REPORT

Comp 472

Department of Computer Science and Software Engineering

Mini Project 2

Concordia University

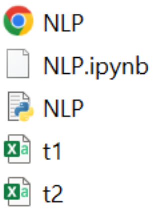
Julia Bazarbachian

2678137

April 22, 2022

The goal of Mini Project 2 was to practice Natural Language Processing on text snippets, produce NER and dependency graphs, and perform Sentiment Analysis on the tokenized text. SpaCy was used to preprocess text documents and produce the necessary graphs. The Afinn Lexicon was used for Sentiment Analysis and Scikit-Learn was used to perform k-means clustering.

Submitted Files:



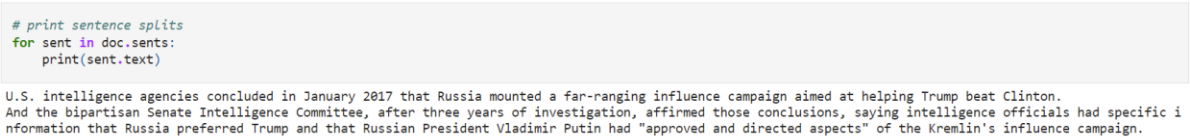
NLP
NLP.ipynb
NLP
t1
t2

NLP.html is the best view the project code and graphs. The code is also submitted in *.ipynb* and *.py* formats. Csv files t1 and t2 show the contents of the T1 and T2 tables for text snippet S1.

Submitted modules:

<u>(a) SpaCy sentence and token splits for S1</u>

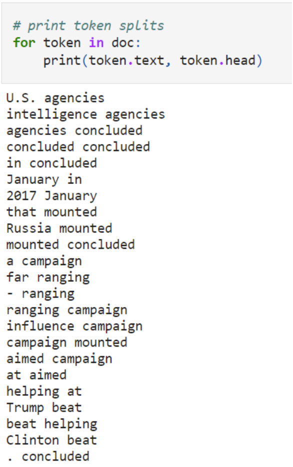
Output of Snippet 1 (S1) preprocessed and split into sentences:



```
# print sentence splits
for sent in doc.sents:
    print(sent.text)
```

```
U.S. intelligence agencies concluded in January 2017 that Russia mounted a far-ranging influence campaign aimed at helping Trump beat Clinton.
And the bipartisan Senate Intelligence Committee, after three years of investigation, affirmed those conclusions, saying intelligence officials had specific i
nformation that Russia preferred Trump and that Russian President Vladimir Putin had "approved and directed aspects" of the Kremlin's influence campaign.
```

Output of S1 preprocessed and split into tokens:

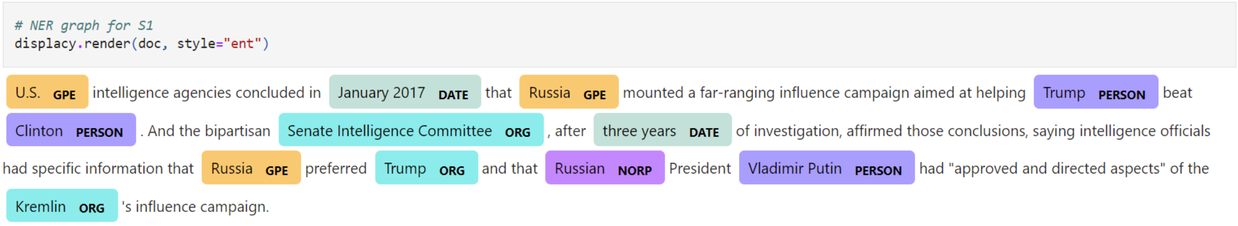


```
# print token splits
for token in doc:
    print(token.text, token.head)
```

```
U.S. agencies
intelligence agencies
agencies concluded
concluded concluded
in concluded
January in
2017 January
that mounted
Russia mounted
mounted concluded
a campaign
far ranging
- ranging
ranging campaign
influence campaign
campaign mounted
aimed campaign
at aimed
helping at
Trump beat
beat helping
Clinton beat
. concluded
```

