NLP Practical Project Documentation

Nenișcă Maria

Problem statement

The goal of this project is to perform sentiment analysis on the script of the animated TV series 'Rick & Morty'. The series has as protagonists the Smith family which is composed of:

- Rick: the grandfather who is a scientist and goes on multiple galaxy adventures with Morty
- Beth: the mother and Rick's daughter
- Jerry: the father
- Summer: the elder daughter
- Morty: the son who has a very close relationship with Rick

The goal is to determine the overall mood throughout the series, and, also, to get a glimpse into each character's personality by analyzing the lines each person says in the script.

Proposed solution

The proposed solution is to use the Bing and NRC lexicons in order to be able to perform sentiment and emotion analysis. The Bing lexicon maps words to two classes: negative and positive, whereas the NRC lexicon maps them to 10 classes: positive, negative, anger, anticipation, disgust, fear, joy, sadness, surprise, and trust.

The way in which we will determine the overall mood, as well as each character's mood is by analyzing the lines of dialogue and applying each lexicon in order to gather the necessary data. Moreover, we will do an unigram, bigram, and trigram analysis on the text in order to determine what are the most frequent words an combination of words used throughout the series in general and by each character.

Implementation

The project consists of a Jupyter notebook. For the implementation we used the following libraries:

- pandas – This helps as deal with the dataset by reading it and manipulating it. For example: applying different filters on the data set to extract the lines said by one character, creating new DataFrames that will help in our analysis.
- string - Used for the text processing, especially in the corpus cleaning part of the analysis.
- nltk.corpus.stopwords – We utilized the stop words available through the nltk library in the corpus cleaning
- nltk.tokenize.word_tokenize – We need a word tokenization function during corpus cleaning
- nltk.util.ngrams – Utilized in the unigram, bigram, and trigram analysis of the script.
- Collections.Counter - Utilized in the unigram, bigram, and trigram analysis of the script to count the frequencies of the constructions.
- from wordcloud import WordCloud, ImageColorGenerator – The wordcloud library is used to generate the word clouds for negative and positive words used in the script.
- numpy – A popular library which will help us in computing various variables needed in the analysis
- import nltk – A popular natural language processing library.

Experiments and results

The solution is divided into multiple parts as follows:

1. **The data** – here we read the necessary data: the script, the bing lexicon, and the NRC lexicon.
2. **Text mining** – here we define the functions that we will need when mining the text:
    a. clean_corpus(text) – this function has the role of cleaning the given text by removing punctuation, extra white spaces and numbers, transforming the text to lower case, and removing stop words. A bit of text pre-processing is done here, namely we make the following transformations: 'gonna'->'gon-na' and 'gotta'->'got-ta'. These two words are very frequent in the scripts (as the analysis shown) and the word_tokenize function from nltk split them as follows: 'gonna'->'gon na', 'gotta'->'got ta'. Because of this, the results were not accurate, so the pre-processing needed to be performed.
    b. frequentTerms(text) – returns a list with all the unigrams in the given text along with their frequencies
    c. frequentBigrams(text) - returns a list with all the bigrams in the given text along with their frequencies
    d. frequentTrigrams(text) - returns a list with all the trigrams in the given text along with their frequencies
3. **Dialogues: Who talks the most?** – in this section we performed an analysis on which characters have the most lines of words in the script. We plotted the first 15 characters with the most lines, and as we can see, Rick and Morty are the two characters who talk the most. This is expected as the story line mainly follows the adventures that these two have throughout galaxies and parallel universes.

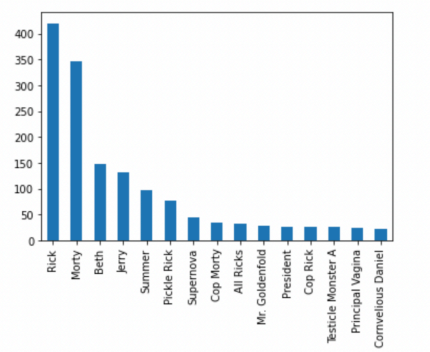


Figure 1 The amount of lines each character has in the script

4. **Bing Lexicon** – in this section we performed the sentiment analysis using the bing lexicon which maps the words to two classes: negative and positive. First, we generated a word cloud with the most frequent words used in the entire script:

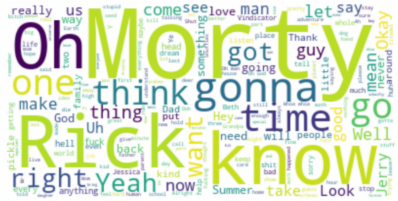


Figure 2 Word cloud with the words from the entire script

Then, we generated two more word clouds for negative and positive words:

Figure 3 Word cloud with positive words

Figure 4 Word cloud for positive words

Then, we wanted to see what the overall mood for each family member was. As we can observe in Figure 5, each character is pretty balanced, and we cannot say that we have negative or positive characters in the Smith family.

