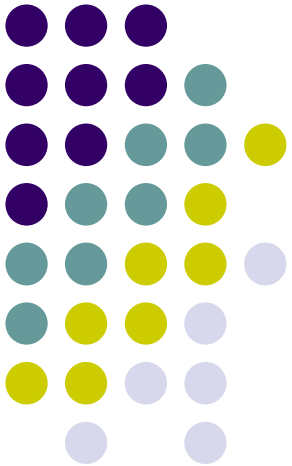
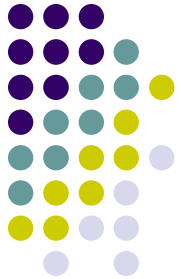
Big Data Analytics

Sixth Assignment



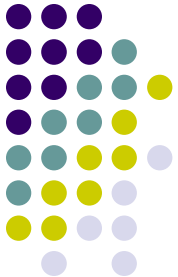
Deadline: 22 February 2021



First part

1) Use the corpus of the US Presidential Inaugural speeches after 1900
2) Remove stopwords, punctuation, numbers, symbols, separators and make everything lower case
3) Keep only features with a number of characters larger than 3
4) Employ the following keywords/seeded words: Government ("laws", "law", "executive"), Congress ("congress", "party"), Peace ("peace", "freedom"), Constitution ("constitution", "rights"), ForeignAffairs ("foreign", "war")

Deadline: 22 February 2021



5) Run a semi-supervised topic analysis via keyATM employing as a covariate the party of the president while also adding 3 non-keywords topics in the analysis

n.b. the original “party” variable is a factor variable with 6 levels. To keep just two levels (i.e., Republican and Democratic) you can do the following. Suppose that the party variable is included in the dataframe `vars_selected` (i.e., `vars_selected$Party`). Then you can write:

```
levels(vars_selected$Party) <- list(Democratic="Dem", Republican="Rep")
```

```
levels(vars_selected$Party ) # only the Democratic and the Republican levels should be displayed
```

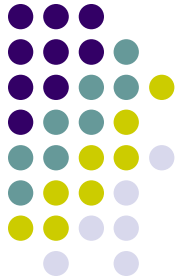
Deadline: 22 February 2021

Comment the results

Do you find any significant difference between Republican and Democratic presidents in terms of the salience given to the different topics?

Do you find any difference between Republican and Democratic presidents in terms of the words employed for the different topics?

Deadline: 22 February 2021



Second part

1. Make any query you like on Twitter
2. Run Newsmap on it by identifying the seed words you think are relevant given your query. Comment the results
3. Then apply a dictionary analysis on the tweets. Briefly comment your results