LabWeek_1

September 3, 2021

[1]:
```python
import urllib.request
response = urllib.request.urlopen('https://www.spacex.com')
html = response.read()
print(html)
```

```
b'<!doctype html>\r\n<!--\r\n-- Space Exploration Technologies Corp.\r\n--
@version 1.0\r\n-- @date 06.02.2020\r\n-- @url
https://spacex.com\r\n-->\r\n<html lang="en">\r\n<head>\r\n\t<meta http-
equiv="Content-Type" content="text/html; charset=utf-8" />\r\n\t<meta http-
equiv="Content-Language" content="en">\r\n\t<meta name="SpaceX" />\r\n\t<meta
name="keywords" content="space,spacex,aerospace,elon musk,mars,falcon 9,falcon
heavy,dragon" />\r\n\t<meta name="description" content="SpaceX designs,
manufactures and launches advanced rockets and spacecraft. The company was
founded in 2002 to revolutionize space technology, with the ultimate goal of
enabling people to live on other planets." />\r\n\t<meta property="og:locale"
content="en_US"/>\r\n\t<meta property="og:type" content="website"/>\r\n\t<meta
property="og:title" content="SpaceX"/>\r\n\t<meta property="og:description"
content="SpaceX designs, manufactures and launches advanced rockets and
spacecraft."/>\r\n\t<meta property="og:site_name" content="SpaceX"/>\r\n\t<meta
property="og:image"
content="https://www.spacex.com/static/images/share.jpg"/>\r\n\t<meta
itemprop="name" content="SpaceX" />\r\n\t<meta itemprop="image"
content="https://www.spacex.com/static/images/share.jpg" />\r\n\t<meta
property="og:url" content="http://www.spacex.com"/>\r\n\t<meta
name="twitter:card" content="summary_large_image"/>\r\n\t<meta
name="twitter:site" content="@spacex" />\r\n\t<meta name="twitter:description"
content="SpaceX designs, manufactures and launches advanced rockets and
spacecraft."/>\r\n\t<meta name="twitter:title" content="SpaceX"/>\r\n\t<meta
name="twitter:image" content="https://www.spacex.com/static/images/share.jpg"/>\
r\n\t<title>SpaceX</title>\r\n\t<link rel="stylesheet" type="text/css"
href="style.min.css" media="screen">\r\n\t<link rel="stylesheet" type="text/css"
href="fix.css" media="screen">\r\n\t<link rel="icon" type="image/x-icon"
href="static/images/favicon.ico">\r\n\t<link rel="Shortcut Icon"
type="image/ico" href="static/images/favicon.ico">\r\n\t<meta name="viewport"
content="width=device-width, maximum-scale=1.0, user-scalable=1">\r\n\t<script
type="text/javascript" src="static/deps-min.js"></script>\r\n\t<script
type="text/javascript" async src="static/player-
min.js"></script>\r\n\r\n\t<style>\r\n\t\t@media (max-width: 700px)
```

```
{\r\n\t\t\t#feature .background {\r\n\t\t\t\tbackground-position-x:
right;\r\n\t\t\t}\r\n\t\t}\r\n\t</style>\r\n</head>\r\n<body
id="home">\r\n\t<div id="wrapper">\r\n\t\t<div id="header" class="section" data-
limit=".15">\r\n  <div class="header-bg"></div>\r\n  <div class="header-
inner">\r\n    <a href="/" id="logo">\r\n      <svg version="1.1" x="0px"
y="0px" viewbox="0 0 400 50">\r\n        <title>SpaceX Logo</title>\r\n
<g class="letter_s">\r\n          <path class="fill-white" d="M37.5,30.5H10.9v-6
.6h34.3c-0.9-2.8-3.8-5.4-8.9-5.4H11.4c-5.7,0-9,2.1-9,6.7v4.9c0,4,3.4,6.3,8.4,6.3
