Lab Assignment 8 (Text Clustering)

This assignment includes implementation of K-means algorithm (an unsupervised learning algorithm) using sklearn library. The Objective of this assignment is to cluster the similar tweets based on similarity of words within the sentences. Use the health tweets dataset of UCI https://archive.ics.uci.edu/ml/datasets/Health+News+in+Twitter

1. Open the tweets file and extract the content line by line and save it in a list.

2. Extract data from the list of tweets. Hint: use regular expression for extracting tweet id, tweet date, news and link. Convert the extracted contents to dataframe.

3. Create a new column in the dataframe and named it as "news_tokens". In this column save the 'news' column contents after doing the pre-processing. [Pre-processing steps: Convert news column content to lower case. Remove the punctations and digits is any from the news columns. Then use from nltk.tokenize import word_tokenize to tokenize the sentences in news column].

4. Use from nltk.tag import pos_tag. Apply pos_tag function on "news_tokens".

5. Use from nltk.stem import WordNetLemmatizer. Apply this Lemmatizer to lemmatize the words.

6. Use TfidVectorizer (from sklearn.feature_extraction.text import TfidfVectorizer), to convert the words into vectors.

7. Now use kmeans algorithm to cluster these vectors of tweets.

8. For Fun: use elbow method to check the appropriate number of clusters. Also you can use silhouette_score() to compare cluster quality for different k values. (vary k from 2 to 15).