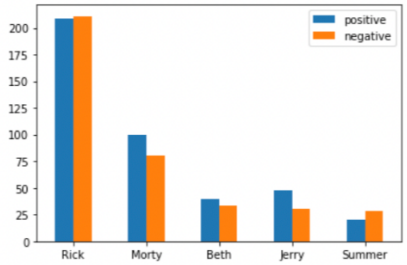


Figure 5 Mood per family member

5. **<u>NRC lexicon</u>** – now it is time to get a more in depth look into the mood of the series and each character's personality by analyzing their lines via the 10 sentiments available in the NRC lexicon. First, we analyzed the overall mood in the series. As can be observed in Figure 6, the overall moods are anticipation and anger. Anticipation was expected as the series has an action-based and adventurous story line. The anger and negative emotions have a high ranking because the majority of the family members do not have the best relationship with Rick, except Morty. Also, Rick has a colorful language and Morty tends to get frustrated and annoyed with rick during their adventures.

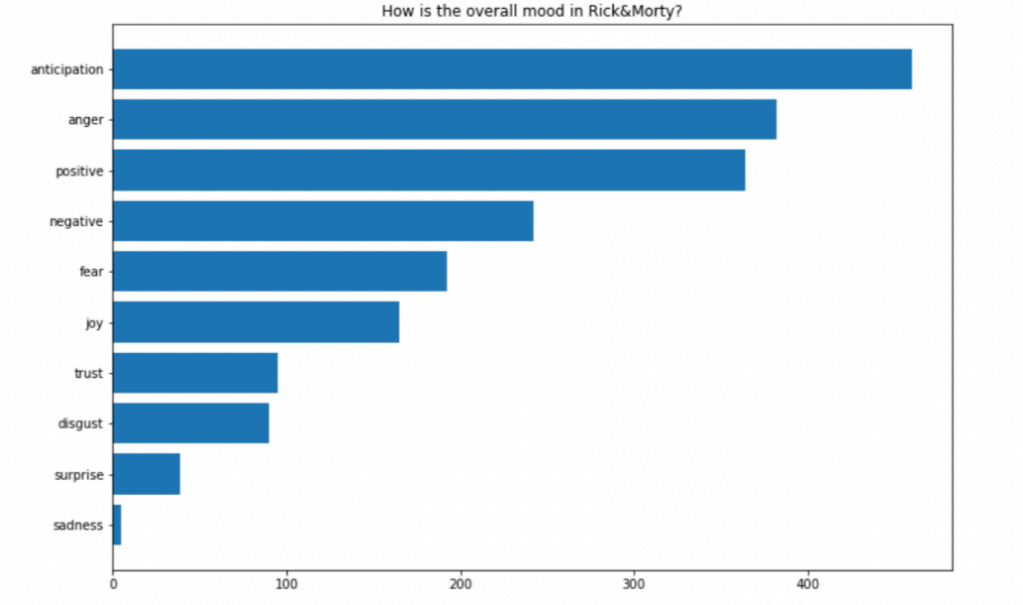
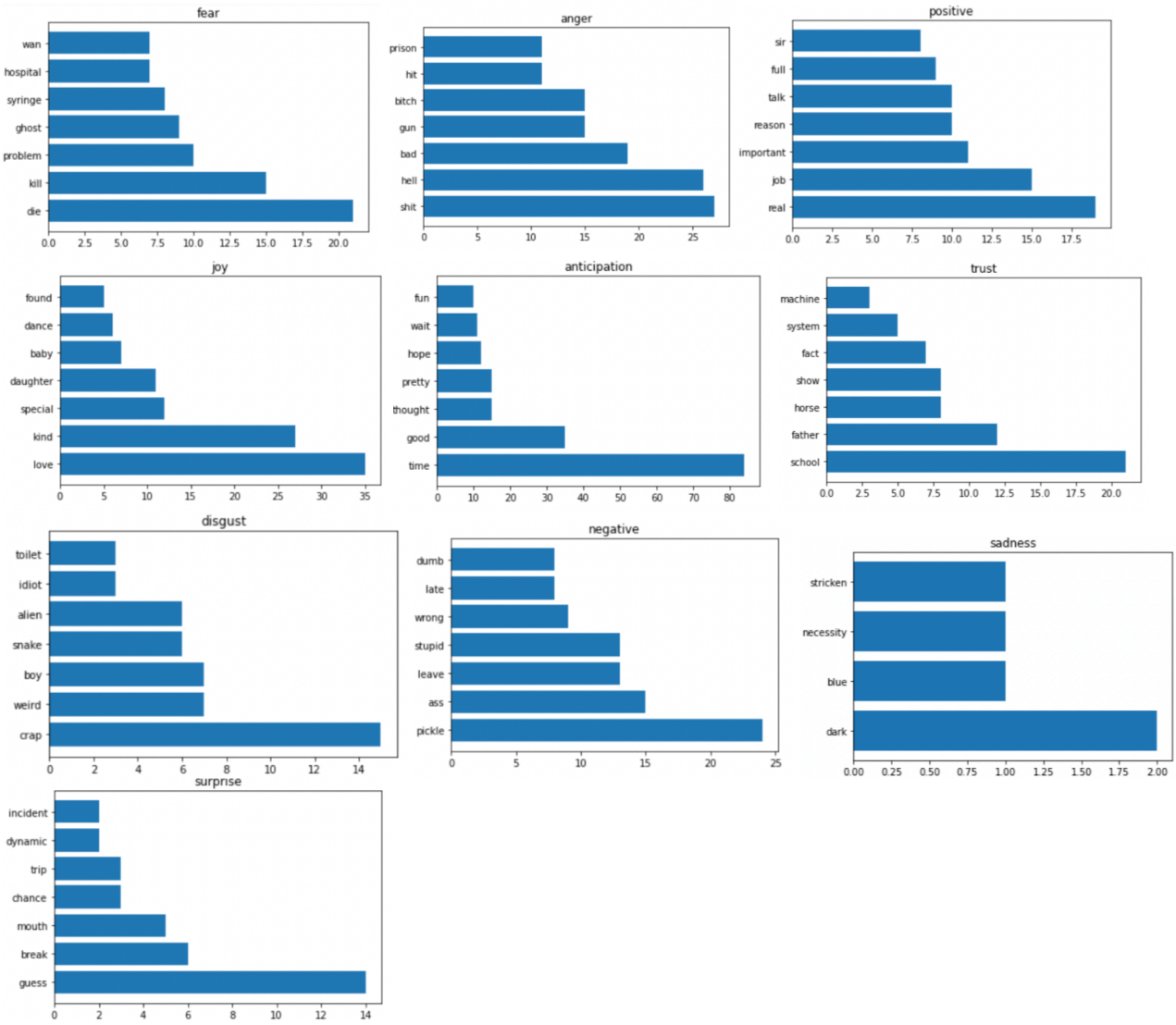


Figure 6 Overall mood in the series

Then we analyzed which words were the most frequently used overall in order to express each emotion from the NRC lexicon.



The most important graph and the one that showcases each character's parsonlaity is a radal chart, seen in Figure 7. There we represented the precentages of emotions each character expressed relative to their lines of dialogue. From there we can see that the ones that show the most anticipation are Summer and Jerry. Because Rick and Morty are the ones that go more frequently in adventures, it would be unnexpected to see this result. However the father and the daughter are the most enthusiastics ones with the 'bubbliest' personalities. Later in the series, the whole family gets included in the space adventures and this contributes as well to the level of anticipations shown by those two characters. As mentioned before and confirmed by this chart, Morty displayes the most anger, which is an expected results. Being the youngest one and of inferior intelligence than the rest of the family members, he gets frustrated often and lashes on different characters, especially Rick. He is a tipical teenager who just got into the first year of high school and that behaviour is expected.