h26.9v7H1.5\r\n          c0.9,3.8,3.8,5.8,9,5.8h27.1c5.7,0,8.5-2.2,8.5-6.9v-4.9C
46.1,33.1,42.8,30.8,37.5,30.5z"/>\r\n        </g>\r\n        <g
class="letter_p">\r\n          <path class="fill-white" d="M91.8,18.6H59v30.7h9.
3V37.5h24.2c6.7,0,10.4-2.3,10.4-7.7v-3.4C102.8,21.4,98.6,18.6,91.8,18.6z
M94.8,28.4\r\n
c0,2.2-0.4,3.4-4,3.4H68.3l0.1-8h22c4,0,4.5,1.2,4.5,3.3V28.4z"/>\r\n
</g>\r\n        <g class="letter_a">\r\n          <polygon class="fill-white"
points="129.9,17.3 124.3,24.2 133.8,37.3 114,37.3 109.1,42.5 137.7,42.5
142.6,49.3 153.6,49.3 \t"/>\r\n        </g>\r\n        <g class="letter_c">\r\n
<path class="fill-white" d="M171.4,23.9h34.8c-0.9-3.6-4.4-5.4-9.4-5.4h-26c-4.5,0
-8.8,1.8-8.8,6.7v17.2c0,4.9,4.3,6.7,8.8,6.7h26.3\r\n
c6,0,8.1-1.7,9.1-5.8h-34.8V23.9z"/>\r\n        </g>\r\n        <g
class="letter_e">\r\n          <polygon class="fill-white" points="228.3,43.5
228.3,34.1 247,34.1 247,28.9 218.9,28.9 218.9,49.3 260.4,49.3 260.4,43.5
\t"/>\r\n          <rect class="fill-white" x="219.9" y="18.6" width="41.9"
height="5.4"/>\r\n        </g>\r\n        <g class="letter_x">\r\n
<path class="fill-white"
d="M287.6,18.6H273l17.2,12.6c2.5-1.7,5.4-3.5,8-5L287.6,18.6z"/>\r\n
<path class="fill-white"
d="M308.8,34.3c-2.5,1.7-5,3.6-7.4,5.4l13,9.5h14.7L308.8,34.3z"/>\r\n
</g>\r\n        <g class="letter_swoosh">\r\n          <path class="fill-white"
d="M399,0.7c-80,4.6-117,38.8-125.3,46.9l-1.7,1.6h14.8C326.8,9.1,384.3,2,399,0.7L
399,0.7z"/>\r\n        </g>\r\n      </svg>\r\n    </a>\r\n    <div
id="navigation">\r\n      <ul class="nav-links">\r\n        <li class="nav-
item"><a href="/vehicles/falcon-9/">Falcon 9</a></li>\r\n        <li class="nav-
item"><a href="/vehicles/falcon-heavy/">Falcon Heavy</a></li>\r\n        <li
class="nav-item"><a href="/vehicles/dragon/">Dragon</a></li>\r\n        <li
class="nav-item"><a href="/vehicles/starship/">Starship</a></li>\r\n        <li
class="nav-item"><a href="/human-spaceflight/">Human Spaceflight</a></li>\r\n
<li class="nav-item"><a href="/rideshare/">Rideshare</a></li>\r\n      </ul>\r\n
</div>\r\n  </div>\r\n\r\n  <div id="navigation-right">\r\n    <ul class="nav-
links">\r\n      <li class="nav-item"><a href="https://shop.spacex.com/"
rel="noopener" target="_blank">SHOP</a></li>\r\n    </ul>\r\n  </div>\r\n\r\n
<div id="menu-close"></div>\r\n  <div id="menu">\r\n    <div id="menu-
background"></div>\r\n    <div id="menu-navigation">\r\n      <ul class="nav-
links">\r\n        <li class="nav-item primary"><a
href="/vehicles/falcon-9/">Falcon 9</a></li>\r\n        <li class="nav-item
primary"><a href="/vehicles/falcon-heavy/">Falcon Heavy</a></li>\r\n        <li
class="nav-item primary"><a href="/vehicles/dragon/">Dragon</a></li>\r\n
<li class="nav-item primary"><a href="/vehicles/starship/">Starship</a></li>\r\n
```

```
<li class="nav-item primary"><a href="/human-spaceflight/">Human
Spaceflight</a></li>\r\n        <li class="nav-item primary"><a
href="/rideshare/">Rideshare</a></li>\r\n        <li class="nav-item
