Topic analysis of movies and TV series

Introduction

This project's goal is to analyse the most common topics present in movies and TV series, and to point out how linked they are to our society and its many problems.

Topic extraction

Our analysis is based on the english subtitles, processed through the following pipeline

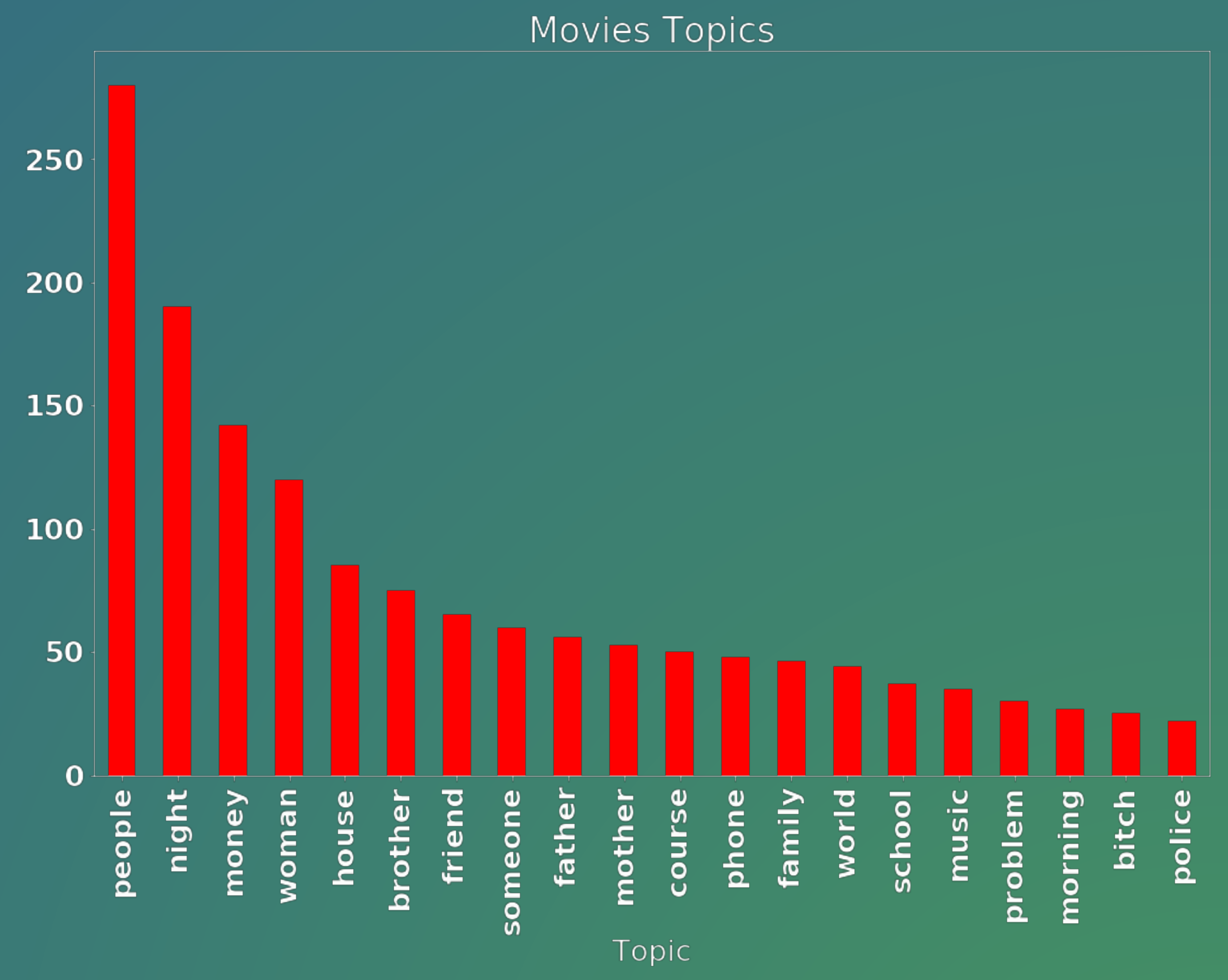
1. Remove small words (less than 4 characters)
2. Remove stop words
3. Remove character names
4. Lemmatise words
5. Remove the words that do not correspond to nouns

We then kept only the 5 most common words that appeared in the filtered subtitle, which correspond to our topics