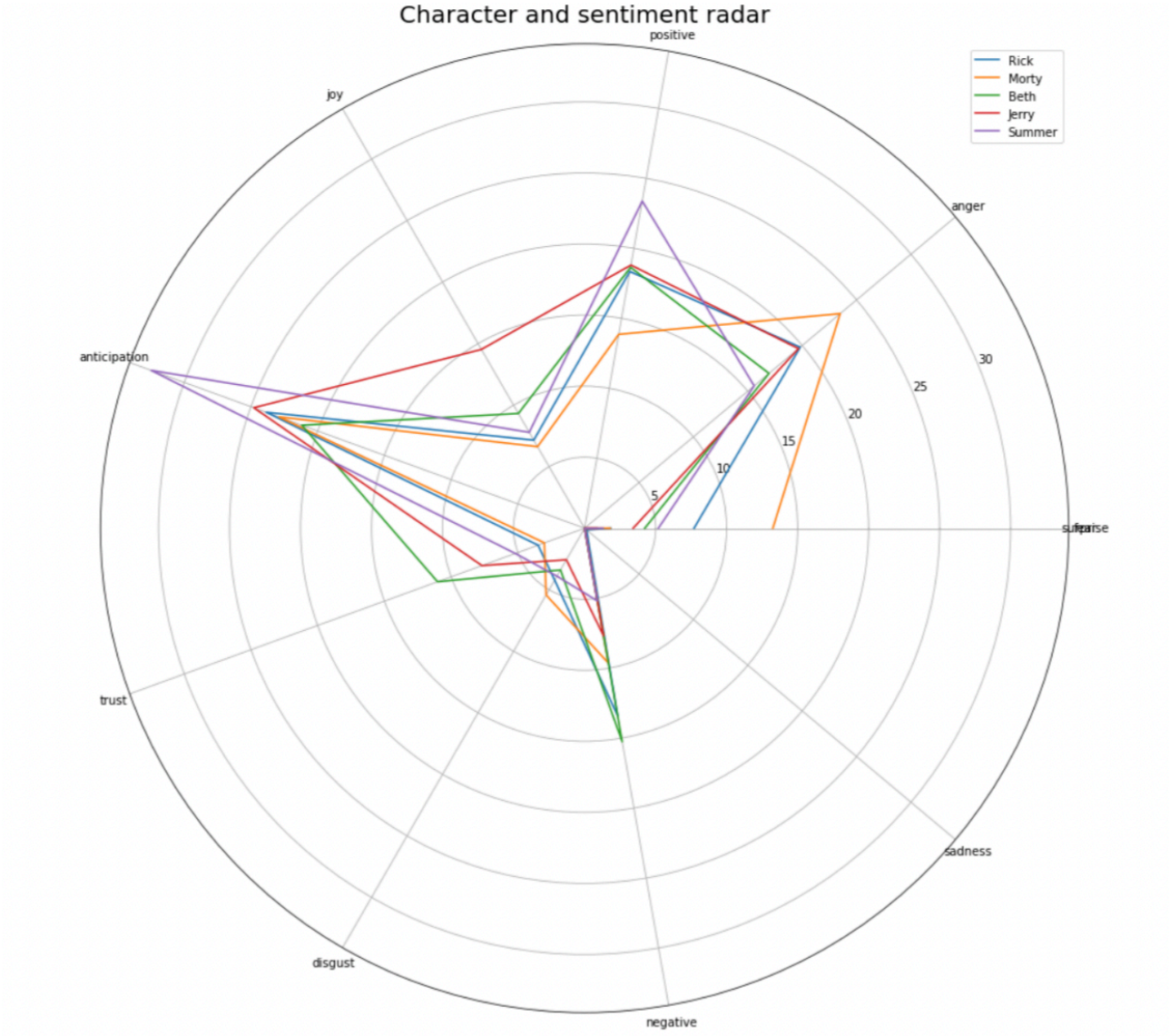
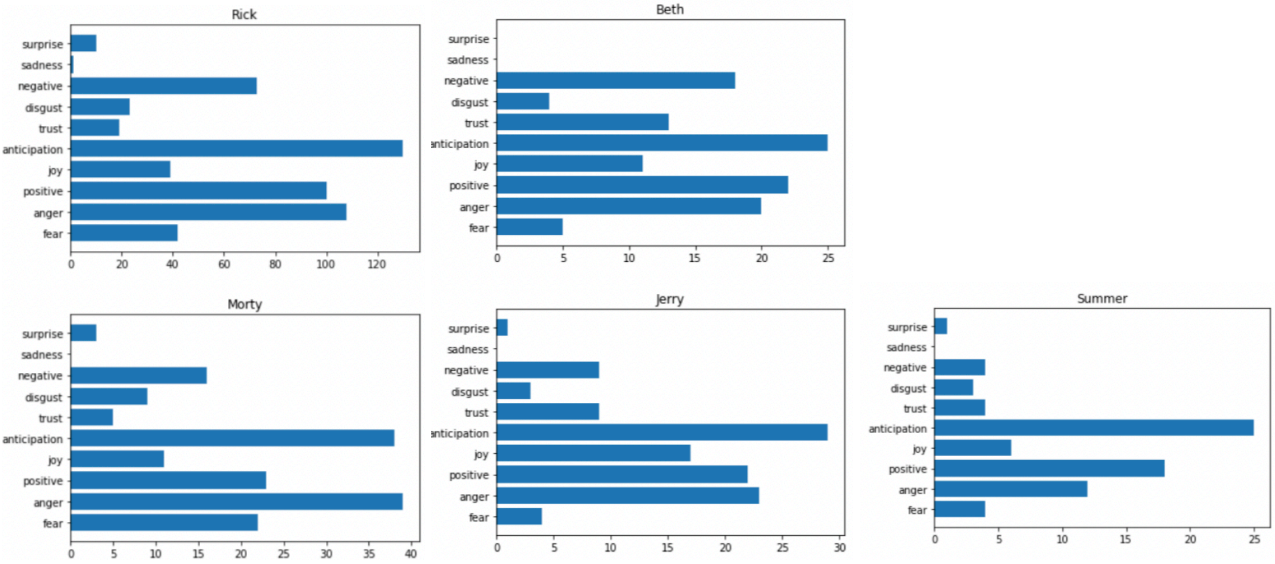
