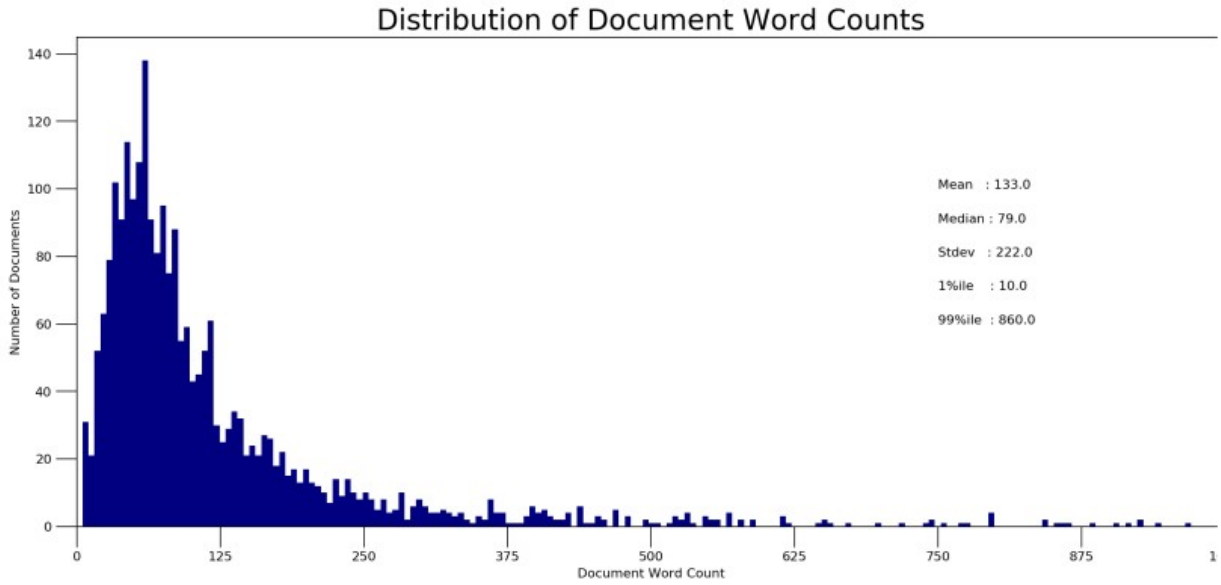
Big Data Analytics

Practical No. 6

AIM: Understand how visualization actually work.

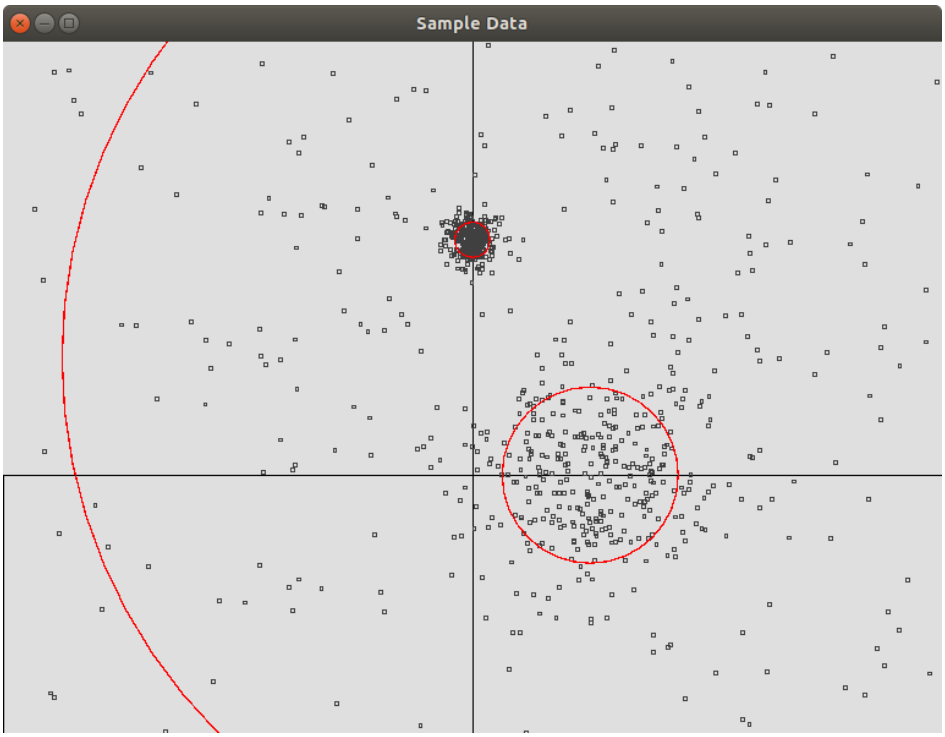
In LDA, we experimented with different visualizations likewise in Gephi.

