Assignment #5

Formatting:

If you're coding in a Jupyter Notebook, just submit your ipynb file with all the outputs. If you're using a py file, include all the outputs in a pdf along with your Python code.

Problem Statement:

Find hot topics of the world news from data available on sklearn:

from sklearn.datasets import fetch_20newsgroups

Find the news topics from this data set by applying Topic Modeling methods. This data already labeled for different topics but do not use those labels and approach this problem like an unsupervised problem.

Tasks:

1- Read fetch_20newsgroups from sklearn.datasets.
2- Perform necessary preprocessing steps for text data such as
    a. Tokenization
    b. Removing all punctuation and lowercase words.
    c. Removing the stop words.
    d. Stemming
3- Create bag of words.
4- Apply one type of the Topic Modeling methods (LDA or LSA) to find the news topics.
5- After finding the “Topics” use word clouds and coherence metrics to modify and get a meaningful set of topics.

Grading Criteria:

Total: 100

1- 0
2- 20
3- 20
4- 40
5- 20