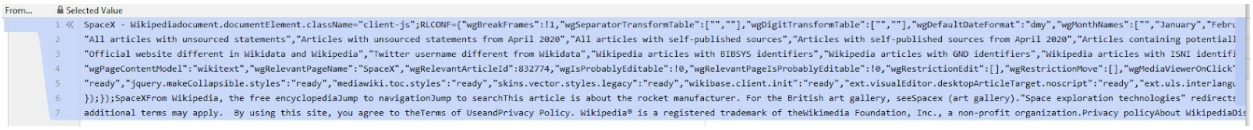
NLP

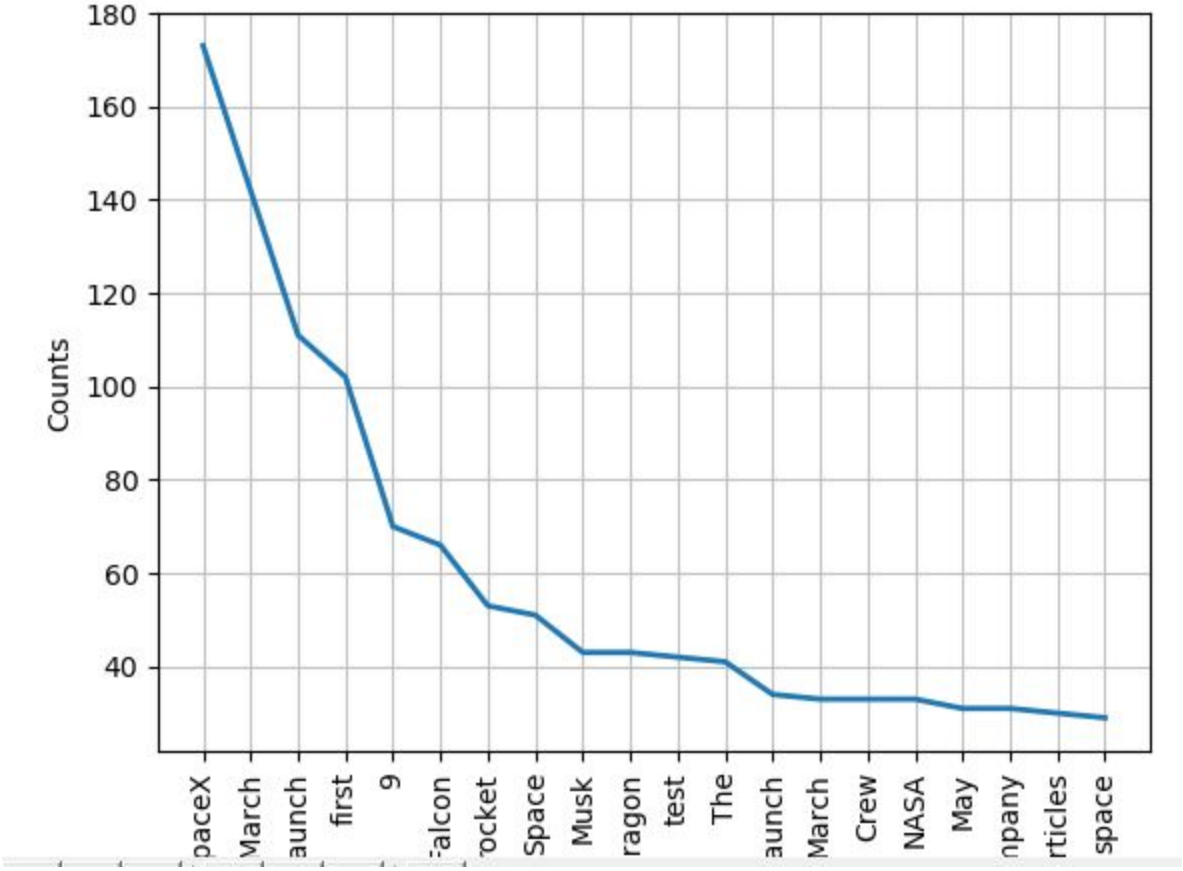
```
# pip install bs4
# pip install beautifulsoup4
# pip install html5lib
import nltk
#nltk.download()
from nltk.book import *
import urllib.request
response =  urllib.request.urlopen('https://en.wikipedia.org/wiki/SpaceX')
html = response.read()
print(html)

from bs4 import BeautifulSoup
soup = BeautifulSoup(html,'html5lib')
text = soup.get_text(strip = True)
print("Text = ")
print(text)

tokens = [t for t in text.split()]
print("Tokens = ")
print(tokens)

from nltk.corpus import stopwords

sr = stopwords.words('english')
clean_tokens = tokens[:]
for token in tokens:
   if token in stopwords.words('english'):
      clean_tokens.remove(token)
freq = nltk.FreqDist(clean_tokens)
for key, val in freq.items():
   print(str(key) + ':' + str(val))
freq.plot(20, cumulative=False)
```

