IR Practical No. 3

Neeraj Appari T073

Aim: Write a program for Pre-processing of a Text Document: stop word removal.

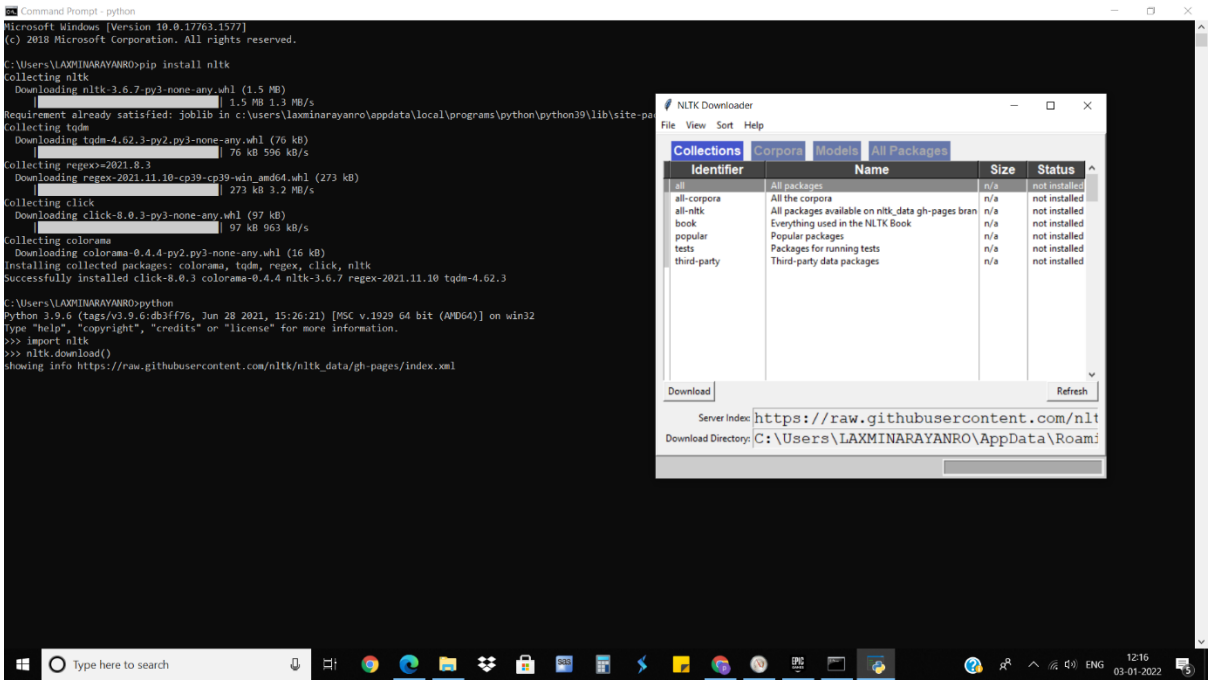
Description:

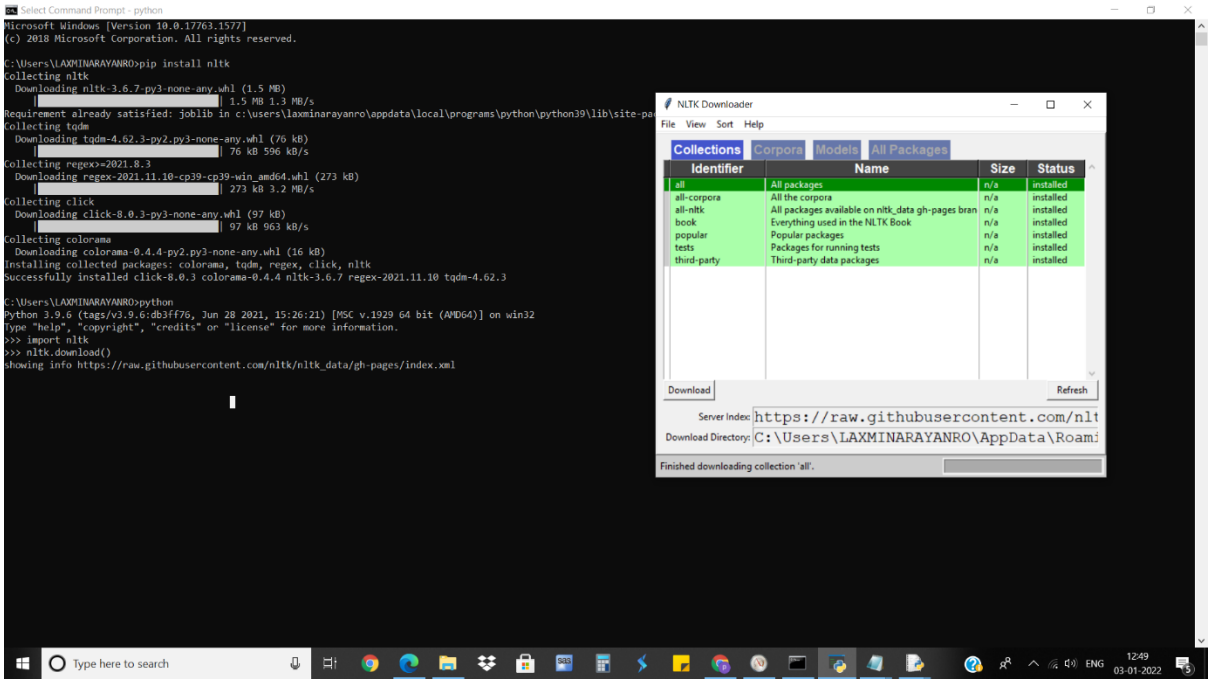
1. NLTK (Natural Language Toolkit)

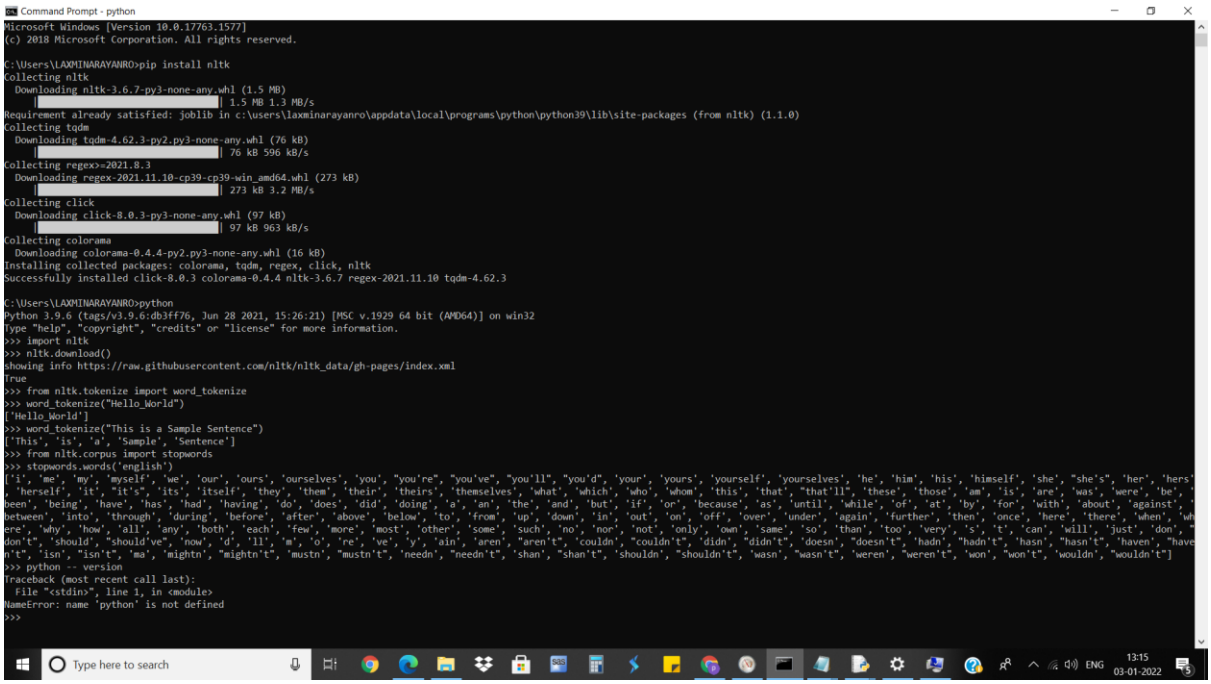
NLTK (Natural Language Toolkit) is the go-to API for NLP (Natural Language Processing) with Python. It is a really powerful tool to preprocess text data for further analysis like with ML models for instance. It helps convert text into numbers, which the model can then easily work with.