secondary"><a href="/mission/">Mission</a></li>\r\n        <li class="nav-item
secondary"><a href="/launches/">Launches</a></li>\r\n        <li class="nav-item
secondary"><a href="/careers/">Careers</a></li>\r\n        <li class="nav-item
secondary"><a href="/updates/">Updates</a></li>\r\n        <li class="nav-item
secondary"><a href="https://shop.spacex.com/" rel="noopener"
target="_blank">Shop</a></li>\r\n      </ul>\r\n    </div>\r\n  </div>\r\n
<button id="hamburger" aria-expanded="false" aria-controls="menu" aria-
label="Menu">\r\n    <div id="bar1" class="bar"></div>\r\n    <div id="bar2"
class="bar"></div>\r\n    <div id="bar3" class="bar"></div>\r\n
</button>\r\n</div>\r\n\r\n\t\t\t<div id="feature"
class="section">\r\n\t\t\t\t<div class="background" data-preload data-desktop="/
static/images/backgrounds-2021/crs-23/HP_CRS23_causeway_DSC_7027_Desktop.jpg"
data-mobile="/static/images/backgrounds-2021/crs-23/HP_CRS23_causeway_DSC_7032_M
obile.jpg">\r\n\t\t\t\t</div>\r\n\t\t\t\t<div class="section-inner
feature">\r\n\t\t\t\t\t<div class="inner-left-bottom">\r\n\t\t\t\t\t\t<h4
class="animate" style="text-transform: uppercase">Recent
Launch</h4>\r\n\t\t\t\t\t\t<h1 class="animate shadowed">CRS-23
Mission</h1>\r\n\t\t\t\t\t\t<a class="btn animate" tabindex="0"
href="/launches/">\r\n\t\t\t\t\t\t\t<div
class="hover"></div>\r\n\t\t\t\t\t\t\t<span class="text">REWATCH</span>\r\n\t\t\
t\t\t\t</a>\r\n\t\t\t\t\t</div>\r\n\t\t\t\t\t<div
id="scrollme">\r\n\t\t\t\t\t\t<svg width="30px"
height="20px">\r\n\t\t\t\t\t\t\t<path stroke="#ffffff" stroke-width="2px"
d="M2.000,5.000 L15.000,18.000 L28.000,5.000 "/>\r\n\t\t\t\t\t\t</svg>\r\n\t\t\t
\t\t</div>\r\n\t\t\t\t</div>\r\n\t\t\t</div>\r\n\r\n\t\t\t<div
class="section">\r\n\t\t\t\t<div class="background" data-preload="" data-
desktop="/static/images/backgrounds-2021/transporter-2/post-
launch/Homepage_Trans2_water_IMG_1484_Desktop.jpg" data-
mobile="/static/images/backgrounds-2021/transporter-2/post-
launch/Homepage_Trans2_presssite_DSC_8385_Mobile.jpg"></div>\r\n\t\t\t\t<div
class="section-inner feature">\r\n\t\t\t\t\t<div class="inner-left-
bottom">\r\n\t\t\t\t\t\t<h4 style="text-transform: uppercase"
class="animate">Recent Launch\r\n\t\t\t\t\t\t</h4><h2 class="animate
shadowed">Transporter-2 Mission</h2>\r\n\t\t\t\t\t\t<a class="btn animate"
tabindex="0" href="/updates/transporter-2-mission">\r\n\t\t\t\t\t\t\t<div
class="hover"></div>\r\n\t\t\t\t\t\t\t<span class="text">REWATCH</span>\r\n\t\t\
t\t\t\t</a>\r\n\t\t\t\t\t</div>\r\n\t\t\t\t</div>\r\n\t\t\t</div>\r\n\r\n\t\t\t<
div class="section">\r\n\t\t\t\t<div class="background" style="background-
position: center top" data-preload data-
desktop="/static/images/backgrounds-2021/hls-resized-2.jpg" data-
mobile="/static/images/backgrounds-2021/HLS_Mobile.jpg"></div>\r\n\t\t\t\t<div
class="section-inner feature">\r\n\t\t\t\t\t<div class="inner-left-
bottom">\r\n\t\t\t\t\t\t<h2 class="animate shadowed">Starship to Land NASA
Astronauts on the Moon</h2>\r\n\t\t\t\t\t\t<a class="btn animate" tabindex="0"
href="/updates/starship-moon-announcement">\r\n\t\t\t\t\t\t\t<div
