GSSOC 2020
20 Newsgroups dataset

What is Topic Modelling?:

Topic Modelling is a popular unsupervised NLP technique to identify the number of topics in a corpus (huge dataset of text elements). So,**topic modelling** is an unsupervised kind of machine learning method that scans various kinds of documents, detects word and phrase patterns within them and automatically creates word groups and similar expressions that characterise the set of documents.

In short, it is a method to automatically detect topics from texts.

There are various algorithms developed for the purpose of Topic Modelling:

- Latent Semantic Analysis (LSA)
- Latent Dirichlet Allocation (LDA)
- Non-negative Matrix Factorization (NFM)...

Our Dataset: 20 Newsgroups

It is a collection of newsgroup documents based on 20 different topics. It is a popularly used dataset for NLP purposes like topic modelling and text classification. The dataset can be found in the **scikit-learn** library of python. It is a collection of roughly 20,000 newsgroup documents split into 2 parts: the *training dataset* and the *test dataset*.

The 20 newsgroups are:

- alt.atheism
- comp.graphics
- comp.os.ms-windows.misc
- comp.sys.ibm.pc.hardware
- comp.sys.mac.hardware
- comp.windows.x rec.autos
- misc.forsale talk.politics.misc
- rec.motorcycles
- rec.sport.baseball
- rec.sport.hockey sci.crypt

- sci.electronics
- sci.med
- sci.space
- soc.religion.christian
- talk.politics.guns
- talk.politics.mideast talk.religion.misc

The dataset can be downloaded from **Kaggle** also: [20 Newsgroups]

Task:

- To find for what value of k the topic modelling works well on the dataset.
- To compare the LSA and LDA model on the dataset.

Implementation of models:

LSA model:

What is LSA?

- LSA is Latent Semantic Analysis.
- We can easily distinguish between these words because we are able to understand the context behind these words. However, a machine would not be able to capture this concept as it cannot understand the context in which the words have been used.
- This is where Latent Semantic Analysis (LSA) comes into play as it attempts to leverage the context around the words to capture the hidden concepts, also known as topics.

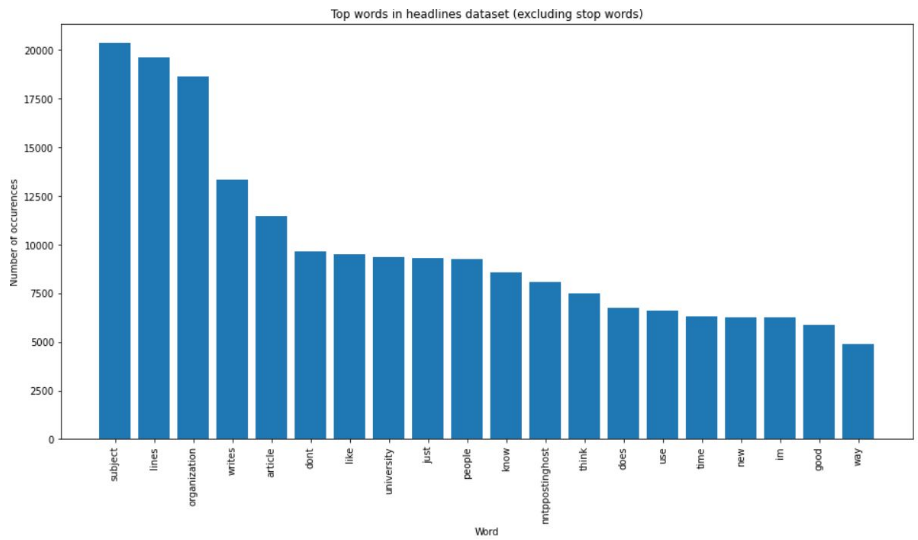
In this model, the task performed are:

1) Preprocess the dataset, i.e.