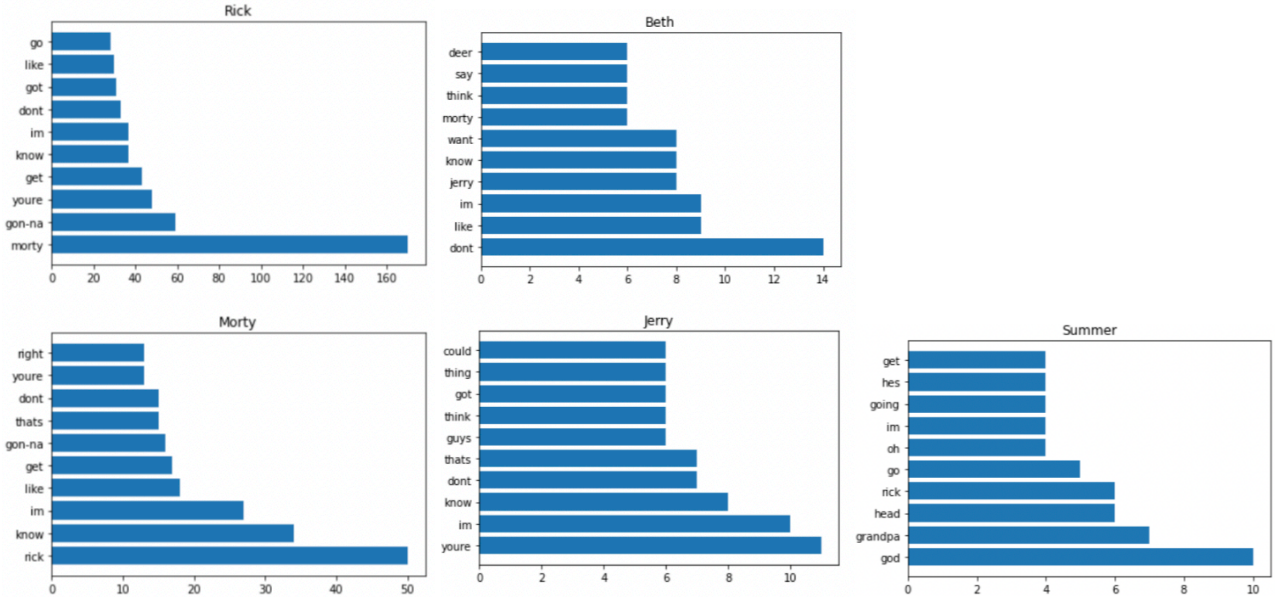