```

```
class="hover"></div>\r\n\t\t\t\t\t\t\t\t<span class="text">LEARN MORE</span>\r\n\t
\t\t\t\t\t</a>\r\n\t\t\t\t\t</div>\r\n\t\t\t\t</div>\r\n\t\t\t</div>\r\n\r\n\t\t
\t<div class="section" data-title="ISS Docking Simulator">\r\n\t\t\t\t<div
class="background" data-preload data-
desktop="/static/images/backgrounds/iss_game.jpg" data-
mobile="/static/images/backgrounds/iss_game_mobile.jpg">\r\n\t\t\t\t\t<video
style="object-fit: cover" autoplay loop muted width="100%" height="100%"
preload="auto" autobuffer="true">\r\n \t\t\t\t\t <source type="video/mp4"
src="/media/ISS-Docking_Simulation-15sec-03-web.mp4">\r\n \t\t\t\t\t <source
type="video/webm" src="/media/ISS-Docking_Simulation-15sec-03-webm.webm">\r\n
\t\t\t\t </video>\r\n\t\t\t\t</div>\r\n\t\t\t\t<div class="section-inner
resize">\r\n\t\t\t\t\t<div class="inner-left-bottom">\r\n\t\t\t\t\t\t<h2
class="animate shadowed">DRAGON DOCKING SIMULATOR</h2>\r\n\t\t\t\t\t\t<p
class="animate shadowed">Dragon is designed to autonomously dock and undock with
the International Space Station. However, the crew can take manual control of
the spacecraft if necessary.</p>\r\n\t\t\t\t\t\t<a class="btn animate"
tabindex="0" href="https://iss-sim.spacex.com/" target="_blank" rel="noopener
noreferrer">\r\n\t\t\t\t\t\t\t<div class="hover"></div>\r\n\t\t\t\t\t\t\t<span
class="text">TRY NOW</span>\r\n\t\t\t\t\t\t</a>\r\n\t\t\t\t\t</div>\r\n\t\t\t\t<
/div>\r\n\t\t\t</div>\r\n\r\n\t\t\t<div id="footer">\r\n  <p>\r\n
<span>SpaceX &copy; 2021</span>\r\n    <a href="https://twitter.com/spacex"
rel="noopener" target="_blank" class="social">TWITTER</a>\r\n    <a
href="https://www.youtube.com/spacex" rel="noopener" target="_blank"
class="social">YOUTUBE</a>\r\n    <a href="https://www.instagram.com/spacex/"
rel="noopener" target="_blank" class="social">INSTAGRAM</a>\r\n    <a
href="https://www.flickr.com/photos/spacex" rel="noopener" target="_blank"
class="social">FLICKR</a>\r\n    <a
href="https://www.linkedin.com/company/spacex" rel="noopener" target="_blank"
class="social">LINKEDIN</a>\r\n    <a href="/media/privacy_policy_spacex.pdf"
target="_blank" class="social">PRIVACY POLICY</a>\r\n  </p>\r\n</div>\r\n\r\n
</div>\r\n\r\n\t<div id="modal" class="modal" aria-modal>\r\n  \t<div
class="modal-transform">\r\n  \t\t<div class="modal-bg"></div>\r\n\t\t\t<div
class="modal-inner">\r\n\t\t\t \t<div class="youtube-
wrapper"></div>\r\n\t\t\t\t<span></span>\r\n\t\t\t</div>\r\n\t\t\t<a href="#"
class="modal-close">\r\n\t\t\t\t<svg version="1.1"
xmlns="http://www.w3.org/2000/svg" xmlns:xlink="http://www.w3.org/1999/xlink"
x="0px" y="0px" viewbox="0 0 50 50" enable-background="new 0 0 50 50"
xml:space="preserve">\r\n\t\t\t\t\t<line class="line1" fill="none" x1="13"
y1="13" x2="33" y2="33" stroke-linecap="square"/>\r\n\t\t\t\t\t<line
class="line2" fill="none" x1="13" y1="33" x2="33" y2="13" stroke-
linecap="square"/>\r\n\t\t\t\t</svg>\r\n\t\t\t</a>\r\n
\t</div>\r\n\t</div>\r\n\t<script type="text/javascript" src="static/core-
min.js"></script>\r\n</body>\r\n</html>\r\n'
```

