INTELLIGENCE SYSTEMS NLP PROJECT

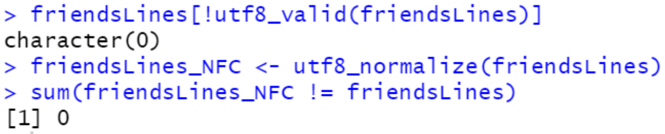
DOGA CENGIZ - https://github.com/dogacengiz

INTRODUCTION

In this paper, application of some natural language processing techniques to a dataset that contains transcripts of the series called "Friends" were explained and the results were presented. The purpose of the study is to analyze the sentence structure and to discover the sentiment of the sentences of each character. Thereby, it can be seen which character is more positive or negative. Also, according to the sentences that they create, a BPE model is create and a clustering algorithm applied to group the characters who are closer to each other. There are 6 main characters in the series and during the NLP experiment we will focus on these characters. All the experiments were held by using R programming language and its packages for NLP [1].

METHODOLOGY AND RESULTS

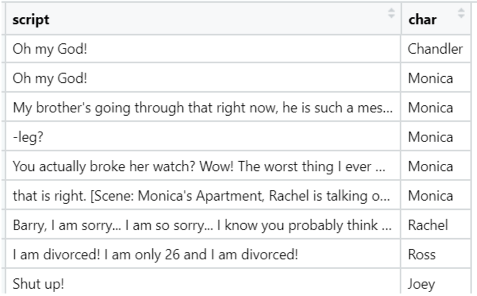
The transcripts of all the episodes found in a website [2]. Only the first season that consists of 24 episodes was used due to the computational limitations. This transcript contains all the cue with the name of their characters. Firstly, the scripts were saved in a .txt file then they were imported to R environment by using readLines() method and the encoding of the text was checked, as it can be seen it the Figure 1, the encoding didn't have any problem. The data was converted into data frame structure. A characters list was created with the name of the characters. A new column was added into the data frame to store the name of the corresponding character and then the name of the character was removed from the script. Thereby, a data frame with the name of the char and its script was obtained as shown in the Table 1. The rows that are not belonging to any main 6 character were removed. After that the scripts were grouped by the characters as in Table 2.



```
> friendsLines[!utf8_valid(friendsLines)]
character(0)
> friendsLines_NFC <- utf8_normalize(friendsLines)
> sum(friendsLines_NFC != friendsLines)
[1] 0
```

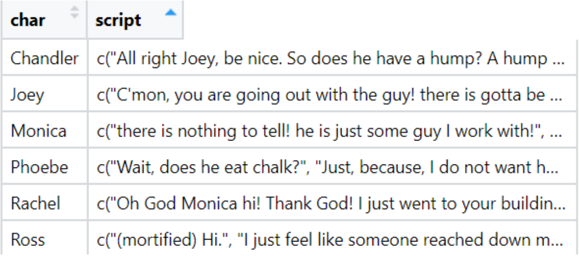
Figure 1 Encoding checks

Table 1 Data frame with characters



script	char
Oh my God!	Chandler
Oh my God!	Monica
My brother's going through that right now, he is such a mes...	Monica
-leg?	Monica
You actually broke her watch? Wow! The worst thing I ever ...	Monica
that is right. [Scene: Monica's Apartment, Rachel is talking o...	Monica
Barry, I am sorry... I am so sorry... I know you probably think ...	Rachel
I am divorced! I am only 26 and I am divorced!	Ross
Shut up!	Joey

Table 2 Grouped data frame



char	script
Chandler	c("All right Joey, be nice. So does he have a hump? A hump ...
Joey	c("C'mon, you are going out with the guy! there is gotta be ...
Monica	c("there is nothing to tell! he is just some guy I work with!", ...
Phoebe	c("Wait, does he eat chalk?", "Just, because, I do not want h...
Rachel	c("Oh God Monica hi! Thank God! I just went to your buildin...
Ross	c("(mortified) Hi.", "I just feel like someone reached down m...

spacyr package was used to tokenize the text and obtain the length of the sentences and the frequency of the words that each character has. Firstly, spacy_tokenize() method was used by setting what=”sentence”.

The histogram of the sentences for each character can be seen below, in Figure 2. The plots show that most of the time the length of the sentence is no longer than 50 characters. The maximum length is around 150 characters which does not happen frequently. For example, plots shows that Ross has not use any sentence that exceeds 150 characters during the first season of the series.