(b) NER and dependency graphs for S1

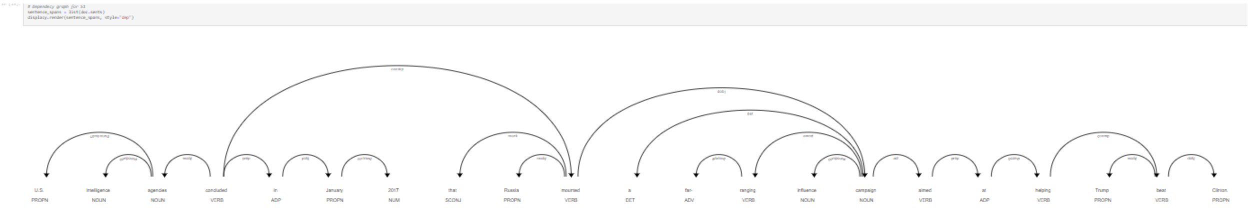
NER graph for S1:



```
# NER graph for S1
displacy.render(doc, style="ent")
```

U.S. GPE intelligence agencies concluded in January 2017 DATE that Russia GPE mounted a far-ranging influence campaign aimed at helping Trump PERSON beat Clinton PERSON . And the bipartisan Senate Intelligence Committee ORG , after three years DATE of investigation, affirmed those conclusions, saying intelligence officials had specific information that Russia GPE preferred Trump ORG and that Russian NORP President Vladimir Putin PERSON had "approved and directed aspects" of the Kremlin ORG 's influence campaign.

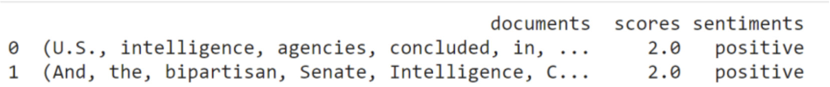
Dependency graph for the first sentence of S1: (the second graph can be found in the code, it was too large to be included here)



```
# Dependency graph for S1
sentence_spans = list(doc.sents)
displacy.render(sentence_spans, style="dep")
```

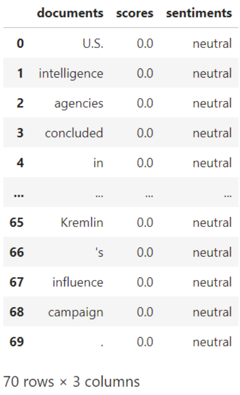
(c) T1$_{S1}$ and T2$_{S1}$ for S1

First, sentiment Analysis is performed for sentences and tokens. Both sentences in S1 were classified as positive.



```
                                          documents  scores sentiments
0  (U.S., intelligence, agencies, concluded, in, ...     2.0   positive
1  (And, the, bipartisan, Senate, Intelligence, C...     2.0   positive
```

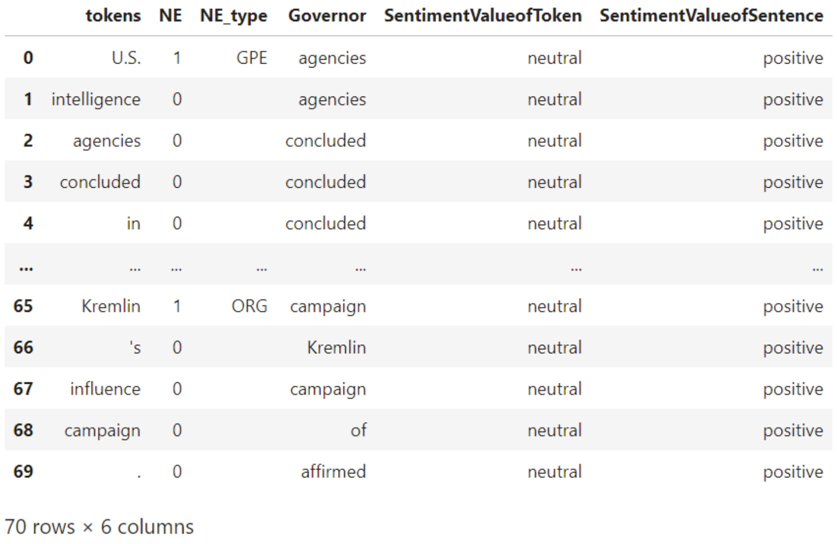
As for the tokens, "helping" and "approved" were found to have positive sentiments while all others were labelled as neutral.



