Homework 1: CS-GY 9223

Exploring 20 NewsGroups

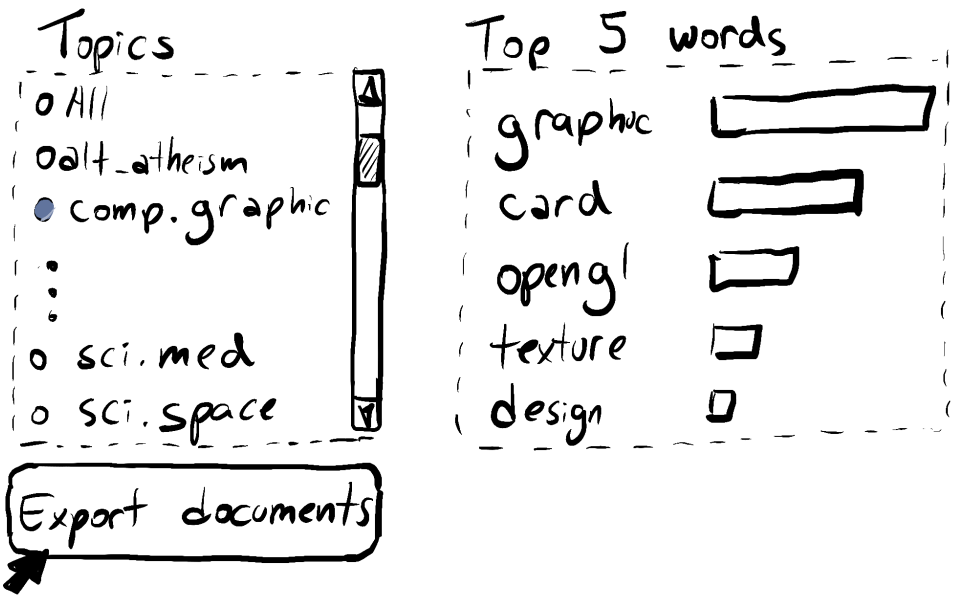
In this homework, you will write a D3 Visualization in Javascript and integrate it in Jupyter Notebook.

The goal of this exercise is to explore the **20 News Groups dataset**, a popular machine learning dataset that contains news articles grouped in 20 topics. Your visualization should receive the dataset and display a bar chart with the top most frequent words in the dataset. The user should be able to filter the data based on topic (for example, by clicking in checkboxes, selecting from a drop down menu, etc.). The user should also be able to export the selected documents from the selected topic back to Python (using a button).

In summary, your visualization should have the following capabilities:

- Display a bar chart with the top K words in the document collection
- Enable the user to filter the documents based on topic, and display a bar chart with the frequency of the top K words from that topic.
- Export the documents from the selected news topic back to python (as a list of strings).
- The visualization has to be integrated with python. The API should have two functions:
  - `plot_top_words(documents, K) # plot top K words using D3 and Javascript`
  - `get_exported_documents() # get the exported documents back to python`

Example of the resulting visualization:



Accessing the data

The data should be accessed from sklearn. In this section we show an example of code for accessing the documents and the document classes.

In [2]:
```
# Fetching the data
from sklearn.datasets import fetch_20newsgroups
import numpy as np
newsgroups = fetch_20newsgroups(subset='test')

# getting the topic ids
topic_idx = np.array(newsgroups.target, dtype=int)

# getting the unique topic names
topic_names = np.array(newsgroups.target_names)

# getting the list of documents
documents = list(newsgroups.data)

# getting the list of topics (in the same order as documents)
topics = list(topic_names[topic_idx])
```

These are the 20 topics in the dataset:

In [3]:
```
topic_names
```

Out[3]:
```
array(['alt.atheism', 'comp.graphics', 'comp.os.ms-windows.misc',
       'comp.sys.ibm.pc.hardware', 'comp.sys.mac.hardware',
       'comp.windows.x', 'misc.forsale', 'rec.autos', 'rec.motorcycle
s',
       'rec.sport.baseball', 'rec.sport.hockey', 'sci.crypt',
       'sci.electronics', 'sci.med', 'sci.space',
       'soc.religion.christian', 'talk.politics.guns',
       'talk.politics.mideast', 'talk.politics.misc',
       'talk.religion.misc'], dtype='<U24')
```

The documents and document topics are assigned to the variables *documents* and *topics*. We print some document examples below.

In [ ]:
```
for i in range(2):
    print("Topic: {}".format(topics[i]))
    print("-"*60)
    print("Document:")
    print(documents[i])
    print("="*60)
    print("")
```