We start with:

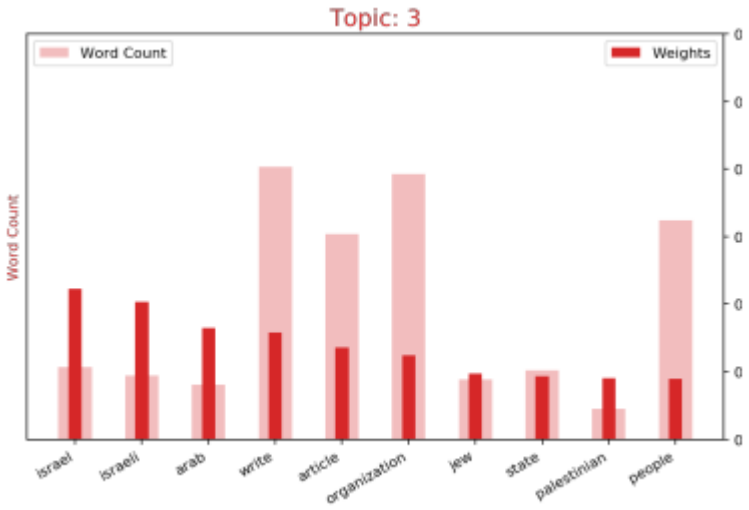


In Python: It shows a very basic histogram of how the document is distributed.

In Python: This shows a word cloud for the topic Middle East news. The larger a word the more dominant it is in that topic.



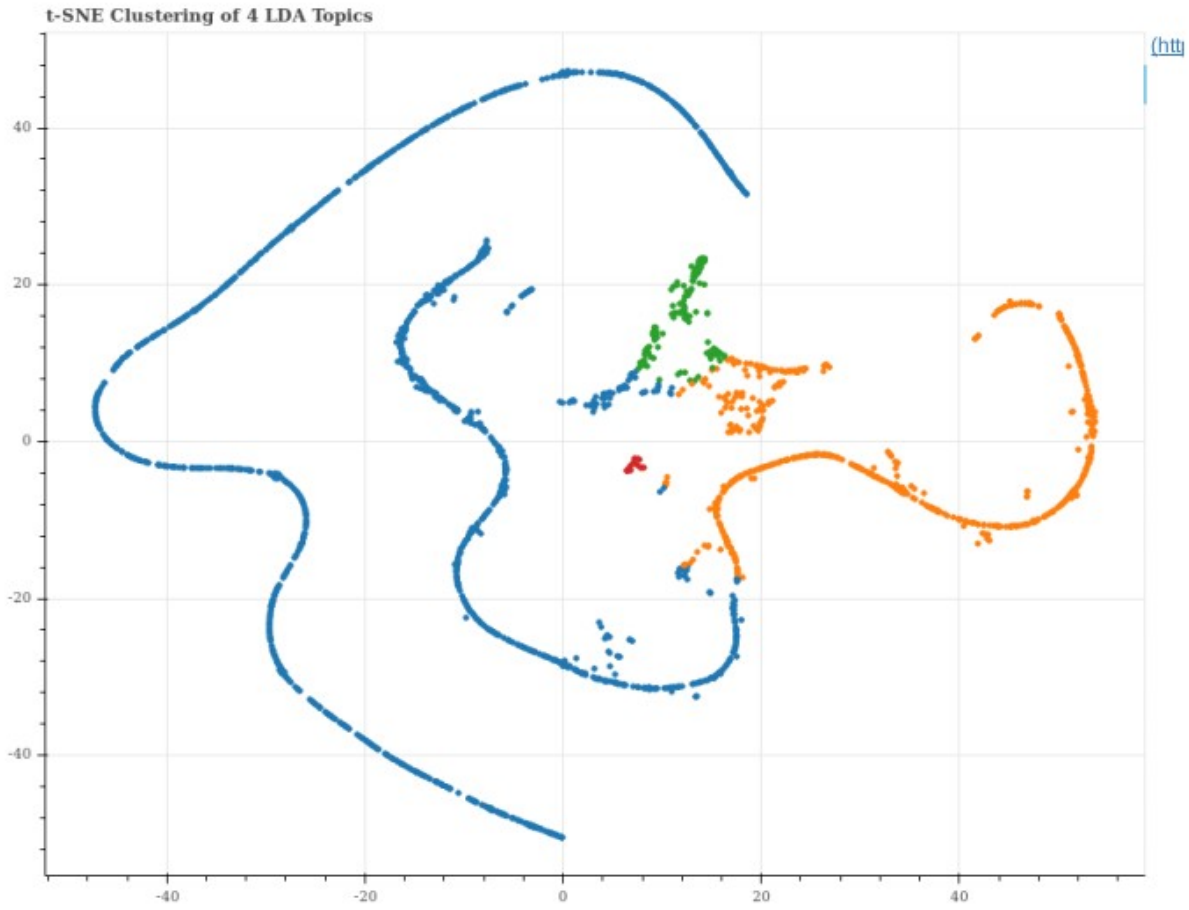
In Mahout: A sample clustering using the command mahout org.apache.mahout.clustering.display.DisplayClustering



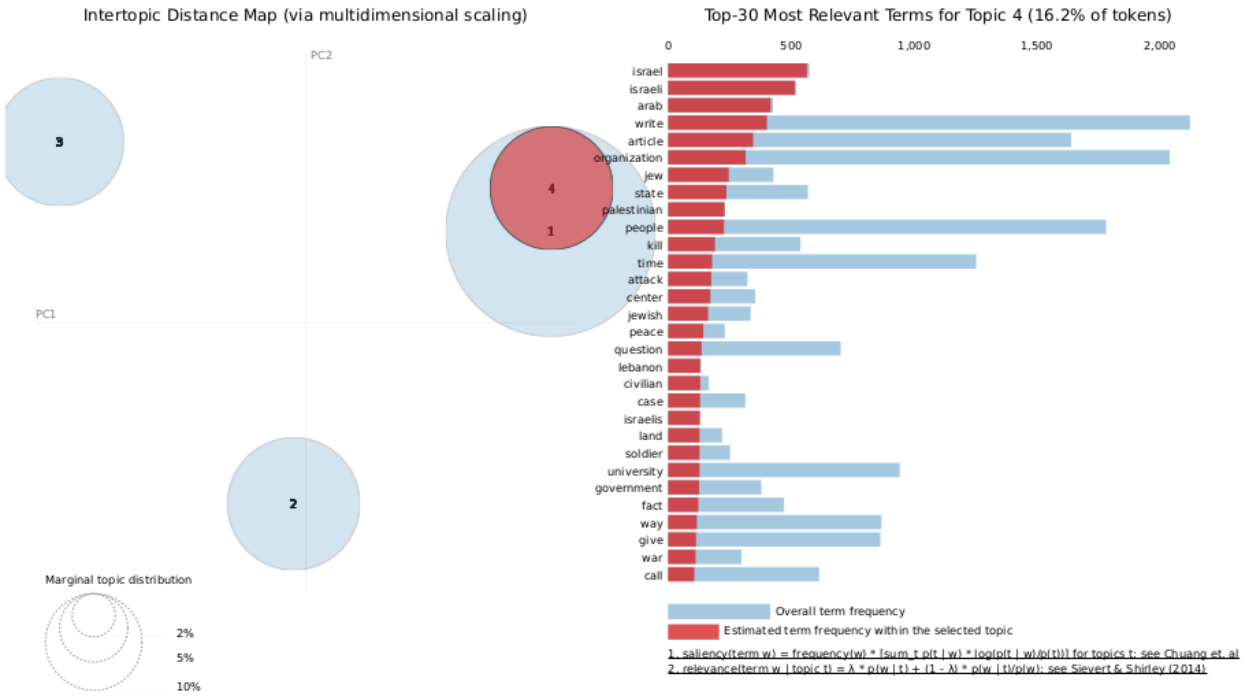
In Python: A word count vs weightage of topic keywords hybrid bar chart.

In Python: Sentence topic clustering. Different colors for different topics.

Big Data Analytics



In Python: T-distributed Stochastic Neighbor Embedding is a machine learning algorithm for visualization. It is a nonlinear dimensionality reduction technique well-suited for embedding high-dimensional data for visualization in a low-dimensional space of two or three dimensions. Specifically, it models each high-dimensional object by a two- or three-dimensional point in such a way that similar objects are modeled by nearby points and dissimilar objects are modeled by distant points with high probability.



In Python: Topics 4 (“Religion”) and 5 (“Middle east”) are very close together in the intertopic distance map, which suggests overlap between words like Jews and Organization.

Conclusion:

We understood the real power and easiness of Python here where we can simply import graphic libraries and then use simple parametric methods to create graphs which are not only visually rich but also interactive.