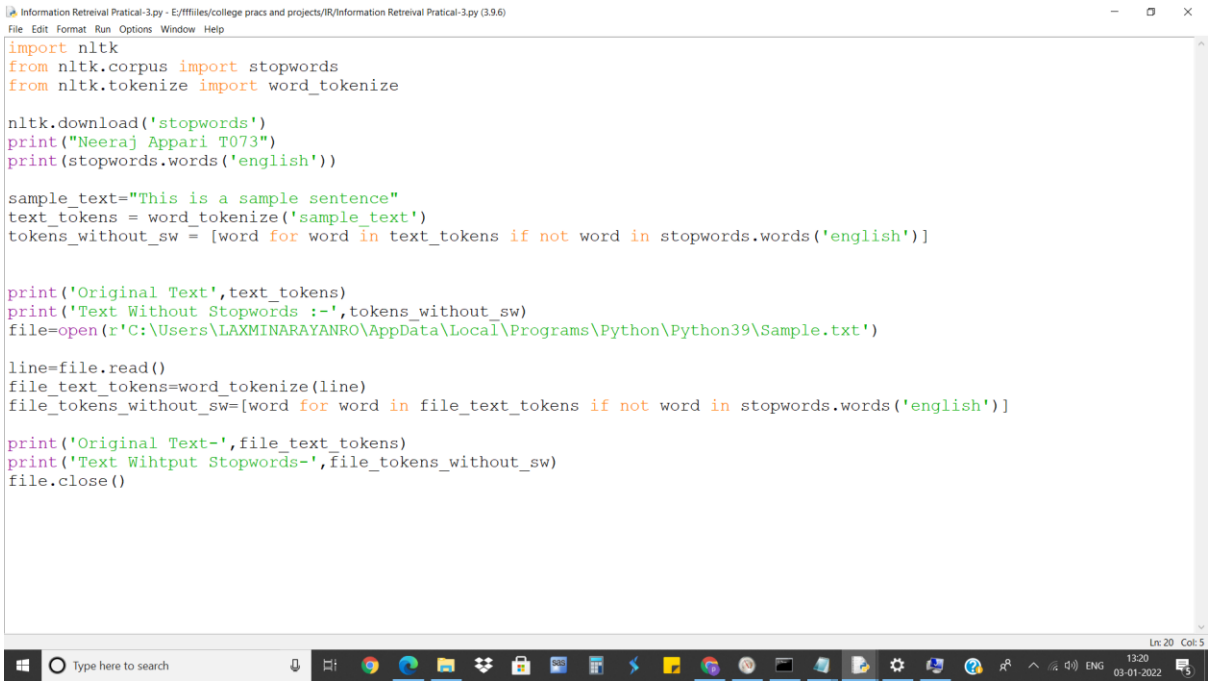
2. Stop words

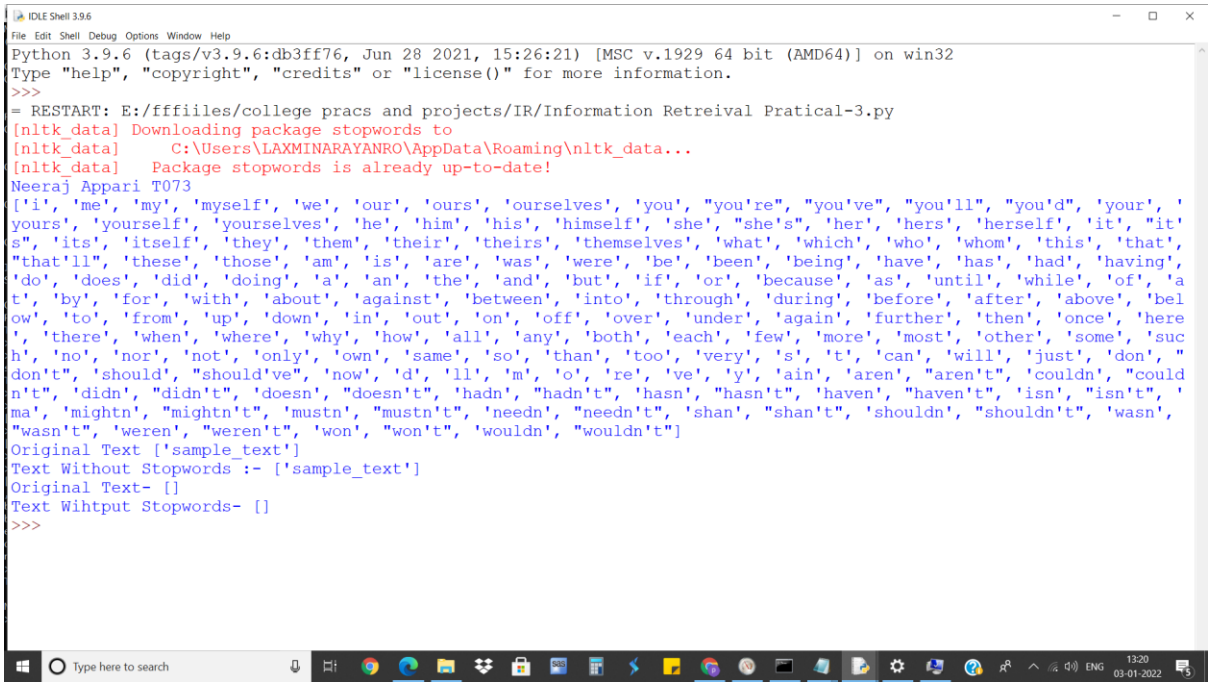
Stopwords are the English words which does not add much meaning to a sentence. They can safely be ignored without sacrificing the meaning of the sentence. For example, the words like the, he, have etc. Such words are already captured this in corpus named corpus. We first download it to our python environment.


Command Prompt - python
Microsoft Windows [Version 10.0.17763.1577]
(c) 2018 Microsoft Corporation. All rights reserved.
C:\Users\LAXMINARAYANRO>pip install nltk
Collecting nltk
Downloading nltk-3.6.7-py3-none-any.whl (1.5 MB)
1.5 MB 1.3 MB/s
Requirement already satisfied: joblib in c:\users\laxminarayanro\appdata\local\programs\python\python39\lib\site-pa
Collecting tqdm
Downloading tqdm-4.62.3-py2.py3-none-any.whl (76 kB)
76 kB 596 kB/s
Collecting regex>=2021.8.3
Downloading regex-2021.11.10-cp39-cp39-win_amd64.whl (273 kB)
273 kB 3.2 MB/s
Collecting click
Downloading click-8.0.3-py3-none-any.whl (97 kB)
97 kB 963 kB/s
Collecting colorama
Downloading colorama-0.4.4-py2.py3-none-any.whl (16 kB)
Installing collected packages: colorama, tqdm, regex, click, nltk
Successfully installed click-8.0.3 colorama-0.4.4 nltk-3.6.7 regex-2021.11.10 tqdm-4.62.3
C:\Users\LAXMINARAYANRO>python
Python 3.9.6 (tags/v3.9.6:db3ff76, Jun 28 2021, 15:26:21) [MSC v.1929 64 bit (AMD64)] on win32
Type "help", "copyright", "credits" or "license" for more information.
>>> import nltk
>>> nltk.download()
showing info https://raw.githubusercontent.com/nltk/nltk_data/gh-pages/index.xml
NLTK Downloader
File View Sort Help
Collections Corpora Models All Packages
Identifier Name Size Status
all All packages n/a not installed
all-corpora All the corpora n/a not installed
all-nltk All packages available on nltk_data gh-pages bran n/a not installed
book Everything used in the NLTK Book n/a not installed
popular Popular packages n/a not installed
tests Packages for running tests n/a not installed
third-party Third-party data packages n/a not installed
Download
Refresh
Server Index: https://raw.githubusercontent.com/nlt
Download Directory: C:\Users\LAXMINARAYANRO\AppData\Roami
Type here to search
ENG
12:16
03-01-2022