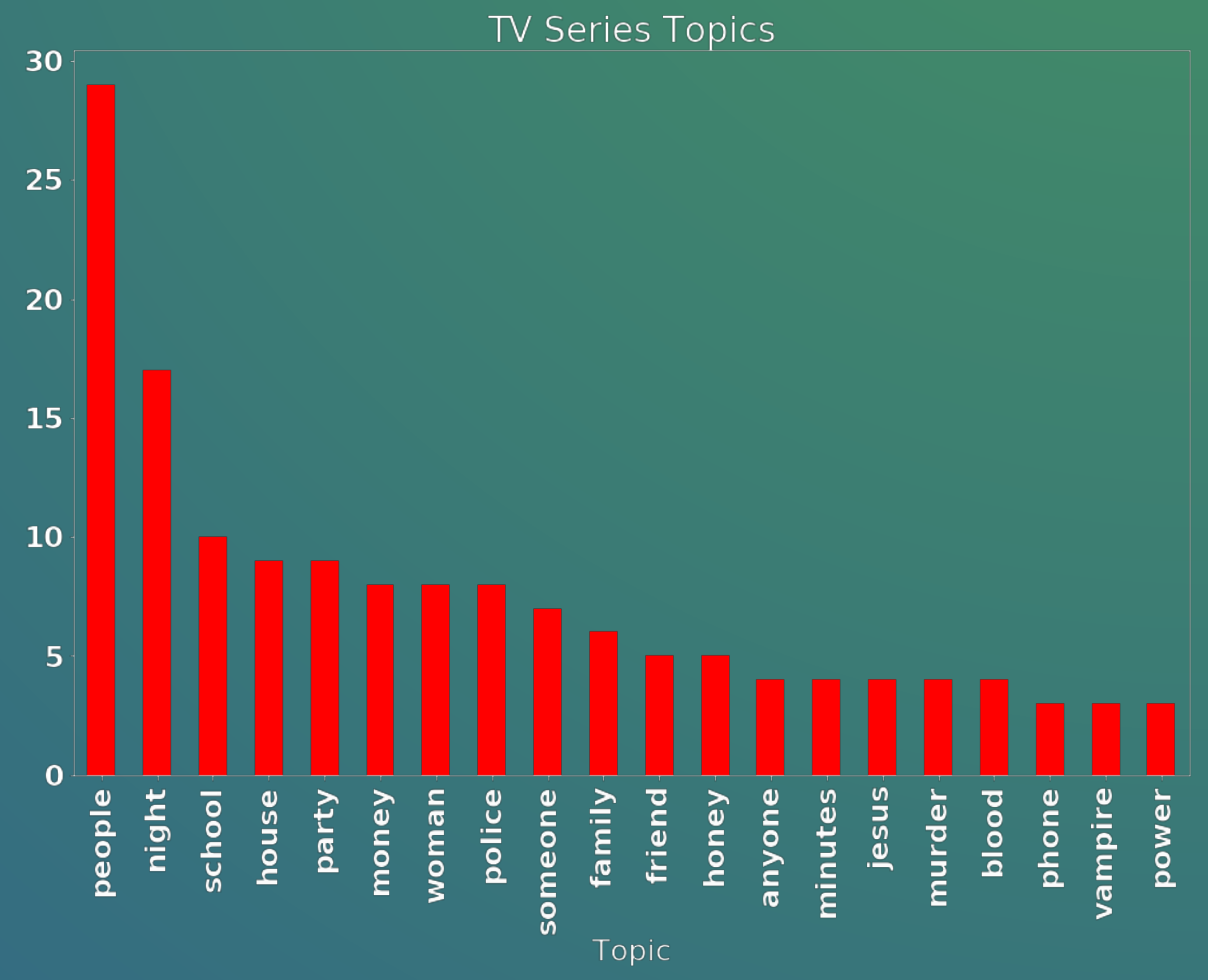
Topics success

Movies

- "People": is the most popular one, which isn't surprising as most of the subtitles consist of dialogues that should often involve references to people.



- "night": but our guess is that a lot of the scenarios have parts that take place at night and the character are likely to mention it.
- "money" is one of those who would expect in the top 3, as it is a recurrent, if not omni-present, in the capitalist society we live in.