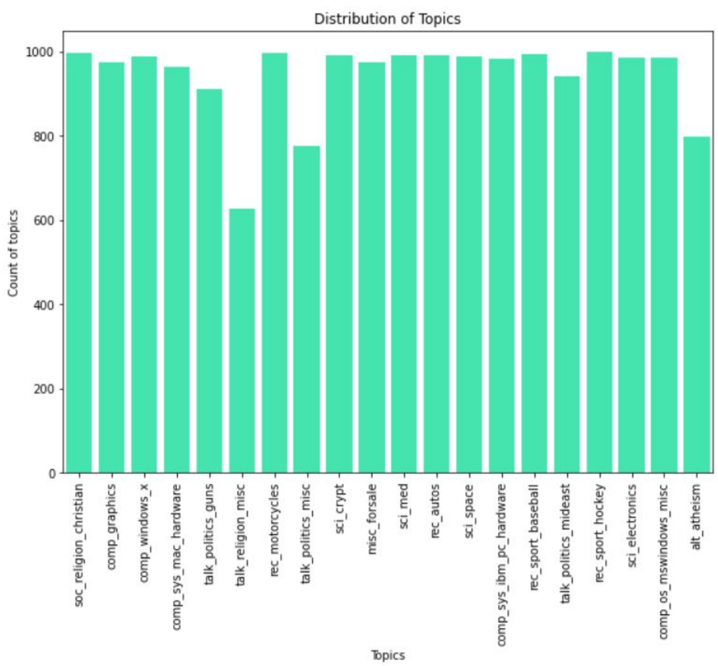
1.To remove the null/empty words

2.to remove the stop words(words which are most commonly used and are not considered to be a topic).

2) Prepare a graph displaying the occurrences of the most common words across all the documents.



3) Display the number of topics available over individual newsgroups

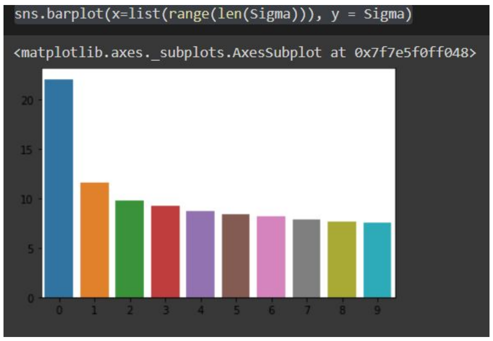


4) Implement the LSA model for different values of k.
    1. Generate a document term mXn matrix having TF-IDF score with k value as an input parameter.
    2. Reduce the dimensions of the matrix to kXk using singular value decomposition(SVD).
    3. Performing SVD will give us vectors(of length k) for each document and term in our dataset.These vectors will be useful for finding the common words using the cosine similarity method.

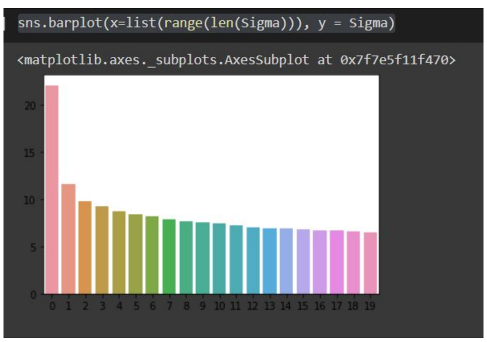
5) Compare the results for different values of k, based on different kinds of graphs and outcomes

1.The topics commonly used in each of the newsgroups are displayed in the form of bar chart:(Here the graph consists of x-axis with values of 0 to k-1 topics and y-axis with rows of dataset which can be included in the ith topic, i.e. 0<i<k-1)

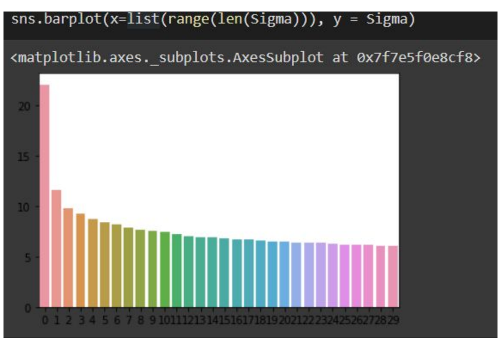
For k = 10:



For k = 20:



For k = 30:



2.The topic names which are commonly used across all newsgroups for different k values are:

For k =10:

Topic 0:

```
like know people think good time thanks
Topic 1:
thanks windows card drive mail file advance
Topic 2:
game team year games season players good
Topic 3:
drive scsi disk hard card drives problem
Topic 4:
windows file window files program using problem
Topic 5:
government chip mail space information encryption data
Topic 6:
like bike know chip sounds looks look
Topic 7:
card sale video offer monitor price jesus
Topic 8:
know card chip video government people clipper
Topic 9:
good know time bike jesus problem work
```

For k = 20:

Topic 0:
like know people think good time thanks
Topic 1:
thanks windows card drive mail file advance
Topic 2:
game team year games season players good
Topic 3:
drive scsi disk hard card drives problem
Topic 4:
windows file window files program using problem
Topic 5:
government chip mail space information encryption data
Topic 6:
like bike know chip sounds looks look
Topic 7:
card sale video offer monitor price jesus
Topic 8:
know card chip video government people clipper
Topic 9:
good know time bike jesus problem work
Topic 10:
think chip good thanks clipper need encryption
Topic 11:
thanks right problem good bike time window
Topic 12:
good people windows know file sale files
Topic 13:
space think know nasa problem year israel
Topic 14:

space good card people time nasa thanks
Topic 15:
people problem window time game want bike
Topic 16:
time bike right windows file need really
Topic 17:
time problem file think israel long mail
Topic 18:
file need card files problem right good
Topic 19:
problem file thanks used space chip sale

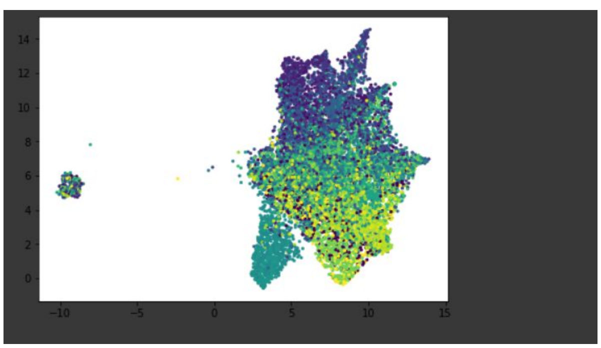
For k = 30:

```
Topic 0:
like know people think good time thanks
Topic 1:
thanks windows card drive mail file advance
Topic 2:
game team year games season players good
Topic 3:
drive scsi disk hard card drives problem
Topic 4:
windows file window files program using problem
Topic 5:
government chip mail space information encryption data
Topic 6:
like bike know chip sounds looks look
Topic 7:
card sale video offer monitor price jesus
Topic 8:
know card chip video government people clipper
Topic 9:
good know time bike jesus problem work
Topic 10:
think chip good thanks clipper need encryption
Topic 11:
thanks right problem good bike time window
Topic 12:
good people windows know file sale files
Topic 13:
space think know nasa problem year israel
Topic 14:
space good card people time nasa thanks
Topic 15:
people problem window time game want bike
Topic 16:
time bike right windows file need really
Topic 17:
time problem file think israel long mail
Topic 18:
file need card files problem right good
```