Output

```
C:\Users\Inna\Anaconda3\envs\tensorflow\python.exe C:/Users/I
*** Introductory Examples for the NLTK Book ***
Loading text1, ..., text9 and sent1, ..., sent9
Type the name of the text or sentence to view it.
Type: 'texts()' or 'sents()' to list the materials.
text1: Moby Dick by Herman Melville 1851
text2: Sense and Sensibility by Jane Austen 1811
text3: The Book of Genesis
text4: Inaugural Address Corpus
text5: Chat Corpus
text6: Monty Python and the Holy Grail
text7: Wall Street Journal
text8: Personals Corpus
text9: The Man Who Was Thursday by G . K . Chesterton 1908
```




```
['SpaceX', '-', 'Wikipediadocument.documentElement.className="client-js";RLCONF={"wgBreakFrames":!1,"wgSeparatorTransformTable":["",""],"wgDigitTransformTable":["",""],"wgDefaultDateFormat":"dmy","wgMonthNames":["","January",
 "February","March","April","May","June","July","August","September","October","November","December"],"wgRequestId":"XpiOEQpAICIAAEaxb3EAAADY","wgCSPNonce":!1,"wgCanonicalNamespace":"","wgCanonicalSpecialPageName":!1,
 "wgNamespaceNumber":0,"wgPageName":"SpaceX","wgTitle":"SpaceX","wgCurRevisionId":951189758,"wgRevisionId":951189758,"wgArticleId":832774,"wgIsArticle":!0,"wgIsRedirect":!1,"wgAction":"view","wgUserName":null,
 "wgUserGroups":["*"],"wgCategories":["CS1', 'maint:', 'BOT:', 'original-url', 'status', 'unknown","CS1', 'maint:', 'unfit', 'url","CS1', 'French-language', 'sources', '(fr)","Articles', 'with', 'short', 'description","Use',
 'American', 'English', 'from', 'August', '2019","All', 'Wikipedia', 'articles', 'written', 'in', 'American', 'English","Use', 'mdy', 'dates', 'from', 'August', '2019","Coordinates', 'not', 'on', 'Wikidata",', '"All',
 'articles', 'with', 'unsourced', 'statements","Articles', 'with', 'unsourced', 'statements', 'from', 'April', '2020","All', 'articles', 'with', 'self-published', 'sources","Articles', 'with', 'self-published', 'sources',
 'from', 'April', '2020","Articles', 'containing', 'potentially', 'dated', 'statements', 'from', 'November', '2017","All', 'articles', 'containing', 'potentially', 'dated', 'statements","Articles', 'containing', 'potentially',
 'dated', 'statements', 'from', 'March', '2013","Articles', 'containing', 'potentially', 'dated', 'statements', 'from', 'May', '2012","Articles', 'containing', 'potentially', 'dated', 'statements', 'from', 'March', '2020",
 "Articles', 'containing', 'potentially', 'dated', 'statements', 'from', 'March', '2017","Wikipedia', 'articles', 'in', 'need'...
```


180
160
140
120
100
80
60
40
Counts
paceX
March
aunch
first
9
alcon
ocket
Space
Musk
ragon
test
The
aunch
March
Crew
NASA
May
npany
rticles
space