Select Command Prompt - python
Microsoft Windows [Version 10.0.17763.1577]
(c) 2018 Microsoft Corporation. All rights reserved.
C:\Users\LAXMINARAYANRO>pip install nltk
Collecting nltk
Downloading nltk-3.6.7-py3-none-any.whl (1.5 MB)
1.5 MB 1.3 MB/s
Requirement already satisfied: joblib in c:\users\laxminarayanro\appdata\local\programs\python\python39\lib\site-pa
Collecting tqdm
Downloading tqdm-4.62.3-py2.py3-none-any.whl (76 kB)
76 kB 596 kB/s
Collecting regex>=2021.8.3
Downloading regex-2021.11.10-cp39-cp39-win_amd64.whl (273 kB)
273 kB 3.2 MB/s
Collecting click
Downloading click-8.0.3-py3-none-any.whl (97 kB)
97 kB 963 kB/s
Collecting colorama
Downloading colorama-0.4.4-py2.py3-none-any.whl (16 kB)
Installing collected packages: colorama, tqdm, regex, click, nltk
Successfully installed click-8.0.3 colorama-0.4.4 nltk-3.6.7 regex-2021.11.10 tqdm-4.62.3
C:\Users\LAXMINARAYANRO>python
Python 3.9.6 (tags/v3.9.6:db3ff76, Jun 28 2021, 15:26:21) [MSC v.1929 64 bit (AMD64)] on win32
Type "help", "copyright", "credits" or "license" for more information.
>>> import nltk
>>> nltk.download()
showing info https://raw.githubusercontent.com/nltk/nltk_data/gh-pages/index.xml
NLTK Downloader
File View Sort Help
Collections Corpora Models All Packages
Identifier Name Size Status
all All packages n/a installed
all-corpora All the corpora n/a installed
all-nltk All packages available on nltk_data gh-pages bran n/a installed
book Everything used in the NLTK Book n/a installed
popular Popular packages n/a installed
tests Packages for running tests n/a installed
third-party Third-party data packages n/a installed
Download
Refresh
Server Index: https://raw.githubusercontent.com/nlt
Download Directory: C:\Users\LAXMINARAYANRO\AppData\Roami
Finished downloading collection 'all'.
Type here to search
ENG
12:49
03-01-2022


Command Prompt - python
Microsoft Windows [Version 10.0.17763.1577]
(c) 2018 Microsoft Corporation. All rights reserved.
C:\Users\LAXMINARAYANRO>pip install nltk
Collecting nltk
Downloading nltk-3.6.7-py3-none-any.whl (1.5 MB)
1.5 MB 1.3 MB/s
Requirement already satisfied: joblib in c:\users\laxminarayanro\appdata\local\programs\python\python39\lib\site-packages (from nltk) (1.1.0)
Collecting tqdm
Downloading tqdm-4.62.3-py2.py3-none-any.whl (76 kB)
76 kB 596 kB/s
Collecting regex>=2021.8.3
Downloading regex-2021.11.10-cp39-cp39-win_amd64.whl (273 kB)
273 kB 3.2 MB/s
Collecting click
Downloading click-8.0.3-py3-none-any.whl (97 kB)
97 kB 963 kB/s
Collecting colorama
Downloading colorama-0.4.4-py2.py3-none-any.whl (16 kB)
Installing collected packages: colorama, tqdm, regex, click, nltk
Successfully installed click-8.0.3 colorama-0.4.4 nltk-3.6.7 regex-2021.11.10 tqdm-4.62.3
C:\Users\LAXMINARAYANRO>python
Python 3.9.6 (tags/v3.9.6:db3ff76, Jun 28 2021, 15:26:21) [MSC v.1929 64 bit (AMD64)] on win32
Type "help", "copyright", "credits" or "license" for more information.
>>> import nltk
>>> nltk.download()
showing info https://raw.githubusercontent.com/nltk/nltk_data/gh-pages/index.xml
True
>>> from nltk.tokenize import word_tokenize
>>> word_tokenize("Hello_World")
['Hello_World']
>>> word_tokenize("This is a Sample Sentence")
['This', 'is', 'a', 'Sample', 'Sentence']
>>> from nltk.corpus import stopwords
>>> stopwords.words('english')
>>> python -- version
Traceback (most recent call last):
File "<stdin>", line 1, in <module>
NameError: name 'python' is not defined
>>>
Type here to search
ENG
13:15
03-01-2022