Figure 7 Radal chart for each character's sentiments

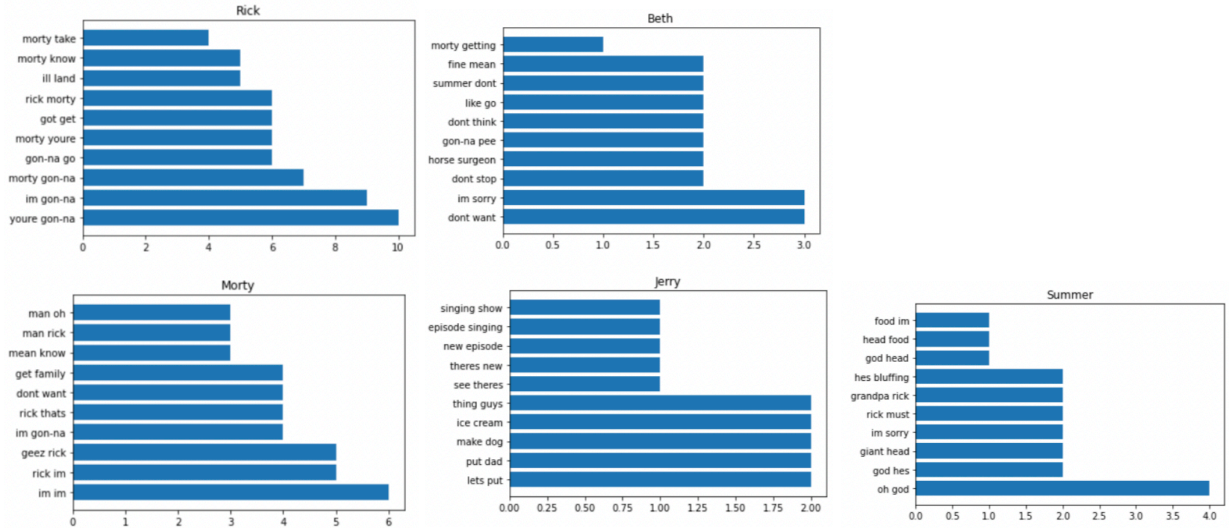
Other important graphs for getting an idea about the personality of each character are those that show for each member the frequency in which they show their emotions by number of lines said. Judging by the number of lines that express one emotion, we can see that the most popular emotion is anticipation for reasons mentioned before.



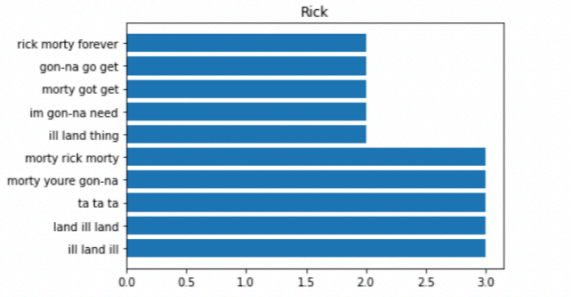
6. **<u>Unigram: Most Frequent words per Smith Family Member</u>** – here we analyzed what were the most frequent words said by each character. We can see that Rick and Morty address each other *very* often. Beth is usually the one that tries to stop different characters from doing something and her most used unigram is ‘dont’.