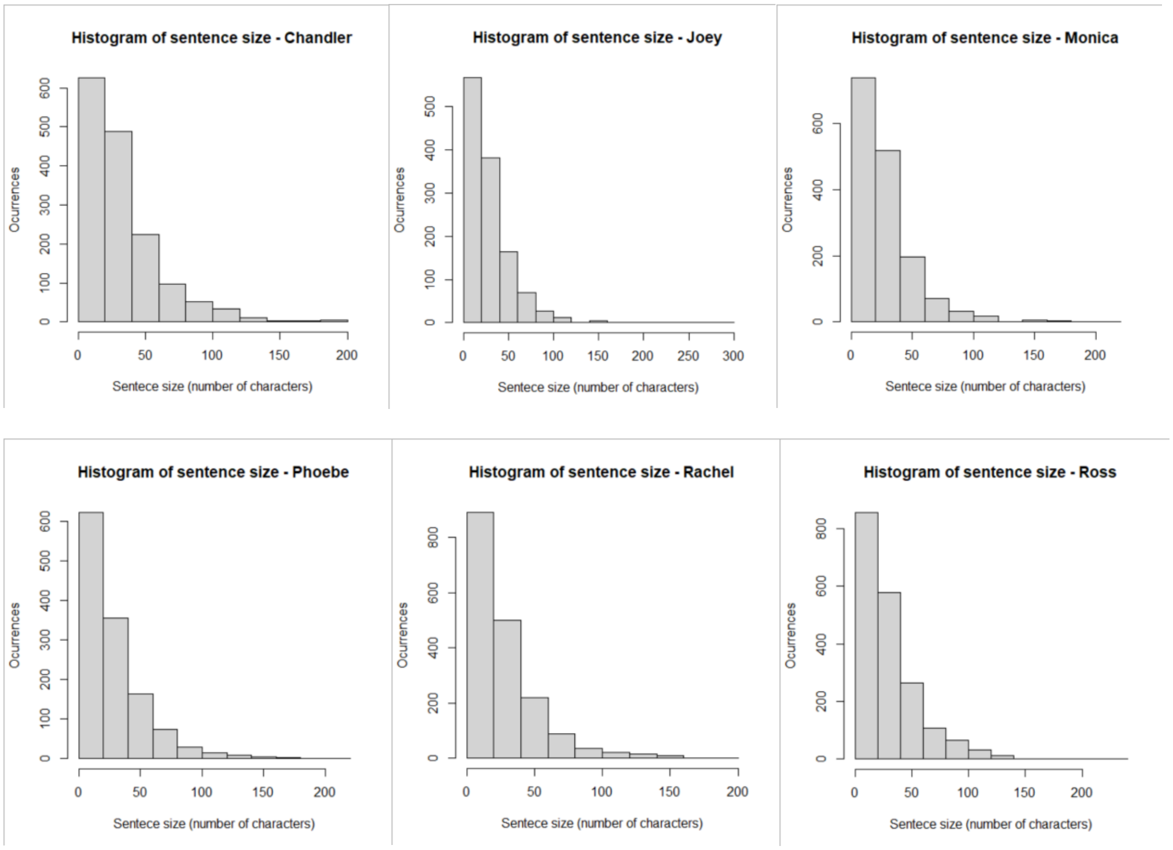
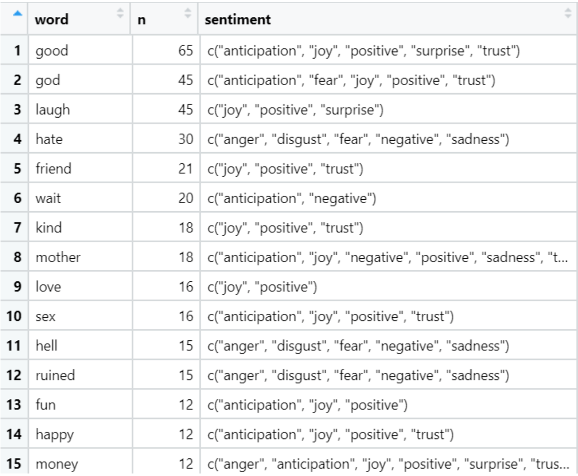


Figure 2 Histogram of size of sentences

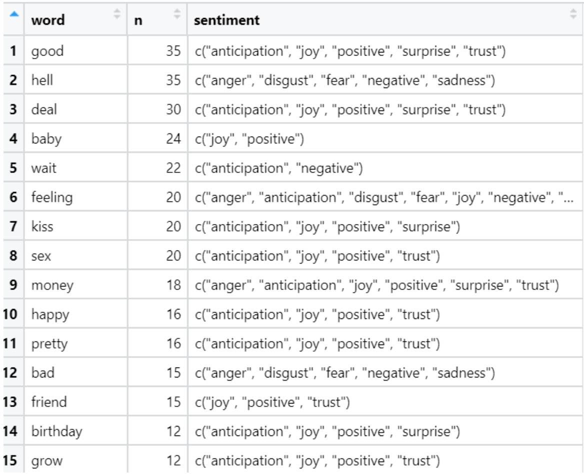
To analyze the sentiments of the sentences, get_sentiments() method used with “nrc” option. This method brings a data frame that has 13.875 words with their sentiments. By using inner_join(), a table that contains all the tokens of the series character was merged with the “nrc” data frame of sentiments. Thereby, we can see the emotions of characters they usually in. Table3 and Table 4 shows the sentiment analysis results of Monica and Joey respectively. It can be seen that in the first 15 frequently used sentimental words of Monica the first three words are positive but in the list of Joey the second most frequently used word contains negative sentiment such as anger.

Table 3 Sentiment analysis of Monica



	word	n	sentiment
1	good	65	c("anticipation", "joy", "positive", "surprise", "trust")
2	god	45	c("anticipation", "fear", "joy", "positive", "trust")
3	laugh	45	c("joy", "positive", "surprise")
4	hate	30	c("anger", "disgust", "fear", "negative", "sadness")
5	friend	21	c("joy", "positive", "trust")
6	wait	20	c("anticipation", "negative")
7	kind	18	c("joy", "positive", "trust")
8	mother	18	c("anticipation", "joy", "negative", "positive", "sadness", "t...
9	love	16	c("joy", "positive")
10	sex	16	c("anticipation", "joy", "positive", "trust")
11	hell	15	c("anger", "disgust", "fear", "negative", "sadness")
12	ruined	15	c("anger", "disgust", "fear", "negative", "sadness")
13	fun	12	c("anticipation", "joy", "positive")
14	happy	12	c("anticipation", "joy", "positive", "trust")
15	money	12	c("anger", "anticipation", "joy", "positive", "surprise", "trus...

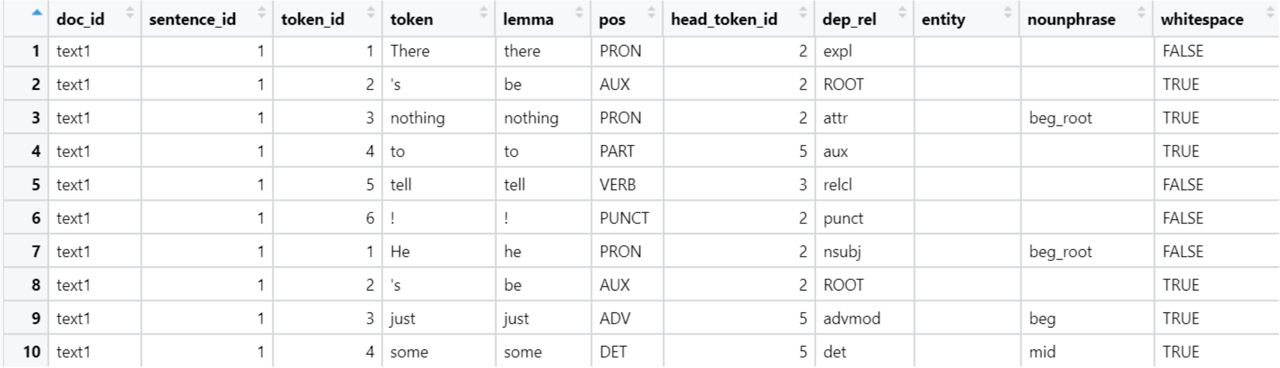
Table 4 Sentiment analysis of Joey



	word	n	sentiment
1	good	35	c("anticipation", "joy", "positive", "surprise", "trust")
2	hell	35	c("anger", "disgust", "fear", "negative", "sadness")
3	deal	30	c("anticipation", "joy", "positive", "surprise", "trust")
4	baby	24	c("joy", "positive")
5	wait	22	c("anticipation", "negative")
6	feeling	20	c("anger", "anticipation", "disgust", "fear", "joy", "negative", "...
7	kiss	20	c("anticipation", "joy", "positive", "surprise")
8	sex	20	c("anticipation", "joy", "positive", "trust")
9	money	18	c("anger", "anticipation", "joy", "positive", "surprise", "trust")
10	happy	16	c("anticipation", "joy", "positive", "trust")
11	pretty	16	c("anticipation", "joy", "positive", "trust")
12	bad	15	c("anger", "disgust", "fear", "negative", "sadness")
13	friend	15	c("joy", "positive", "trust")
14	birthday	12	c("anticipation", "joy", "positive", "surprise")
15	grow	12	c("anticipation", "joy", "positive", "trust")

Table 5 shows a sentence analysis. The first 100 phrases of Monica was used to perform this analysis.

