Part 1: Reflection

What challenges did you face for implementing the LDA?

Finding a method to achieve topic word distribution and document topic distribution was one of the obstacles we encountered while implementing the LDA. This was a challenging as we had a large corpus to deal with. It took about 30 min to execute the model which was time consuming to go back and edit the code.

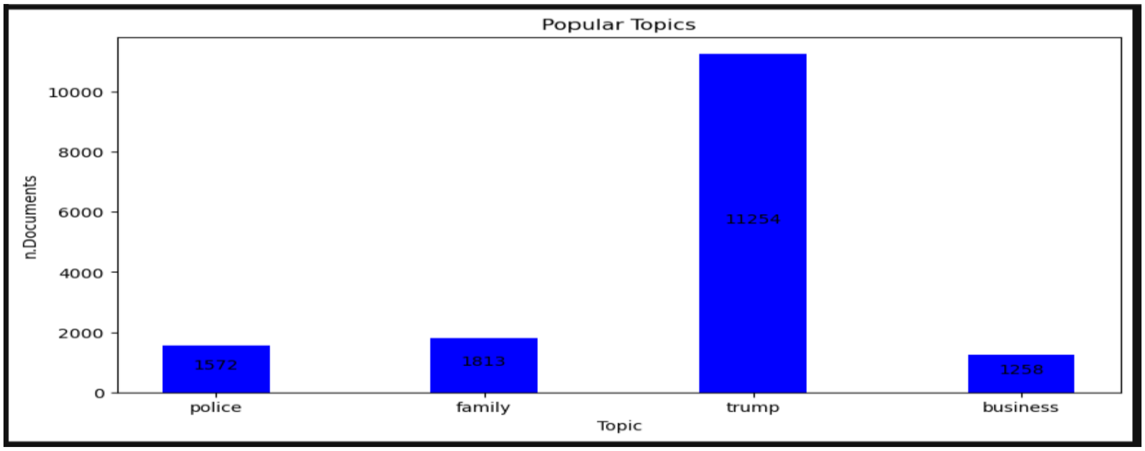
What are the main differences between LDA and other text clustering techniques, i.e., K-means?

LDA and other clustering methods like K-mean are also examples of unsupervised learning models. The main difference between LDA and k-cluster is the assignment of documents to a variety of topics in LDA. Each document will be assigned to one topic by K-cluster and one or more topics by LDA. For instance, if N is set to 3, k-mean will group the documents into one of the three topics. On the other hand, for LDA it will do the following: document A belongs to 30% of topic 1, 50% to topic 2, and 20% to topic 3. LDA will give a more realistic result compared to k-mean cluster for topic assignments.

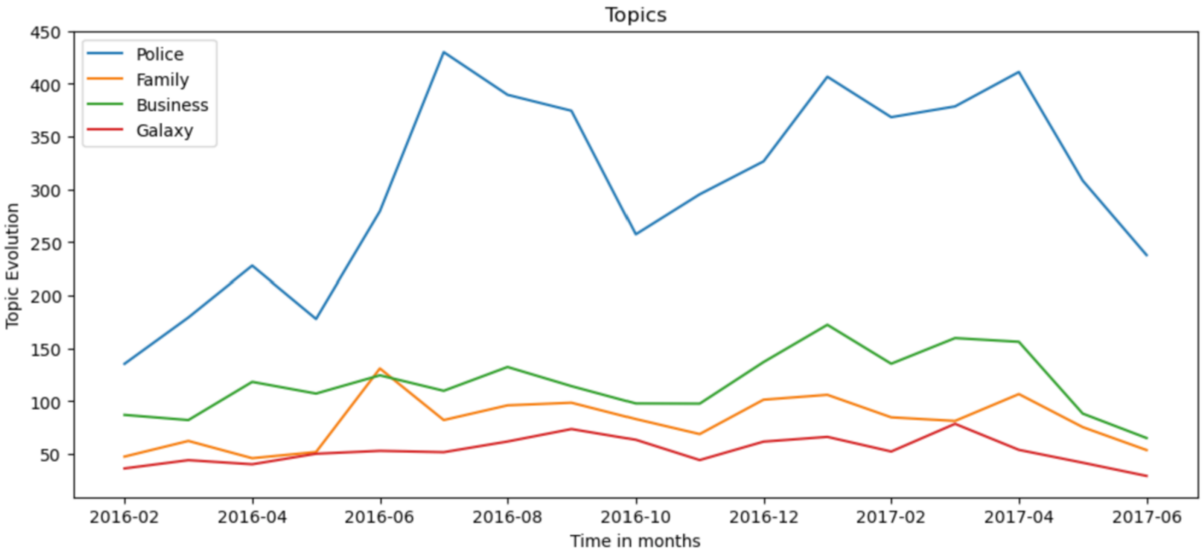
Part 2: Visualization Questions

What is the most popular topic? Draw a bar chart to visualize the topic distribution in documents. A sample chart is given below. The horizonal axes shows the topics and the vertical axes shows the documents containing that topic.

The four most popular subjects across all corpus documents are displayed in the bar graph below. Trump is the most talked-about topic. There are around 11 254 documents that contains the topic of Trump. This makes sense given that Trump was running for president in the years between 2016 and 2017 and major new article revolved around this topic.

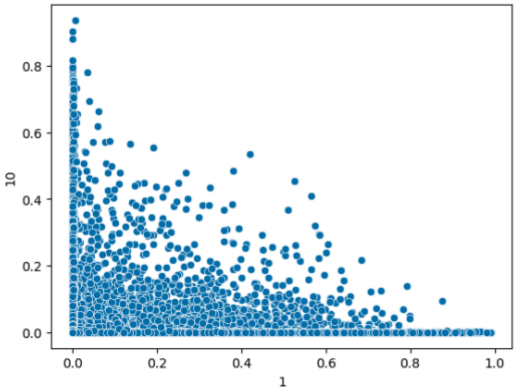


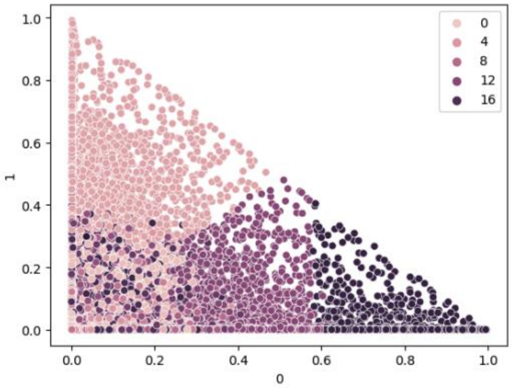
Select four popular topics in the corpus and draw a chart to show the popularity (trend) of the topics throughout the months of 2016 and 2017. Your chart should look like the sample chart presented in the lecture of week 9.



Use Silhouette Scores or Elbow Curves to justify the number of topics in your project. Explain what would be best number of topics in this dataset and why.

The number of topics that is used for LDA model is 20. The output for Silhouette shows 0.25 which is doable but naive. This indicates that the number of topics used is not sufficient for the number of documents in the corpus. The closer it is to 1 the better the model works and the closer to 0 the worse the model works.


10
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
1


1
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
0
0
4
8
12
16