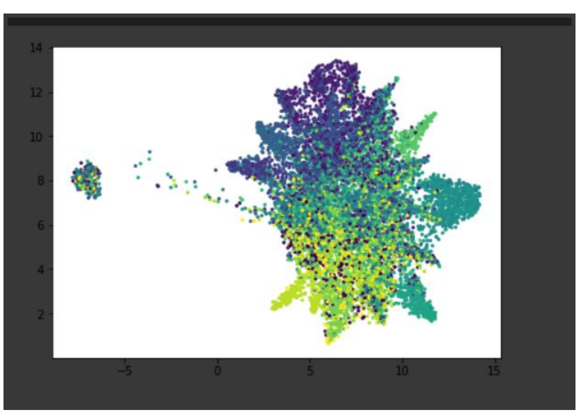
```
Topic 19:
problem file thanks used space chip sale
Topic 20:
problem mail windows need address send really
Topic 21:
need space game israel want windows really
Topic 22:
year said need bike armenian armenians window
Topic 23:
year need make time offer israel monitor
Topic 24:
right good space government jesus window problem
Topic 25:
sure make really window said thanks government
Topic 26:
team bike window list jesus players file
Topic 27:
game bike looking window year israel mail
Topic 28:
sure work make program jesus works email
Topic 29:
email article need window scsi post believe
```

3.The topics clusters across the newsgroups are:

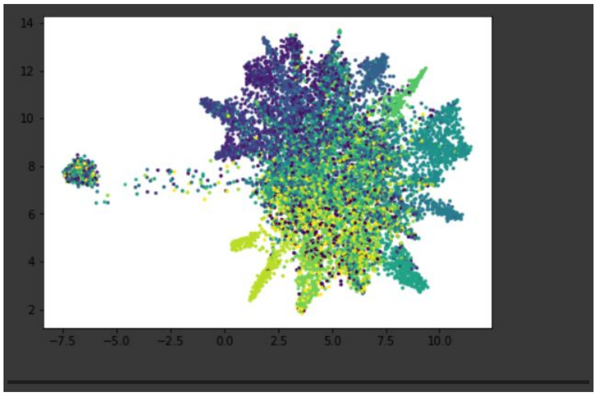
For k = 10:



For k = 20:



For k = 30:



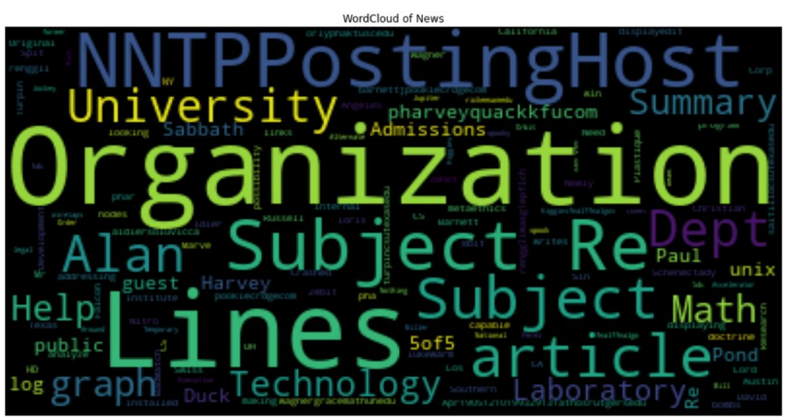
Results:

We have been able to implement LSA and LDA algorithms on the dataset.

LSA Implementation

LSA implementation gave us promising results with the value of k between **25 to 30**.

So, for the value of k=30 we have implemented the word cloud(cloud which provides the most commonly used words across all the newsgroups)



LDA Implementation

LDA implementation gave us good results:

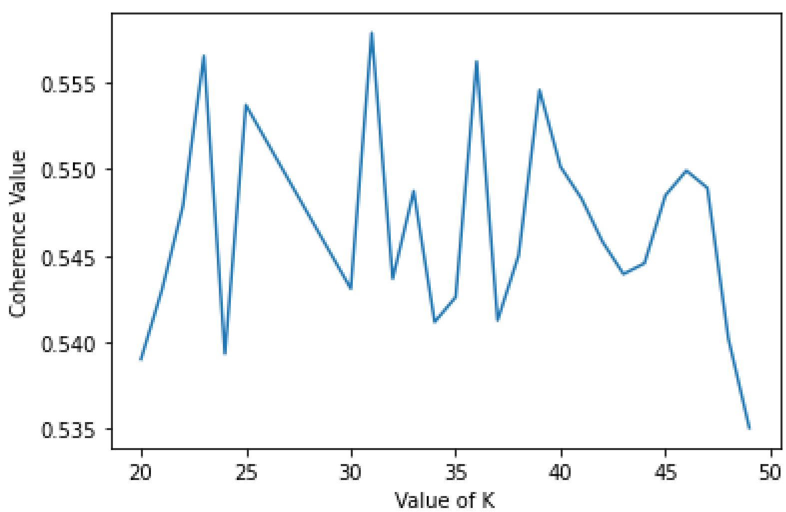
- Perplexity = -8.602970553126184
- Coherence Score = 0.5578795935507215 (k=31, Mallet LDA)

Sample Topics:

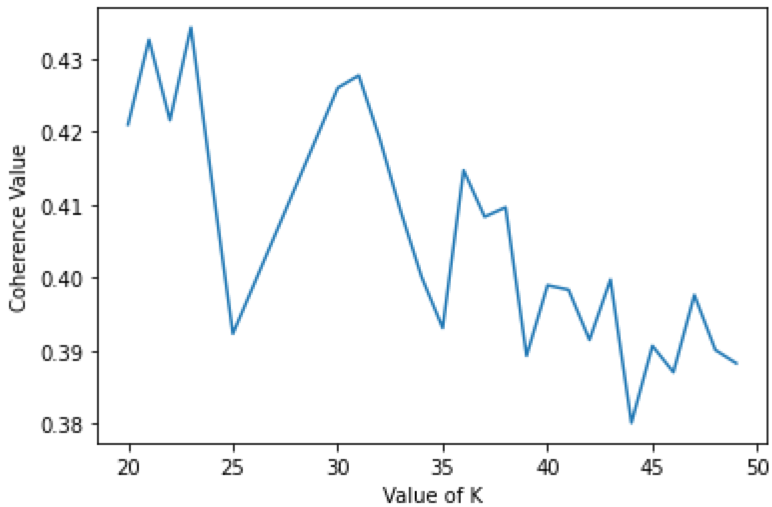
```
[(0,

  '0.336*"sci_electronics" + 0.334*"sci_med" + 0.329*"comp_graphics" + 0.000*"soc_religion_christian" + 0.000*"comp_windows_x" + 0.000*"talk_politics_guns" + 0.000*"rec_motorcycles" + 0.000*"talk_politics_misc" + 0.000*"sci_crypt" + 0.000*"misc_forsale"'),

 (1,

  '0.998*"rec_motorcycles" + 0.000*"misc_forsale" + 0.000*"comp_graphics" + 0.000*"sci_med" + 0.000*"comp_os_mswindows_misc" + 0.000*"rec_autos" + 0.000*"comp_windows_x" + 0.000*"talk_politics_mideast" + 0.000*"talk_politics_guns" + 0.000*"sci_electronics"'),

 (2,

  '0.514*"comp_windows_x" + 0.485*"talk_politics_guns" + 0.000*"rec_autos" + 0.000*"alt_atheism" + 0.000*"comp_sys_mac_hardware" + 0.000*"talk_politics_misc" + 0.000*"comp_os_mswindows_misc" + 0.000*"talk_religion_misc" + 0.000*"talk_politics_mideast" + 0.000*"sci_med"'),

 (3,

  '0.548*"sci_crypt" + 0.451*"alt_atheism" + 0.000*"rec_autos" + 0.000*"comp_sys_ibm_pc_hardware" + 0.000*"rec_sport_hockey" + 0.000*"comp_sys_mac_hardware" + 0.000*"talk_politics_misc" + 0.000*"rec_motorcycles" + 0.000*"comp_graphics" + 0.000*"sci_electronics"'),

 (4,

  '0.998*"sci_space" + 0.000*"soc_religion_christian" + 0.000*"comp_sys_mac_hardware" + 0.000*"talk_politics_mideast" + 0.000*"talk_religion_misc" + 0.000*"comp_sys_ibm_pc_hardware" + 0.000*"sci_med" + 0.000*"sci_electronics" + 0.000*"rec_sport_hockey" + 0.000*"rec_motorcycles"'),

 (5,

  '0.998*"talk_politics_mideast" + 0.000*"sci_med" + 0.000*"talk_religion_misc" + 0.000*"sci_space" + 0.000*"comp_sys_ibm_pc_hardware" + 0.000*"rec_autos" + 0.000*"sci_electronics" + 0.000*"misc_forsale" + 0.000*"talk_politics_guns" + 0.000*"rec_motorcycles"'),

 (6,

  '0.344*"rec_sport_hockey" + 0.329*"comp_sys_mac_hardware" + 0.327*"comp_os_mswindows_misc" + 0.000*"comp_graphics" + 0.000*"rec_motorcycles" + 0.000*"sci_space" + 0.000*"alt_atheism" + 0.000*"comp_sys_ibm_pc_hardware" + 0.000*"sci_crypt" + 0.000*"talk_religion_misc"'),
