Group: Alejandro Munoz, Rafael Alves Moreira, Kevin Raj, Robert Linhart

Coding Exercise: Write a code to tokenize the text and grab contents from a webpage where you can find information about SpaceX. Use stopwords and use the above strategy to calculate the frequency. The graph should display the first 10 high distribution words in the webpage while ignoring others. If the frequency of the word is less the 5 times ignore those words as well.

Questions:

1. Why we use stopwords? Why stopwords are not necessary for NLP frequency distribution.
   a. Stop words are words such as 'and', 'the', 'but'. These words are marked as unimportant words and are eliminated during pre-processing that allows the NLP algorithm to focus on the important words that add meaning. Stop words are not necessary for NLP frequency because they add clutter to the results, and the important words used for analysis could be missed.

2. Based on high frequency words what information you can extract from the graph?
   a. The Wikipedia article for SpaceX mentions the word – original the most with 175 occurrences. This is due to the citations at the bottom of the article archiving from the original article. SpaceX is ranked second with 144 occurrences which makes sense since the subject of the article is about SpaceX. Other interesting words were launch with 96 occurrences which makes sense for shuttles launching. Falcon was mentioned 92 times, and Falcon is the name of the rockets used for SpaceX launches.
3. Can you provide different visualization for frequency distribution? If yes, please perform. If no, why?
   a. Here is another visualization to show the top 10 words by occurrence in a tabular format:

Another visualization for NLP code without stop words

```
4]: freq.tabulate(10, cumulative=False)
```

original	SpaceX	launch	Falcon	Retrieved1	first	9	Space	May	March
175	144	96	92	87	80	75	72	69	63

Another visualization for NLP code with stop words

```
[15]: freq.tabulate(10, cumulative=False)
      the     to     of     on    and    in original      a  SpaceX    for
      533    276    258    245    218   195      175     159    144    109
```
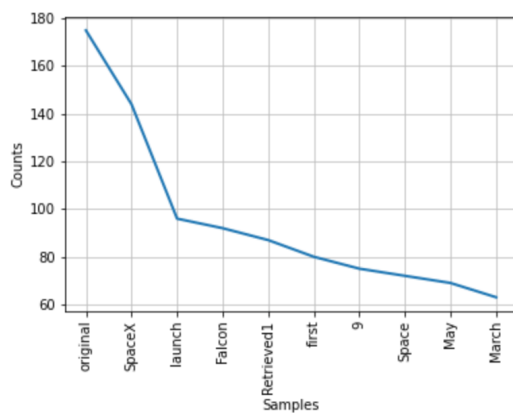
Frequency plot without stop words

Frequency plot without stop words

```
[23]: freq.plot(10, cumulative=False)
```



```
[23]: <matplotlib.axes._subplots.AxesSubplot at 0x7fdb75e88d60>
```
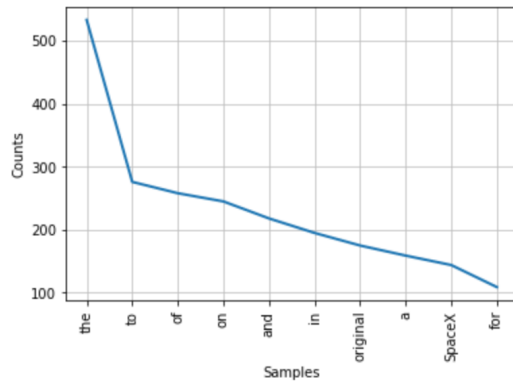
Frequency plot with stop words

Frequency plot with stop words

```
[14]: freq.plot(10, cumulative=False)
```



```
[14]: <matplotlib.axes._subplots.AxesSubplot at 0x7fdb77627df0>
```

Please upload you code along with outputs to Canvas. The submission should include code file, and word file for explanation. Make sure to comment your code properly.

Github Link: https://github.com/kevinraj127/CSCE5290/tree/main/ICE%231