TV Series

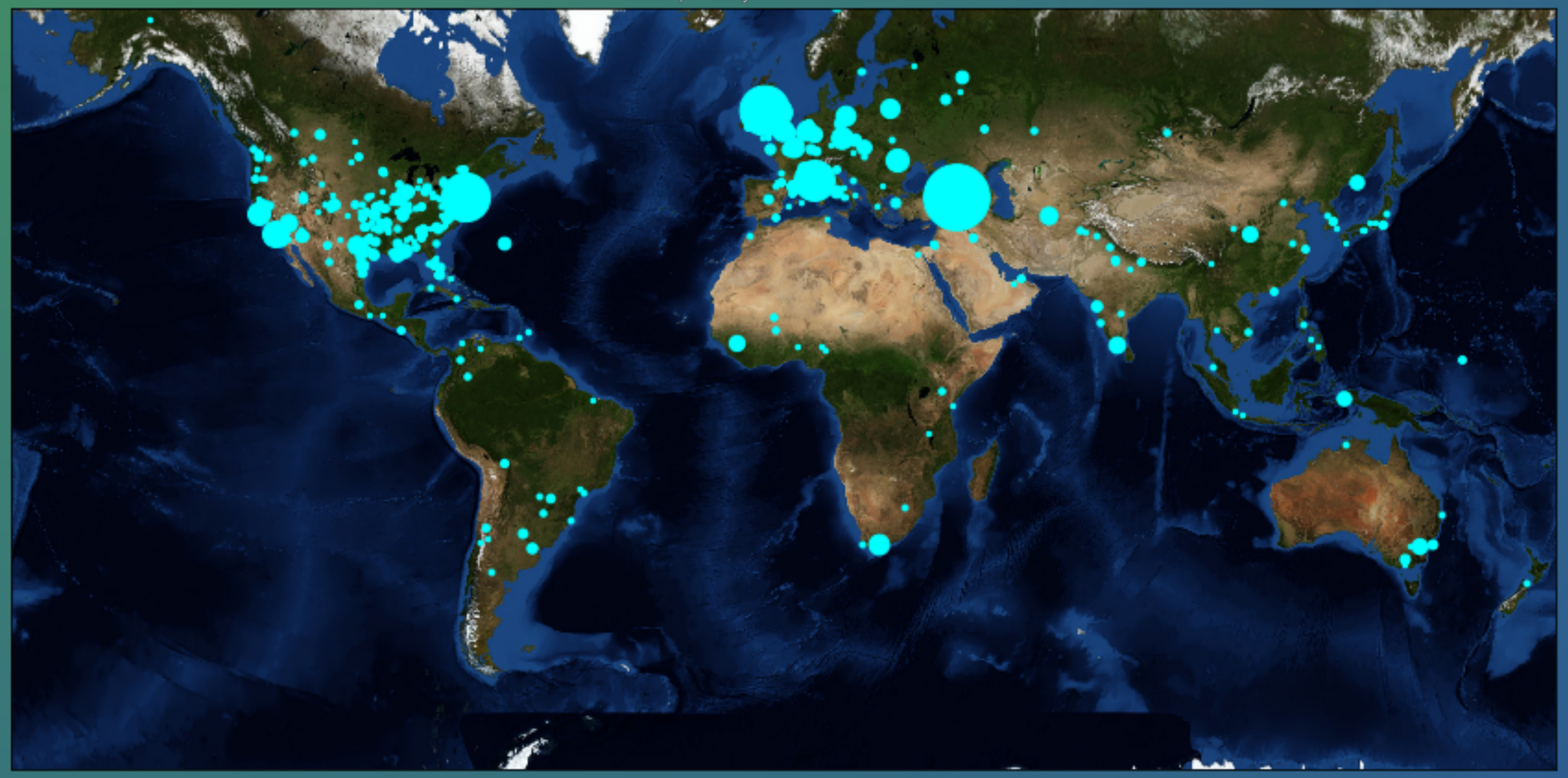
The TV series topics mostly align with the movie topics, we see some similarities - "people" and "night" being at the top. "Police" and "school" seem to be more prevalent in the series compared to the movies.

Locations success

Most of the mentioned locations are set either in the USA or in Europe, which makes sense as most of the movies and TV series are produced by those countries,

We can see a majority of mentions for the USA (New York especially), but we can still see some mentions of some other big cities around the world, which proves that TV Series are not only an American business. Considering that we analysed english subtitles, our data analysis might be biased to show more importance towards english speaking countries.

Most frequently mentioned locations



Conclusion

We can see that the popularity of the topics treated in movies not very surprising, they represent very well the society we currently live in. We have to note that due to data collection and filtering by number of votes our dataset was extremely reduced, and the missing data reduced it even more, therefore the results are to be taken with caution.