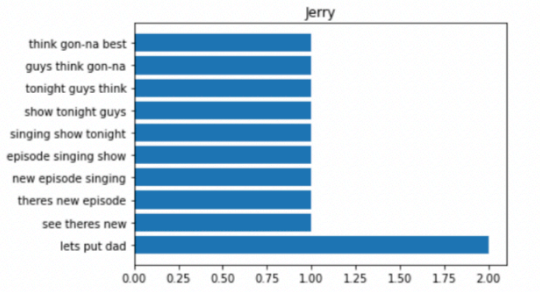


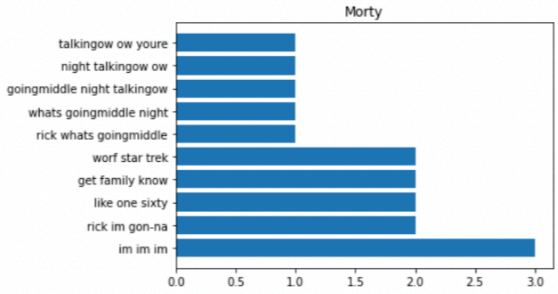
7. **<u>The Bigram and Trigram: most frequent combinations per family member</u>**: - From the most used bigrams we can see that Rick is very assertive and likes to tell others what to do, especially Morty. His most used bigram is ‘you’re gonna’, and the others have a very imperative tone: ‘morty take’, ‘morty youre’, ‘morty gonna’. Again, we can see that Beth opposes a lot of ideas and action that come from other members by expressing that she doesn’t want something to happen through ‘don’t want’, ‘summer don’t’. We can see that she says her profession a lot: ‘horse surgeon’.

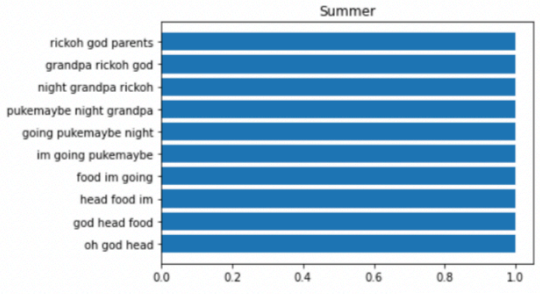


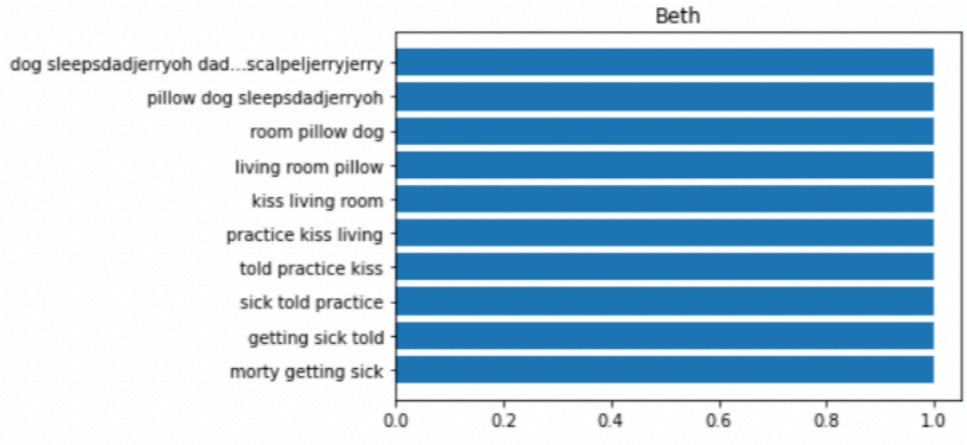
Looking at the most used trigrams by each characters we can see again that Rick tells Morty what to do a lot and that he is landing his space craft often: ‘land I’ll land’, ‘I’ll land I’ll’. We can see that Jerry likes a singing show very much and wants the entire family to watch it with him: ‘new episode singing’, ‘singing show tonight’, ’show tonight guys’.


Rick
rick morty forever
gon-na go get
morty got get
im gon-na need
ill land thing
morty rick morty
morty youre gon-na
ta ta ta
land ill land
ill land ill
0.0
0.5
1.0
1.5
2.0
2.5
3.0


Jerry
think gon-na best
guys think gon-na
tonight guys think
show tonight guys
singing show tonight
episode singing show
new episode singing
theres new episode
see theres new
lets put dad
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00


Morty
talkingow ow youre
night talkingow ow
goingmiddle night talkingow
whats goingmiddle night
rick whats goingmiddle
worf star trek
get family know
like one sixty
rick im gon-na
im im im
0.0
0.5
1.0
1.5
2.0
2.5
3.0


Summer
rickoh god parents
grandpa rickoh god
night grandpa rickoh
pukemaybe night grandpa
going pukemaybe night
im going pukemaybe
food im going
head food im
god head food
oh god head
0.0
0.2
0.4
0.6
0.8
1.0


Beth
dog sleepsdadjerryoh dad...scalpeljerryjerry
pillow dog sleepsdadjerryoh
room pillow dog
living room pillow
kiss living room
practice kiss living
told practice kiss
sick told practice
getting sick told
morty getting sick
0.0
0.2
0.4
0.6
0.8
1.0