0.0.1 Use BeautifulSoup to clean the grabbed text like

[2]:

```python
from bs4 import BeautifulSoup
soup = BeautifulSoup(html,"html5lib")

text = soup.get_text(strip=True)
print(text)
```

```
SpaceX@media (max-width: 700px) {
                        #feature .background {
                                background-position-x: right;
                        }
                }SpaceX LogoFalcon 9Falcon HeavyDragonStarshipHuman
SpaceflightRideshareSHOPFalcon 9Falcon HeavyDragonStarshipHuman
SpaceflightRideshareMissionLaunchesCareersUpdatesShopRecent LaunchCRS-23
MissionREWATCHRecent LaunchTransporter-2 MissionREWATCHStarship to Land NASA
Astronauts on the MoonLEARN MOREDRAGON DOCKING SIMULATORDragon is designed to
autonomously dock and undock with the International Space Station. However, the
crew can take manual control of the spacecraft if necessary.TRY NOWSpaceX ©
2021TWITTERYOUTUBEINSTAGRAMFLICKRLINKEDINPRIVACY POLICY
```

0.0.2 Clean text from the crawled web page

[3]:

```python
tokens = [t for t in text.split()]
print(tokens)
```

```
['SpaceX@media', '(max-width:', '700px)', '{', '#feature', '.background', '{',
'background-position-x:', 'right;', '}', '}SpaceX', 'LogoFalcon', '9Falcon',
'HeavyDragonStarshipHuman', 'SpaceflightRideshareSHOPFalcon', '9Falcon',
'HeavyDragonStarshipHuman',
'SpaceflightRideshareMissionLaunchesCareersUpdatesShopRecent', 'LaunchCRS-23',