Table 5 Sentence analysis



	doc_id	sentence_id	token_id	token	lemma	pos	head_token_id	dep_rel	entity	nounphrase	whitespace
1	text1	1	1	There	there	PRON	2	expl			FALSE
2	text1	1	2	's	be	AUX	2	ROOT			TRUE
3	text1	1	3	nothing	nothing	PRON	2	attr		beg_root	TRUE
4	text1	1	4	to	to	PART	5	aux			TRUE
5	text1	1	5	tell	tell	VERB	3	relcl			FALSE
6	text1	1	6	!	!	PUNCT	2	punct			FALSE
7	text1	1	1	He	he	PRON	2	nsubj		beg_root	FALSE
8	text1	1	2	's	be	AUX	2	ROOT			TRUE
9	text1	1	3	just	just	ADV	5	advmod		beg	TRUE
10	text1	1	4	some	some	DET	5	det		mid	TRUE

After analyzing the sentences, BPE model was created by using bpe() method and choosing the phrases of Monica and Joey. Quanteda library was used to calculate the differences between the texts of each character. Then the corpus was converted into dfm by using dfm() method. This dfm object was used to create the dendrogram of the corpus. While dendrogram was generating method for distance was selected as “euclidean”. Hierarchical clustering technique was used with “ward.D” method to create the clusters. In the Figure 3 the dendrogram of the characters of the series can be seen. Table 6 shows the name of the characters, and the label represents it in the dendrogram. According to the dendrogram, we can say that the Ross have the most distance to the rest of the group. On the other hand, Joey-Phoebe and Chandler-Monica pairs are the ones closer to each other.

Table 6 Characters name matching

Chandler	Joey	Monica	Phoebe	Rachel	Ross
Char1	Char2	Char3	Char4	Char5	Char6

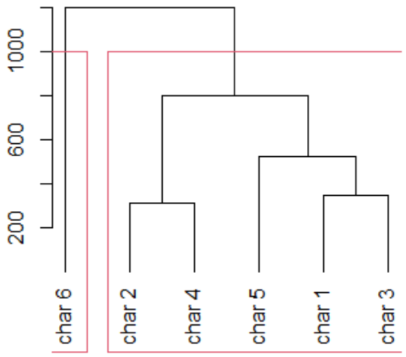
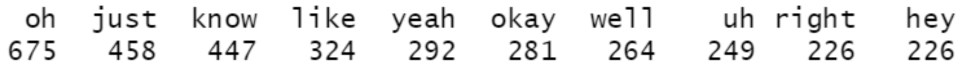


Figure 3 Dendrogram of the corpus

In Figure 4, we can see the most used in the first season of the series by 6 main characters when we remove the stop words. In order to remove the stop words, dfm_remove() method used by selecting the stop word language as English, stopwords("en").



```
  oh  just  know  like  yeah  okay  well    uh right   hey 
 675   458   447   324   292   281   264   249   226   226 
```

CONCLUSION

As it explained above, text processing, text analysis, sentiment analysis and text clustering algorithms were used in the project. The results of the sentiment analysis show us insights about the emotional profile of each main characters of the series. Moreover, results of the clustering shoed that Chandler and Monica have similarities to each other that cause a smaller distance between them, and we know that this two character got married in the next seasons. Also, Ross who is the char6 in the dendrogram, has different lifestyle then the other characters in the series therefore more distances between him and others can be explained by that.

In conclusion, NLP techniques help on the character analysis of a series. Also, these methods can be applied to the real life by using the conversation channels such as email, messages to obtain more information about people, to analyze the emotions during a period of a time, and to calculate the distances to other people.

References

[1] "R: The R Project for Statistical Computing," [Online]. Available: https://www.r-project.org/. [Accessed 30 01 2022].

[2] "Friends Transcripts," [Online]. Available: https://fangj.github.io/friends/. [Accessed 29 01 2022].