directed Graph