'MissionREWATCHRecent', 'LaunchTransporter-2', 'MissionREWATCHStarship', 'to',
'Land', 'NASA', 'Astronauts', 'on', 'the', 'MoonLEARN', 'MOREDRAGON', 'DOCKING',
'SIMULATORDragon', 'is', 'designed', 'to', 'autonomously', 'dock', 'and',
'undock', 'with', 'the', 'International', 'Space', 'Station.', 'However,',
'the', 'crew', 'can', 'take', 'manual', 'control', 'of', 'the', 'spacecraft',
'if', 'necessary.TRY', 'NOWSpaceX', '©',
'2021TWITTERYOUTUBEINSTAGRAMFLICKRLINKEDINPRIVACY', 'POLICY']
```

0.0.3 Count word frequency

[5]:

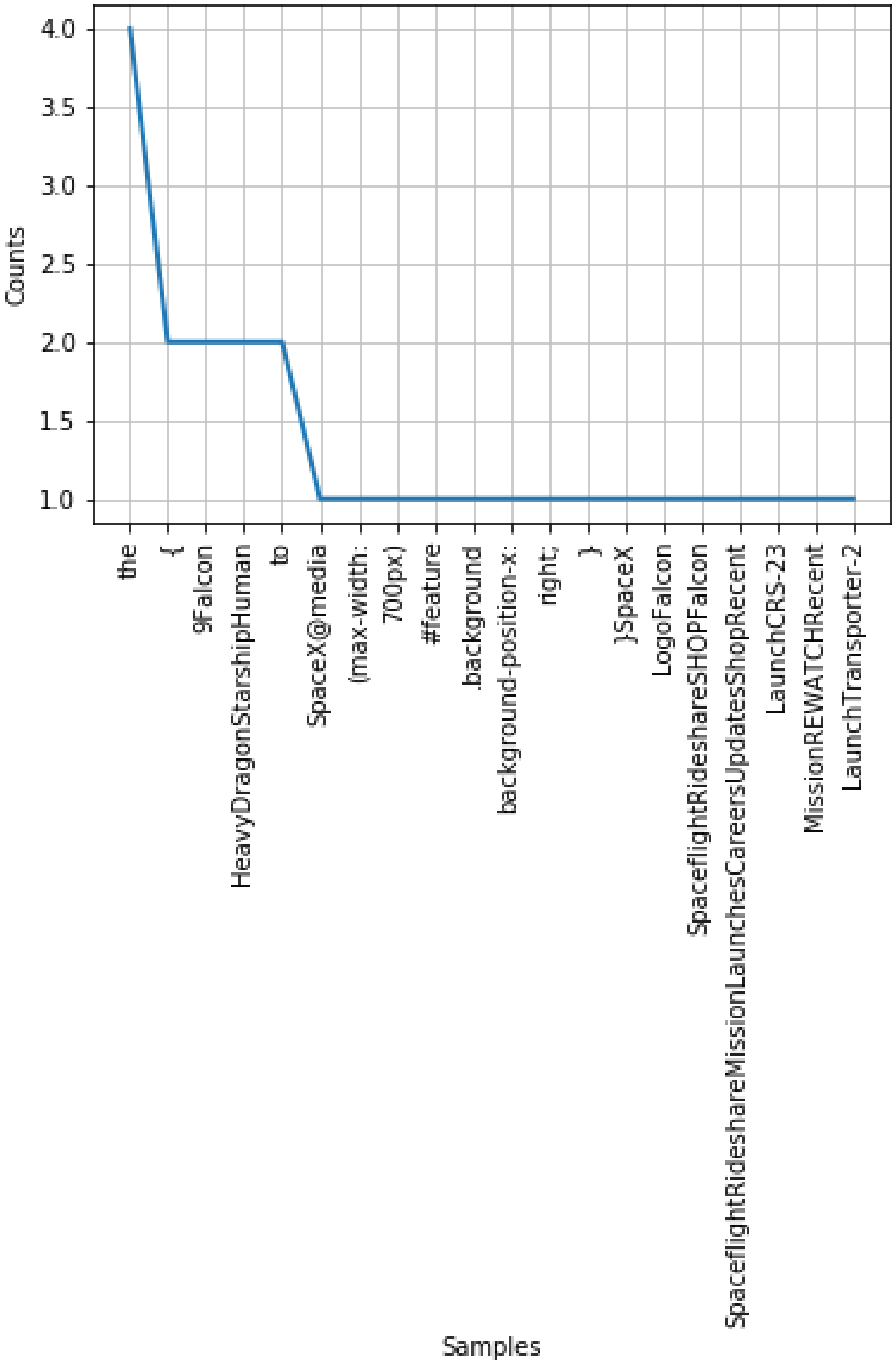
```python
import nltk
freq = nltk.FreqDist(tokens)

for key,val in freq.items():

    print (str(key) + ':' + str(val))
```

```
SpaceX@media:1
(max-width::1
700px):1
{:2
#feature:1
.background:1
background-position-x::1
right;:1
}:1
}SpaceX:1
LogoFalcon:1
9Falcon:2
HeavyDragonStarshipHuman:2
SpaceflightRideshareSHOPFalcon:1
SpaceflightRideshareMissionLaunchesCareersUpdatesShopRecent:1
LaunchCRS-23:1
MissionREWATCHRecent:1
LaunchTransporter-2:1
MissionREWATCHStarship:1
to:2
Land:1
NASA:1
Astronauts:1
on:1
the:4
MoonLEARN:1
MOREDRAGON:1
DOCKING:1
SIMULATORDragon:1
is:1
designed:1
autonomously:1
dock:1
and:1
undock:1
with:1
International:1
Space:1
Station.:1
However,:1
crew:1
can:1
take:1
manual:1
control:1
of:1
spacecraft:1
if:1
```

```
necessary.TRY:1
NOWSpaceX:1
©:1
2021TWITTERYOUTUBEINSTAGRAMFLICKRLINKEDINPRIVACY:1
POLICY:1
```