```

```
 (7,
          '0.359*"soc_religion_christian" + 0.357*"rec_autos" +
0.283*"talk_politics_misc" + 0.000*"rec_sport_hockey" + 0.000*"alt_atheism" +
0.000*"comp_windows_x" + 0.000*"comp_sys_ibm_pc_hardware" +
0.000*"rec_motorcycles" + 0.000*"comp_os_mswindows_misc" +
0.000*"talk_politics_mideast"'),

 (8,
          '0.620*"rec_sport_baseball" + 0.379*"talk_religion_misc" +
0.000*"talk_politics_misc" + 0.000*"rec_autos" + 0.000*"sci_crypt" +
0.000*"comp_windows_x" + 0.000*"comp_sys_ibm_pc_hardware" +
0.000*"rec_sport_hockey" + 0.000*"comp_os_mswindows_misc" +
0.000*"sci_space"'),

 (9,
          '0.503*"misc_forsale" + 0.496*"comp_sys_ibm_pc_hardware" +
0.000*"comp_sys_mac_hardware" + 0.000*"rec_sport_hockey" + 0.000*"sci_space" +
0.000*"comp_windows_x" + 0.000*"talk_religion_misc" + 0.000*"rec_motorcycles" +
0.000*"sci_crypt" + 0.000*"talk_politics_guns"')]
```

Coherence Graph for Mallet LDA:



Coherence Graph for LDA:



Inferences:

- LSA offers lower accuracy as compared to LDA.
- LSA is more difficult to implement than LDA.
- LDA provides far more better results as compared to that of LDA.

REFERENCES

- **Topic Modeling:**
  - http://derekgreene.com/papers/greene14topics.pdf
- **LSA model:**
  - https://www.datacamp.com/community/tutorials/discovering-hidden-topics-python
  - https://www.analyticssteps.com/blogs/introduction-latent-semantic-analysis-lsa-and-latent-dirichlet-allocation-lda
  - https://towardsdatascience.com/latent-semantic-analysis-sentiment-classification-with-python-5f657346f6a3
  - https://www.analyticsvidhya.com/blog/2018/10/stepwise-guide-topic-modeling-latent-semantic-analysis/
  - https://www.kaggle.com/rcushen/topic-modelling-with-lsa-and-lda
- **LDA model:**

- https://www.tutorialspoint.com/gensim/gensim_creating_lda_mallet_model.htm
- https://programminghistorian.org/en/lessons/topic-modeling-and-mallet
- https://medium.com/analytics-vidhya/how-to-perform-topic-modeling-using-mallet-abc43916560f