	documents	scores	sentiments
0	U.S.	0.0	neutral
1	intelligence	0.0	neutral
2	agencies	0.0	neutral
3	concluded	0.0	neutral
4	in	0.0	neutral
...	...	...	...
65	Kremlin	0.0	neutral
66	's	0.0	neutral
67	influence	0.0	neutral
68	campaign	0.0	neutral
69	.	0.0	neutral

70 rows × 3 columns

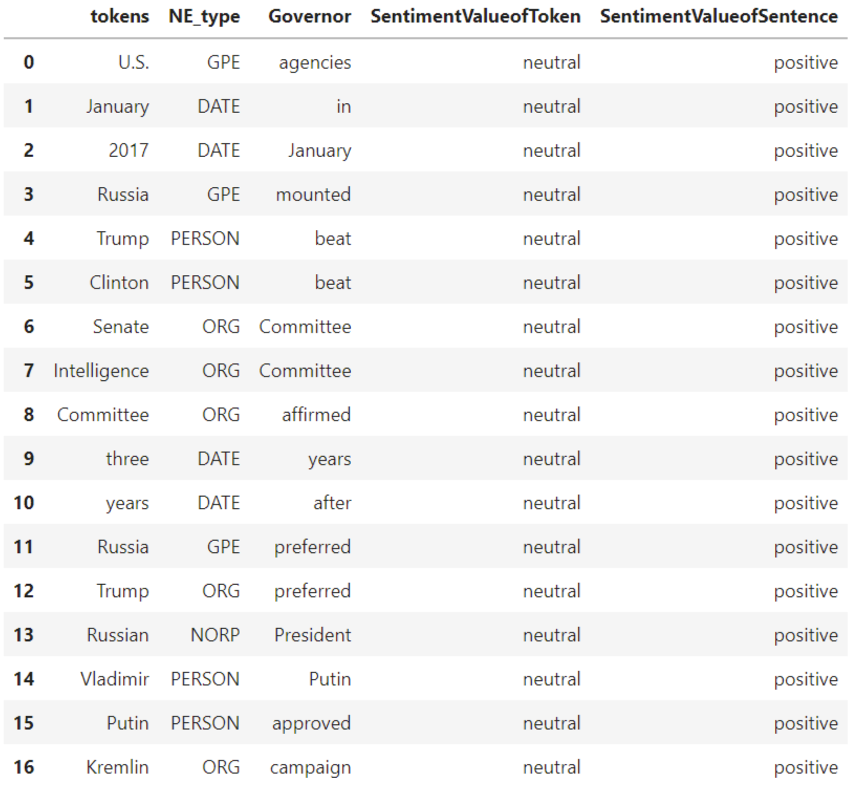
Table 1 for S1 has feature columns NE, NEtype, Governor, SentimentValueofToken, SentimentValueofSentence for all tokens:



	tokens	NE	NE_type	Governor	SentimentValueofToken	SentimentValueofSentence
0	U.S.	1	GPE	agencies	neutral	positive
1	intelligence	0		agencies	neutral	positive
2	agencies	0		concluded	neutral	positive
3	concluded	0		concluded	neutral	positive
4	in	0		concluded	neutral	positive
...	...	...	...	...	...	...
65	Kremlin	1	ORG	campaign	neutral	positive
66	's	0		Kremlin	neutral	positive
67	influence	0		campaign	neutral	positive
68	campaign	0		of	neutral	positive
69	.	0		affirmed	neutral	positive