Information Retreival Pratical-3.py - E:/fffiiles/college pracs and projects/IR/Information Retreival Pratical-3.py (3.9.6)
File Edit Format Run Options Window Help
import nltk
from nltk.corpus import stopwords
from nltk.tokenize import word_tokenize
nltk.download('stopwords')
print("Neeraj Appari T073")
print(stopwords.words('english'))
sample_text="This is a sample sentence"
text_tokens = word_tokenize('sample_text')
tokens_without_sw = [word for word in text_tokens if not word in stopwords.words('english')]
print('Original Text',text_tokens)
print('Text Without Stopwords :-',tokens_without_sw)
file=open(r'C:\Users\LAXMINARAYANRO\AppData\Local\Programs\Python\Python39\Sample.txt')
line=file.read()
file_text_tokens=word_tokenize(line)
file_tokens_without_sw=[word for word in file_text_tokens if not word in stopwords.words('english')]
print('Original Text-',file_text_tokens)
print('Text Wihtput Stopwords-',file_tokens_without_sw)
file.close()
Ln: 20 Col: 5
Type here to search
ENG
13:20
03-01-2022


IDLE Shell 3.9.6
File Edit Shell Debug Options Window Help
Python 3.9.6 (tags/v3.9.6:db3ff76, Jun 28 2021, 15:26:21) [MSC v.1929 64 bit (AMD64)] on win32
Type "help", "copyright", "credits" or "license()" for more information.
>>>
= RESTART: E:/fffiiles/college pracs and projects/IR/Information Retreival Pratical-3.py
[nltk_data] Downloading package stopwords to
[nltk_data] C:\Users\LAXMINARAYANRO\AppData\Roaming\nltk_data...
[nltk_data] Package stopwords is already up-to-date!
Neeraj Appari T073
['i', 'me', 'my', 'myself', 'we', 'our', 'ours', 'ourselves', 'you', "you're", "you've", "you'll", "you'd", 'your', '
yours', 'yourself', 'yourselves', 'he', 'him', 'his', 'himself', 'she', "she's", 'her', 'hers', 'herself', 'it', "it'
s", 'its', 'itself', 'they', 'them', 'their', 'theirs', 'themselves', 'what', 'which', 'who', 'whom', 'this', 'that',
"that'll", 'these', 'those', 'am', 'is', 'are', 'was', 'were', 'be', 'been', 'being', 'have', 'has', 'had', 'having',
'do', 'does', 'did', 'doing', 'a', 'an', 'the', 'and', 'but', 'if', 'or', 'because', 'as', 'until', 'while', 'of', 'a
t', 'by', 'for', 'with', 'about', 'against', 'between', 'into', 'through', 'during', 'before', 'after', 'above', 'bel
ow', 'to', 'from', 'up', 'down', 'in', 'out', 'on', 'off', 'over', 'under', 'again', 'further', 'then', 'once', 'here
', 'there', 'when', 'where', 'why', 'how', 'all', 'any', 'both', 'each', 'few', 'more', 'most', 'other', 'some', 'suc
h', 'no', 'nor', 'not', 'only', 'own', 'same', 'so', 'than', 'too', 'very', 's', 't', 'can', 'will', 'just', 'don', "
don't", 'should', "should've", 'now', 'd', 'll', 'm', 'o', 're', 've', 'y', 'ain', 'aren', "aren't", 'couldn', "could
n't", 'didn', "didn't", 'doesn', "doesn't", 'hadn', "hadn't", 'hasn', "hasn't", 'haven', "haven't", 'isn', "isn't", '
ma', 'mightn', "mightn't", 'mustn', "mustn't", 'needn', "needn't", 'shan', "shan't", 'shouldn', "shouldn't", 'wasn',
"wasn't", 'weren', "weren't", 'won', "won't", 'wouldn', "wouldn't"]
Original Text ['sample_text']
Text Without Stopwords :- ['sample_text']
Original Text- []
Text Wihtput Stopwords- []
>>>
Type here to search
ENG
13:20
03-01-2022