0.0.4 plot a graph for those tokens

```
[6]: freq.plot(20, cumulative=False)
```


4.0
3.5
3.0
2.5
2.0
1.5
1.0
Counts
the
{
9Falcon
HeavyDragonStarshipHuman
to
SpaceX@media
(max-width:
700px)
#feature
.background
background-position-x:
right;
}
}SpaceX
LogoFalcon
SpaceflightRideshareSHOPFalcon
SpaceflightRideshareMissionLaunchesCareersUpdatesShopRecent
LaunchCRS-23
MissionREWATCHRecent
LaunchTransporter-2
Samples

```
[6]: <matplotlib.axes._subplots.AxesSubplot at 0x7f847ac39bd0>
```

0.0.5 Remove stop words using NLTK

[7]:

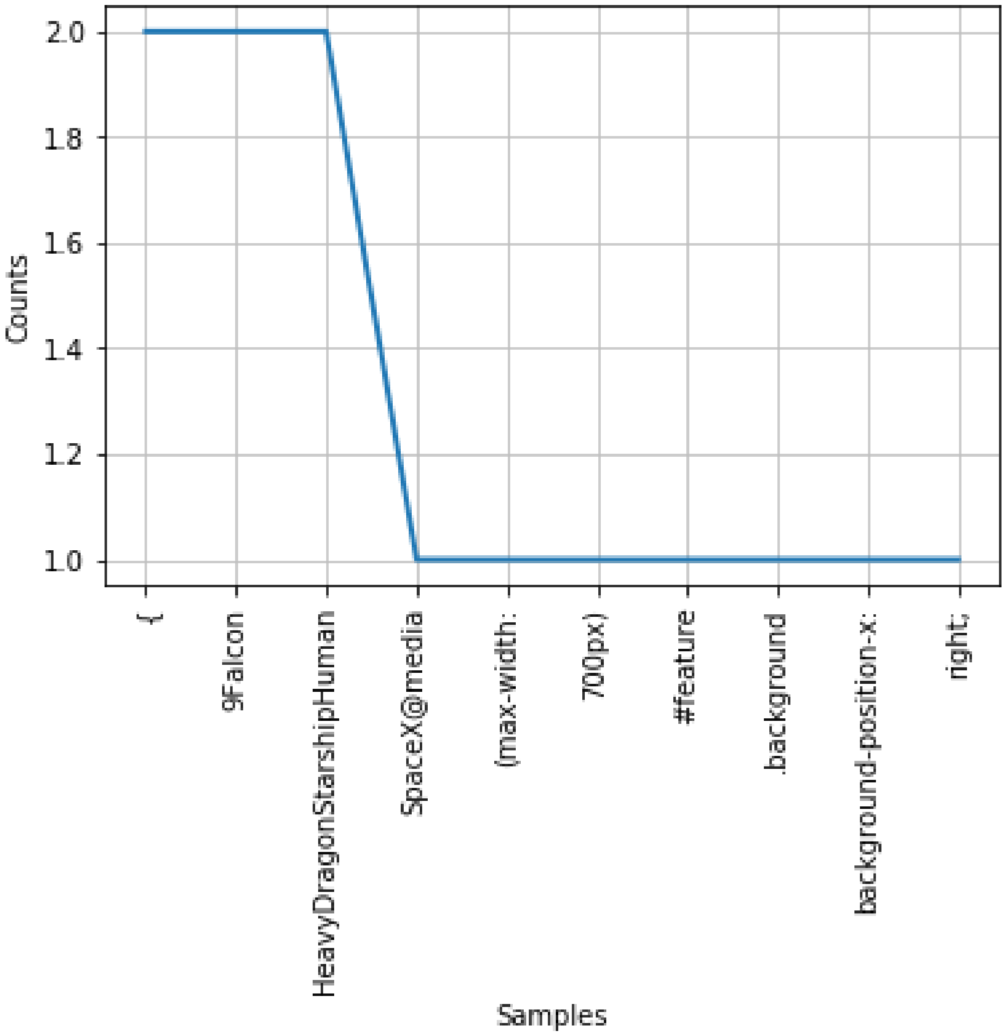
```python
from nltk.corpus import stopwords
clean_tokens = tokens[:]

sr = stopwords.words('english')

for token in tokens:

    if token in sr:
        clean_tokens.remove(token)
print(clean_tokens)
```

```
['SpaceX@media', '(max-width:', '700px)', '{', '#feature', '.background', '{',
'background-position-x:', 'right;', '}', '}SpaceX', 'LogoFalcon', '9Falcon',
'HeavyDragonStarshipHuman', 'SpaceflightRideshareSHOPFalcon', '9Falcon',
'HeavyDragonStarshipHuman',
'SpaceflightRideshareMissionLaunchesCareersUpdatesShopRecent', 'LaunchCRS-23',
'MissionREWATCHRecent', 'LaunchTransporter-2', 'MissionREWATCHStarship', 'Land',
'NASA', 'Astronauts', 'MoonLEARN', 'MOREDRAGON', 'DOCKING', 'SIMULATORDragon',
'designed', 'autonomously', 'dock', 'undock', 'International', 'Space',
'Station.', 'However,', 'crew', 'take', 'manual', 'control', 'spacecraft',
'necessary.TRY', 'NOWSpaceX', '©',
'2021TWITTERYOUTUBEINSTAGRAMFLICKRLINKEDINPRIVACY', 'POLICY']
```

0.0.6 Plot graph with stopwords

[8]:

```python
freq = nltk.FreqDist(clean_tokens)
freq.plot(10, cumulative=False)
```



```
[8]: <matplotlib.axes._subplots.AxesSubplot at 0x7f847abcf2d0>
```