70 rows × 6 columns

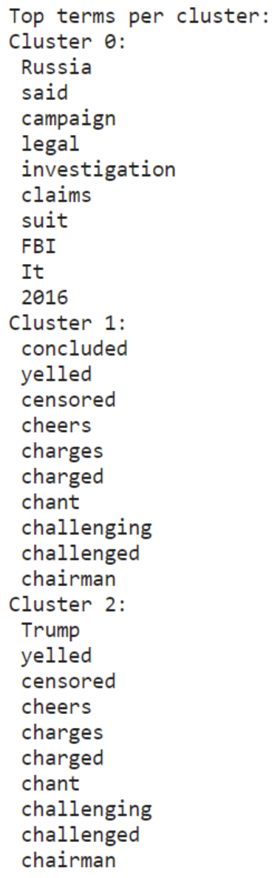
Table 2 for S1 has feature columns NEtype, Governor, SentimentValueofToken, SentimentValueofSentence only for Named Entity tokens.



	tokens	NE_type	Governor	SentimentValueofToken	SentimentValueofSentence
0	U.S.	GPE	agencies	neutral	positive
1	January	DATE	in	neutral	positive
2	2017	DATE	January	neutral	positive
3	Russia	GPE	mounted	neutral	positive
4	Trump	PERSON	beat	neutral	positive
5	Clinton	PERSON	beat	neutral	positive
6	Senate	ORG	Committee	neutral	positive
7	Intelligence	ORG	Committee	neutral	positive
8	Committee	ORG	affirmed	neutral	positive
9	three	DATE	years	neutral	positive
10	years	DATE	after	neutral	positive
11	Russia	GPE	preferred	neutral	positive
12	Trump	ORG	preferred	neutral	positive
13	Russian	NORP	President	neutral	positive
14	Vladimir	PERSON	Putin	neutral	positive
15	Putin	PERSON	approved	neutral	positive
16	Kremlin	ORG	campaign	neutral	positive

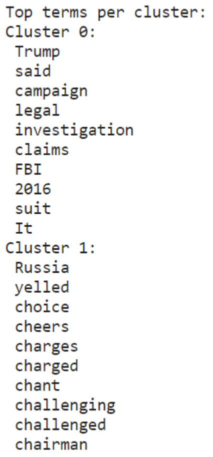
(d) 3-means clusters for the entire text

Lastly 3-means and 2-means clustering was performed on the full AP Text, and the Elbow Curve was used to find a good value k for k-means clusters.

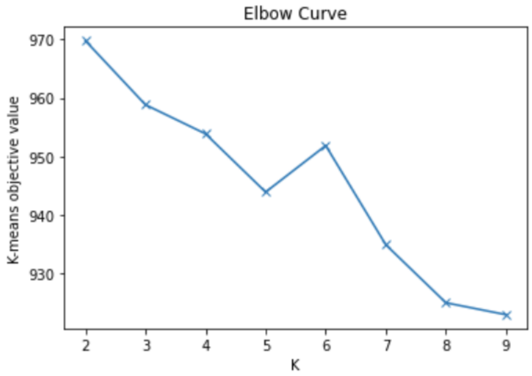


```
Top terms per cluster:
Cluster 0:
 Russia
 said
 campaign
 legal
 investigation
 claims
 suit
 FBI
 It
 2016
Cluster 1:
 concluded
 yelled
 censored
 cheers
 charges
 charged
 chant
 challenging
 challenged
 chairman
Cluster 2:
 Trump
 yelled
 censored
 cheers
 charges
 charged
 chant
 challenging
 challenged
 chairman
```

(e) 2-means clusters for the entire text



```
Top terms per cluster:
Cluster 0:
 Trump
 said
 campaign
 legal
 investigation
 claims
 FBI
 2016
 suit
 It
Cluster 1:
 Russia
 yelled
 choice
 cheers
 charges
 charged
 chant
 challenging
 challenged
 chairman
```



The top 10 words in each cluster were printed to screen. At the word-level, this shows that the tokens in the same cluster belong to the same semantic domain. By analysing them we can create labels for each cluster. Based on the curve above, we can observe from the point of inflection (the elbow) that the optimal value of k for our model is 5.