

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/:
```

```
*** 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
```

```
From... Selected Value
1 << SpaceX - Wikipediadocument.documentElement.className="client-js";RLCONF={"wgBreakFrames":1,"wgSeparatorTransformTable":["",""],"wgDigitTransformTable":["",""],"wgDefaultDateFormat":"day","wgMonthNames":["","January","Febr
2 "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 potential
3 "Official website different in Wikidata and Wikipedia","Twitter username different from Wikidata","Wikipedia articles with B18595 identifiers","Wikipedia articles with Q40 identifiers","Wikipedia articles with ISNI identi
4 "wgPageContentModel":"wikitext","wgRelevantPageName":"SpaceX","wgRelevantArticleId":"832774","wgIsProbablyEditable":10,"wgEleventPagesProbablyEditable":10,"wgRestrictionEdit":[""],"wgRestrictionMove":[""],"wgMediaViewerOnClick
5 "ready","jquery.makeCollapsible.styles":"ready","mediawiki.toc.styles":"ready","skins.vector.styles.legacy":"ready","wikibase.client.init":"ready","ext.visualEditor.desktopArticleTarget.noscript":"ready","ext.uls.interlang
6 });});SpaceXfrom Wikipedia, the free encyclopediaJump to navigationJump to searchThis article is about the rocket manufacturer. For the British art gallery, seeSpaceX (art gallery)."Space exploration technologies" redirects
7 additional terms may apply. By using this site, you agree to theTerms of UseandPrivacy Policy. Wikipedia® is a registered trademark of theWikimedia Foundation, Inc., a non-profit organization.Privacy policyAbout Wikipedia01
```

```
<< ["SpaceX", "", "Wikipediadocument.documentElement.className="client-js";RLCONF={"wgBreakFrames":1,"wgSeparatorTransformTable":["",""],"wgDigitTransformTable":["",""],"wgDefaultDateFormat":"day","wgMonthNames":["","January",
"February","March","April","May","June","July","August","September","October","November","December"],"wgRequestId":"Xp10EQaICIAAEaxb3EAAADY","wgCSPNonce":1,"wgCanonicalNamespace":"","wgCanonicalSpecialPageName":1,
"wgNamespacesNumber":0,"wgPageName":"SpaceX","wgTitle":"SpaceX","wgCurRevisionId":1951189758,"wgRevisionId":1951189758,"wgArticleId":832774,"wgIsArticle":10,"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","ndy","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"...
```