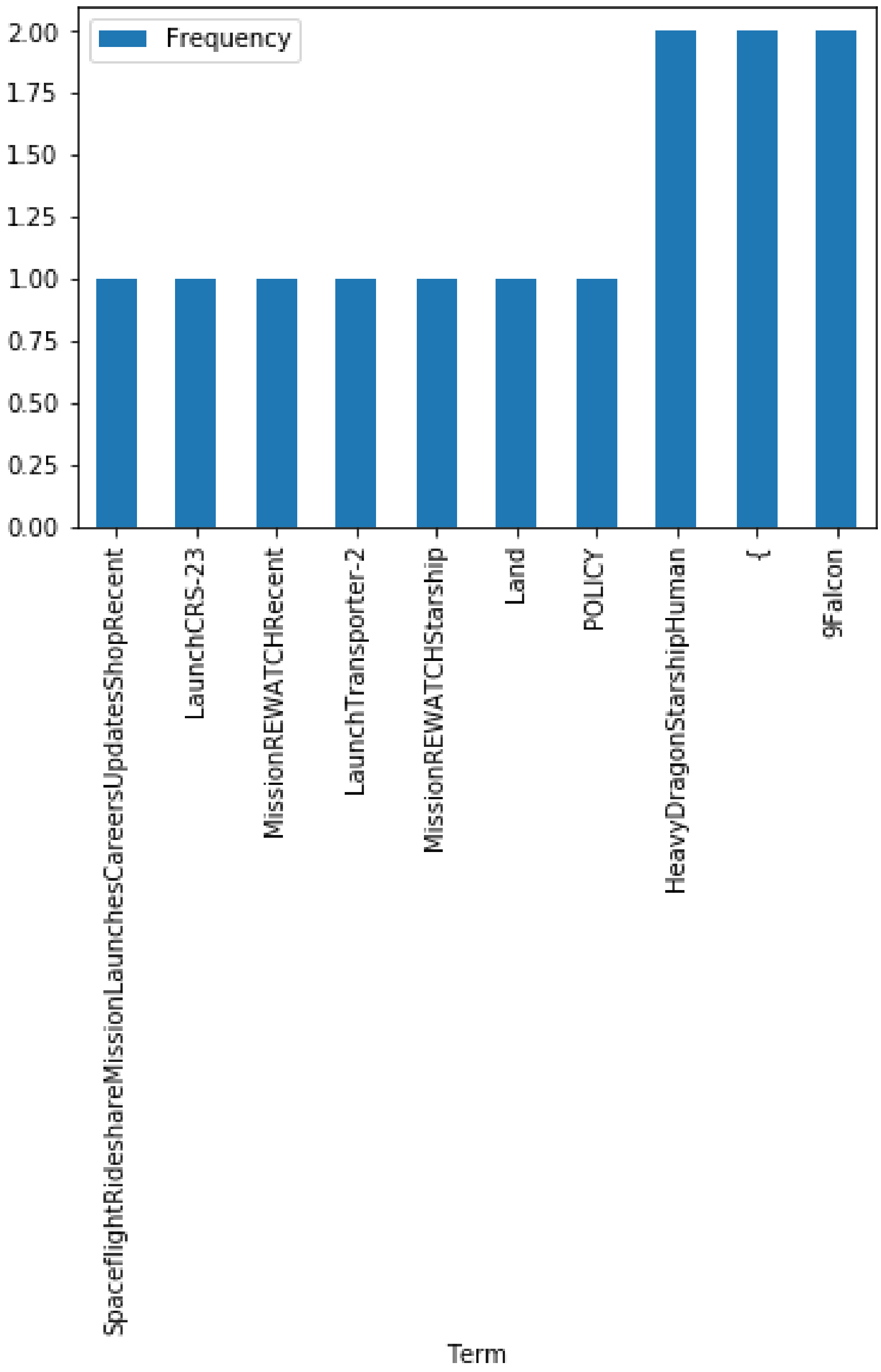
0.0.7 Created Bar graph

```
[43]: import pandas as pd
df_fdist = pd.DataFrame.from_dict(freq, orient='index')
df_fdist.columns = ['Frequency']
df_fdist.index.name = 'Term'
# print(df_fdist)
df_fdist = df_fdist.sort_values(by=['Frequency'])
df_fdist=df_fdist[-10:]
print(df_fdist)
df_fdist.plot.bar()
```

```
                                        Frequency
```

```
Term
SpaceflightRideshareMissionLaunchesCareersUpdat…          1
LaunchCRS-23                                                1
MissionREWATCHRecent                                        1
LaunchTransporter-2                                         1
MissionREWATCHStarship                                      1
Land                                                        1
POLICY                                                      1
HeavyDragonStarshipHuman                                    2
{                                                           2
9Falcon                                                     2
```

```
[43]: <matplotlib.axes._subplots.AxesSubplot at 0x7f847f6e9d50>
```


Frequency
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
SpaceflightRideshareMissionLaunchesCareersUpdatesShopRecent
LaunchCRS-23
MissionREWATCHRecent
LaunchTransporter-2
MissionREWATCHStarship
Land
POLICY
HeavyDragonStarshipHuman
{
9Falcon
Term

```
[ ]:
```