Topic Modeling

TA: Asmaa Kassem

Team Number: T085

level	Department	ID	Section	Name
3	CS	20201700954	10	هايدى حسن سعد عبدالله
3	CS	20201700944	10	نورهان محمد عبدالله محمد حماد
3	CS	20201700814	8	مريم علاء الدين فضل
3	CS	20201701153	10	هناء احمد حامد
3	CS	20201700107	2	اسامة محمد الغريب
3	CS	20201700530	5	عمرو العزونى محمد

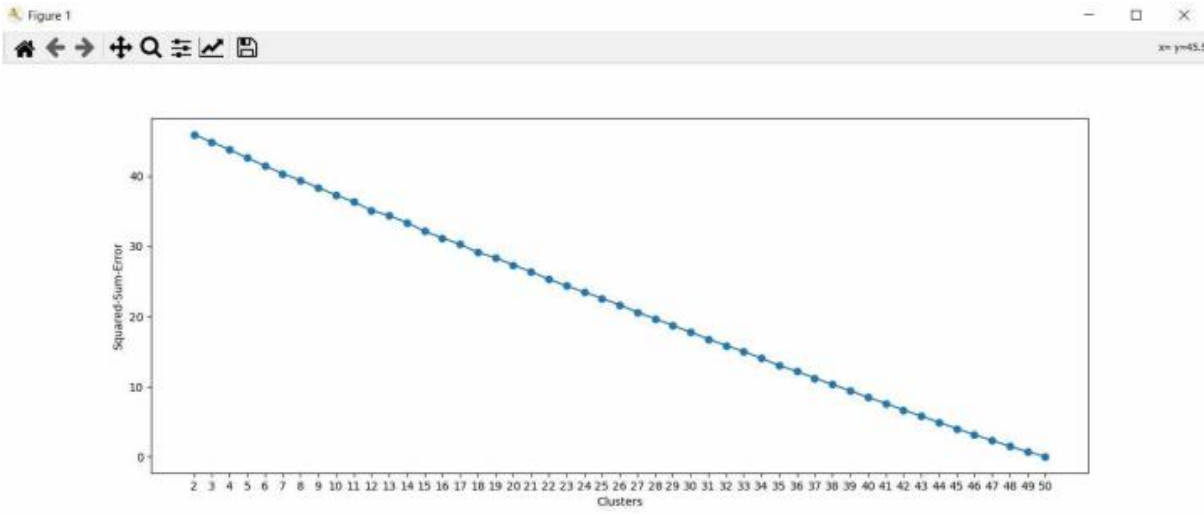
1) Data Reading:
    - Reading data using read from pandas library
    - Return it into data frame df

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:

-

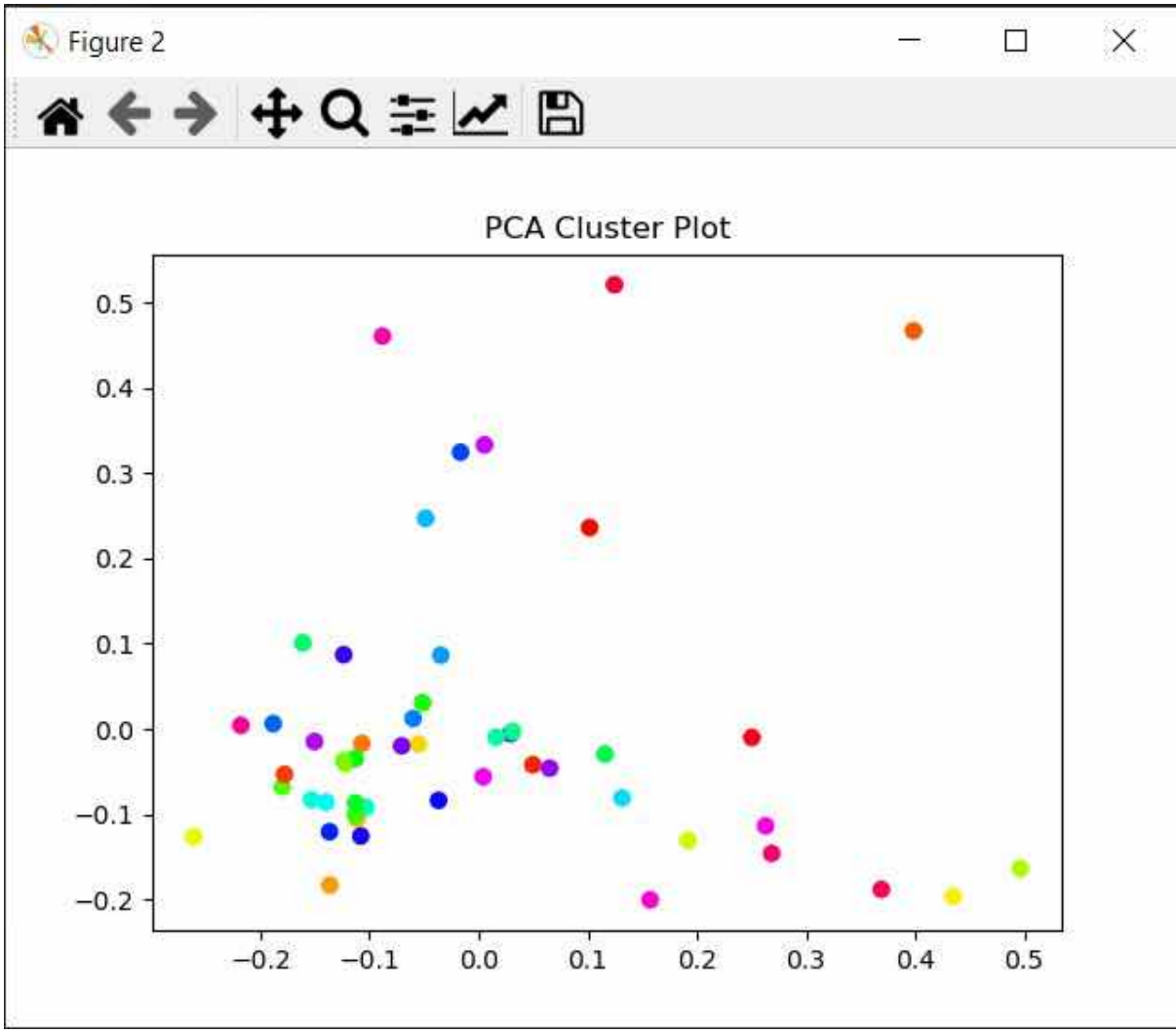


Figure 2
PCA Cluster Plot
0.5
0.4
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.2
−0.1
0.0
0.1
0.2
0.3
0.4
0.5