

Topic Modeling

1) Data Reading:

- Reading data using read from pandas library
- Return it into data frame df

2) Data Preprocessing:

- Use sample size of data, sample size = 200
- Import stop words, Punctuation and remove them from sample
- Loop on sample size , get each raw, then tokenize data
- Lemmetize tokens using WordNetLemmatizer
- Append them to Corpus dataframe

3) Features Extraction:

- Create TF-IDF vectorizer object
- Fit training data using fit_transform
- Convert it to dataframe tfidfmatrix

4) Kmeans:

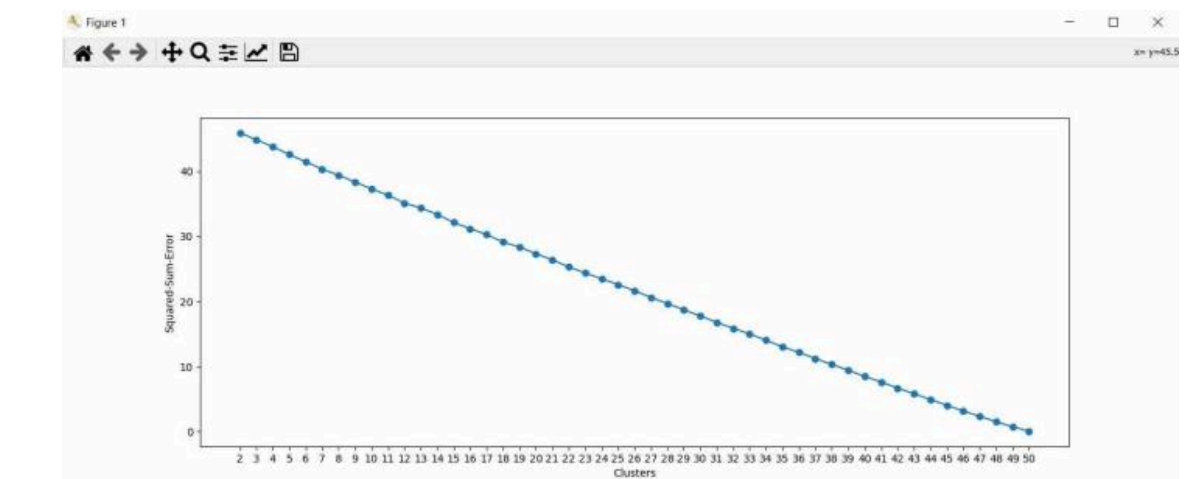
- Do clustering for optimal K
- Append them in Squared_Sum_Error

5) MiniBatchKMeans:

- Fit data using fit_transform

6) Visualising data:

- Kmeans:



- MiniBatchKMeans:

