

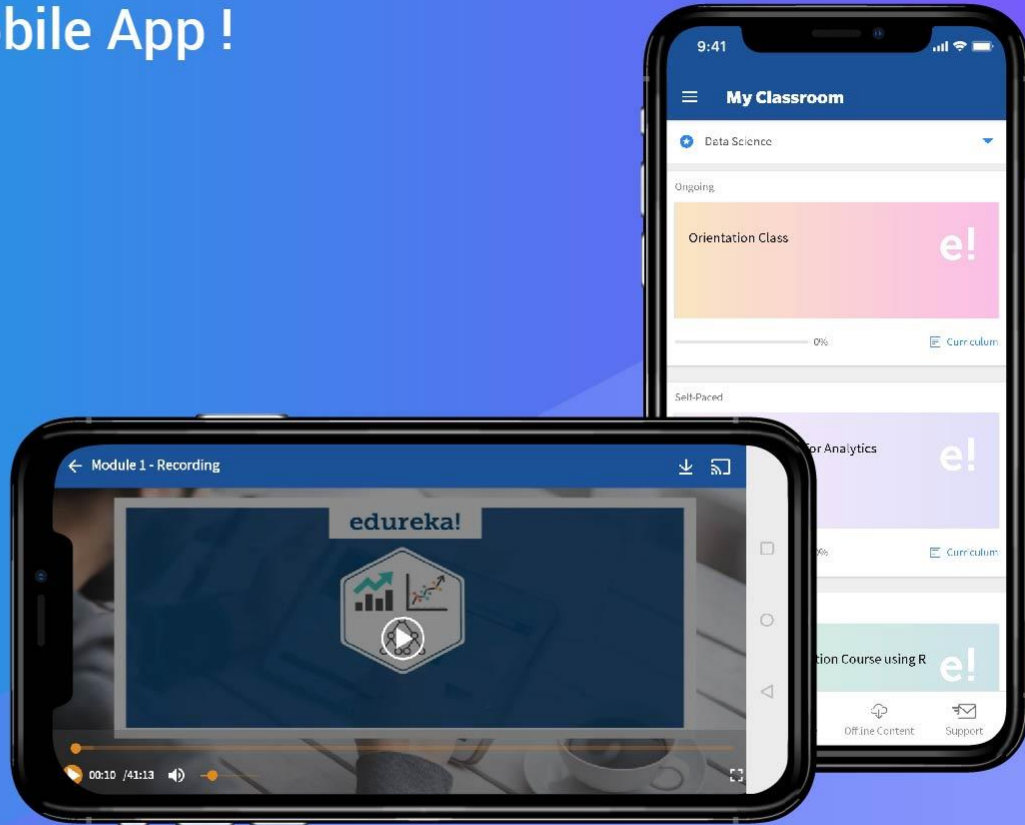
How to make the best use of Live Sessions

- Please login on time
- Please do a check on your network connection and audio before the class to have a smooth session
- All participants will be on mute, by default. You will be unmuted when requested or as needed
- Please use the “Questions” panel on your webinar tool to interact with the instructor at any point during the class
- Ask and answer questions to make your learning interactive
- Please have the support phone number (US : 1855 818 0063 (toll free), India : +91 90191 17772) and raise tickets from LMS in case of any issues with the tool
- Most often logging off or rejoining will help solve the tool related issues

Download Edureka Mobile App !

- ✓ Watch course videos offline
- ✓ Get class reminders
- ✓ Contact customer support

Learn on the go



edureka!

A decorative graphic in the top right corner consisting of a network of white dots connected by thin white lines, forming a complex, interconnected web-like structure.

Text Mining And NLP

A decorative graphic in the bottom left corner consisting of a network of white dots connected by thin white lines, forming a complex, interconnected web-like structure.

Objectives

After completing this module, you should be able to:

- Gain an understanding of Text Mining & NLP
- Manipulate various file types
- Use the 'NLTK' library



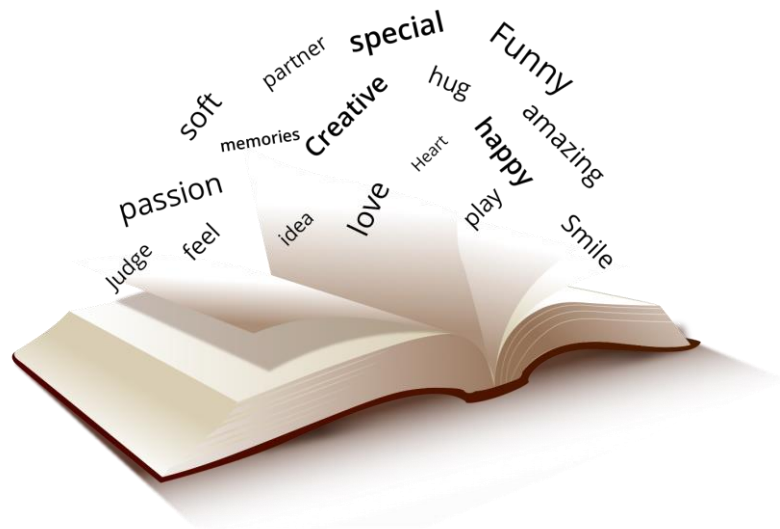


What Is Text Mining?

What Is Text Mining?


- Text Mining / Text Analytics is the process of deriving meaningful information from natural language text
- Text mining usually involves the process of structuring the input text, deriving patterns within the structured data, and finally evaluating and interpreting the output

Compared with the kind of data stored in databases, text is unstructured, amorphous, and difficult to deal with algorithmically. Nevertheless, in modern culture, text is the most common vehicle for the formal exchange of information





Need Of Text Mining


- With the advancement of technology, more and more data is available in digital form. Among which, most of the data (approx. 85%) is in unstructured textual form
- So it has become essential to develop better techniques and algorithms to extract useful and interesting information from this large amount of textual data. Hence, the area of text mining and information extraction has become popular areas of research, to extract interesting and useful information



3,574,010,156
Internet Users in the world


1,153,752,381
Total number of Websites

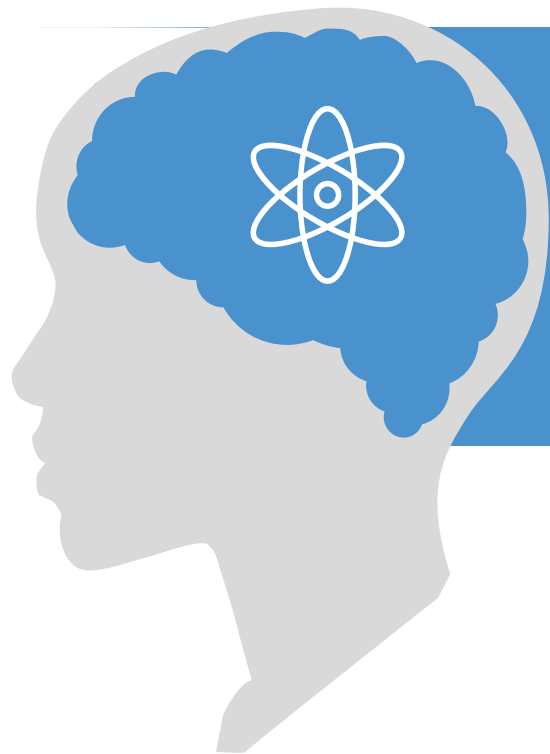

120,991,996,102
Emails sent [today](#)


2,670,891,356
Google searches [today](#)


2,493,471
Blog posts written [today](#)


340,974,567
Tweets sent [today](#)

Text Mining And NLP



As, Text Mining refers to the process of deriving high quality information from the text . The overall goal is, essentially to turn text into data for analysis, via application of Natural Language Processing (NLP)

Natural Language Processing

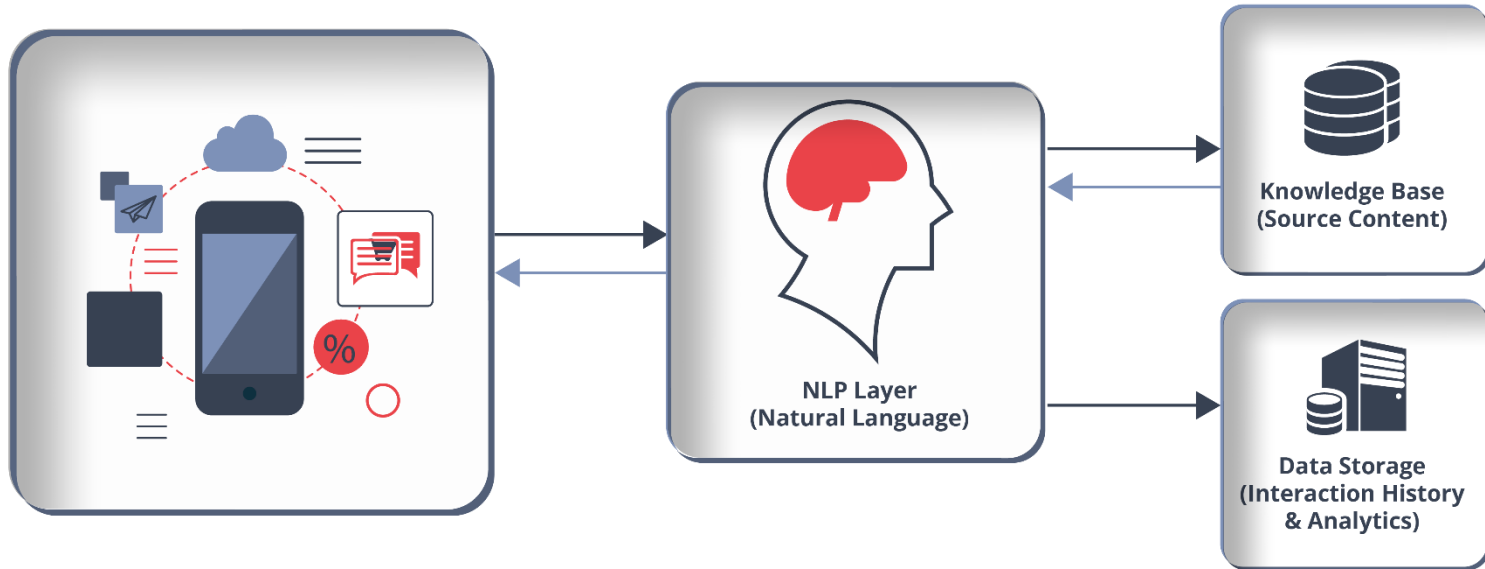




What Is NLP?

Basic Structure Of A NLP Application

Chatbot considered here:



Basic Structure Of A NLP Application



Knowledge Base
(Source Content)

Knowledge Base: It contains the database of information that is used to equip chatbots with the information needed to respond to queries of customers request.

Basic Structure Of A NLP Application



Knowledge Base
(Source Content)

Knowledge Base: It contains the database of information that is used to equip chatbots with the information needed to respond to queries of customers request.



Data Storage
(Interaction History
& Analytics)

Data Store: It contains interaction history of chatbot with users.

Basic Structure Of A NLP Application



NLP Layer: It translates user queries into information that can be used for appropriate responses.

Application Layer: It is the application interface that is used to interact with the user.



Applications Of NLP

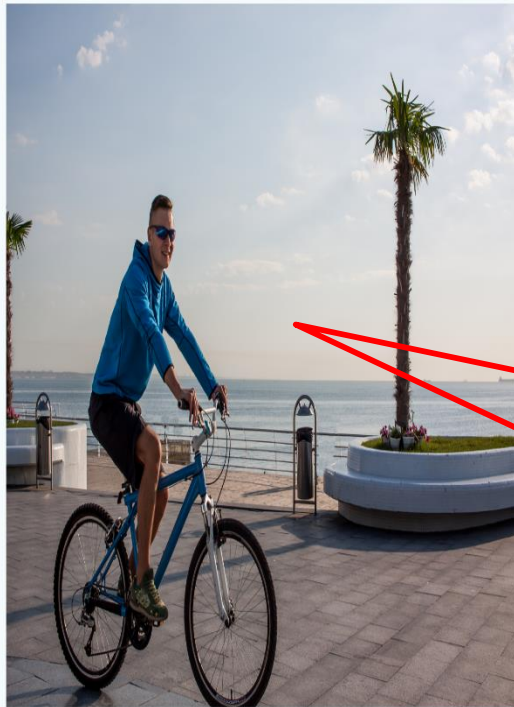


**Speech
Recognition**

Text Classification



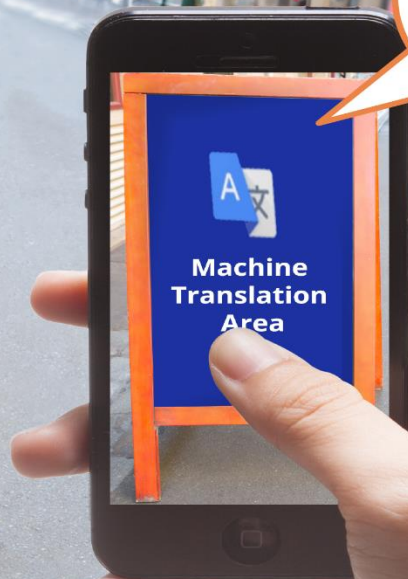
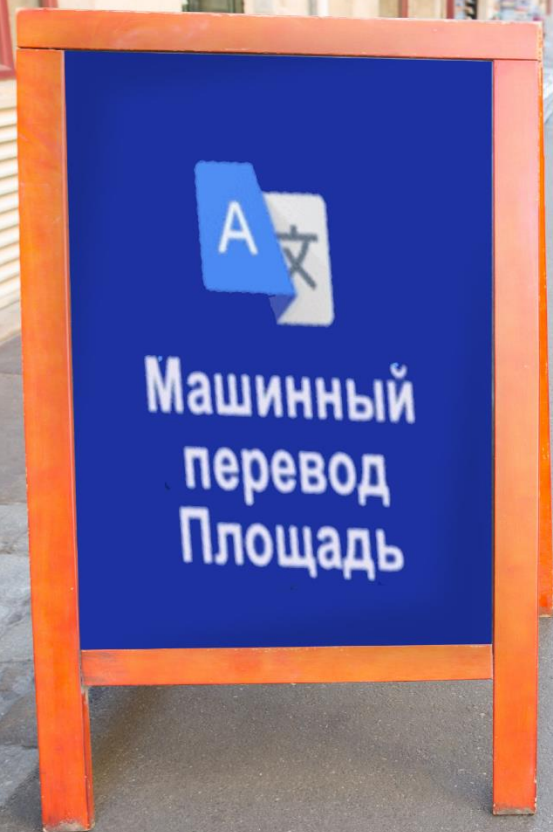
A living room with a couch &
a television



A man riding a bike
on a beach



A man is walking down the
street with a suitcase



Messaging bots

A Computer program that can interact with one or multiple humans through the chat interface of an online messaging platform.

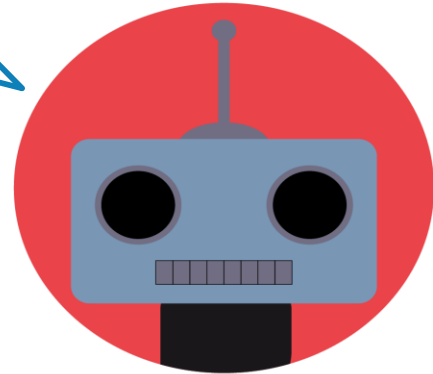



hey! I need a taxi home.

**Question
Answering**

**Question
Answering**

Sure. A taxi is on its way!





Now, let's understand
how to import text for
text mining using
Python

Reading External Files – OS Module

You can use the OS module in python to read external files based on your working directory. Let's check the code below to get the current working directory:

```
import os  
os.getcwd()
```

```
'C:\\Users\\atul'
```

To change the current working directory:

```
path = 'D:\\NLP'  
os.chdir(path)  
os.getcwd()
```

```
'D:\\NLP'
```

Reading Text(.txt) Files

In order to read a text file, we'll use the 'open' function in python and pass in the name of the text file:

```
text_file = open("textfile.txt", 'r')  
text_file.read()
```

read() to read the .txt files

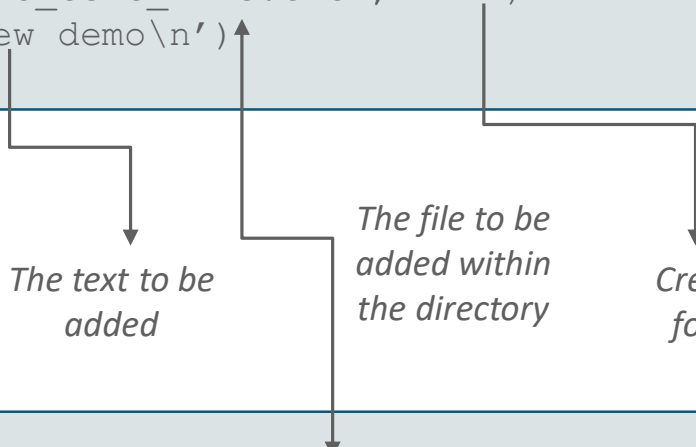
'textfile.txt' stored in current working directory

```
"My app's associated domain doesn't work on iOS 11.4 and iOS 12 beta. After some research I found that after iOS 11.4 only request."
```

Writing .txt Files

Let us consider, you want to create a new text file within the working directory and write text into the same:

```
text_file2=open("demo_text_file.txt", "w+")  
text_file2.write('New demo\n')  
text_file2.close()
```



The text to be added

The file to be added within the directory

Creates new file, if no such file is found in the working directory

```
text_file3 = open('demo_text_file.txt', 'r')  
text_file3.read()
```

```
'New demo\n'
```


Reading MS Word(.docx) Documents In Python

Let us consider, you want to import a word doc within the working directory:

```
import docx
doc=docx.Document('Demo_Word.docx')
```

Let us now check the total number of paragraphs within the working doc:

```
len(doc.paragraphs)
```

1

Accessing text within the within the doc:

```
doc.paragraphs[0].text
```

```
"My app's associated domain doesn't work on iOS 11.4 and iOS 12 beta. After some research I found that after iOS 11.4 only requires."
```

Accessing The Paragraph

To access the number of sentences in the paragraph:

```
len(doc.paragraphs[0].runs)
```

1

Getting the specific sentences within the paragraph:

```
doc.paragraphs[0].runs[0].text
```

```
"My app's associated domain doesn't work on iOS 11.4 and iOS 12 beta. After some research I found that after iOS 11.4 only require."
```

Adding A New Paragraph

Consider, you want to add a new paragraph within a word doc, "ios 11.4 is better than ios beta"

```
len(doc.paragraphs[0].runs)
```

Let's recheck the number of paragraphs in the doc:

```
len(doc.paragraphs)
```

```
2
```

Save the updated word doc:

```
doc.save('Demo_Word.docx')
```

Loading CSV Data Into Python As DataFrames

Data can be loaded into DataFrames from input data stored in the CSV format using the **read_csv()** function

```
table = pandas.read_csv("/home/edupy/Datasets/USArrests.csv")
```

Path to file

[illegible]

```
<DOCTYPE html>  
<html lang="en">  
<head>  
    <title>My perfect website</title>  
    <meta charset="utf-8">  
  
    <link rel="stylesheet" href="/css/mysql.css?v">  
    <link rel="stylesheet" href="/css/www.mysql.com/v">  
  
    <script name="script.js" content="script.js&id=6049, initial script"></script>  
  
    <script>  
        var mytag = mytag[0];  
        mytag.end = mytag.end + 1;  
        [Function]:  
            var gds = document.createElement('script');  
            gds.async = true;  
            gds.src = "/script.js";  
            var ujsurl = "https://document.location.protocol";  
            gds.src += ujsurl + "https://www.mysql.com/js/script.js";  
            var doc = document.getElementsByTagName("script");  
            node.appendChild(doc[doc.length - 1]);  
            node.appendChild(gds);  
        ]  
    </script>  
    mytag.doc.push(function() {  
        //for homepage/jqueryShippings = mytag.shipping();  
        addTitle(404, "404 Not Found");  
        addMeta(0, 0, 1000, 2000);  
        mytag.defer(mytag["http://localhost/mysql.com/js/"]([0], 250), [200, 300], "success doc V")
```

CSV File

Program Data

Program

Adding A New Column To An Existing DataFrame

A new column can be added to a DataFrame when the data is passed as a Series

```
import numpy as np
import pandas as pd
table = pd.read_csv("data.csv")
df2 = table[['id','diagnosis']]
length = len(table)
df2["new_column"]=pd.Series(np.random.randn(length), index=df2.index)
df2
```

	id	diagnosis
0	842302	M
1	842517	M
2	84300903	M
3	84348301	M

df2 initially



	id	diagnosis	new_column
0	842302	M	-0.100364
1	842517	M	-0.685319
2	84300903	M	-0.475807
3	84348301	M	-0.244345

df2 post column addition

Storing Data In CSV Files

- Data present in DataFrames can be written to a CSV file using the **to_csv()** function
- If the specified path doesn't exist, a file of the same name is automatically created

```
table.to_csv("/home/edupy/Datasets/USArrests2.csv")
```

Path to file

[illegible][illegible]

Program

Program Data

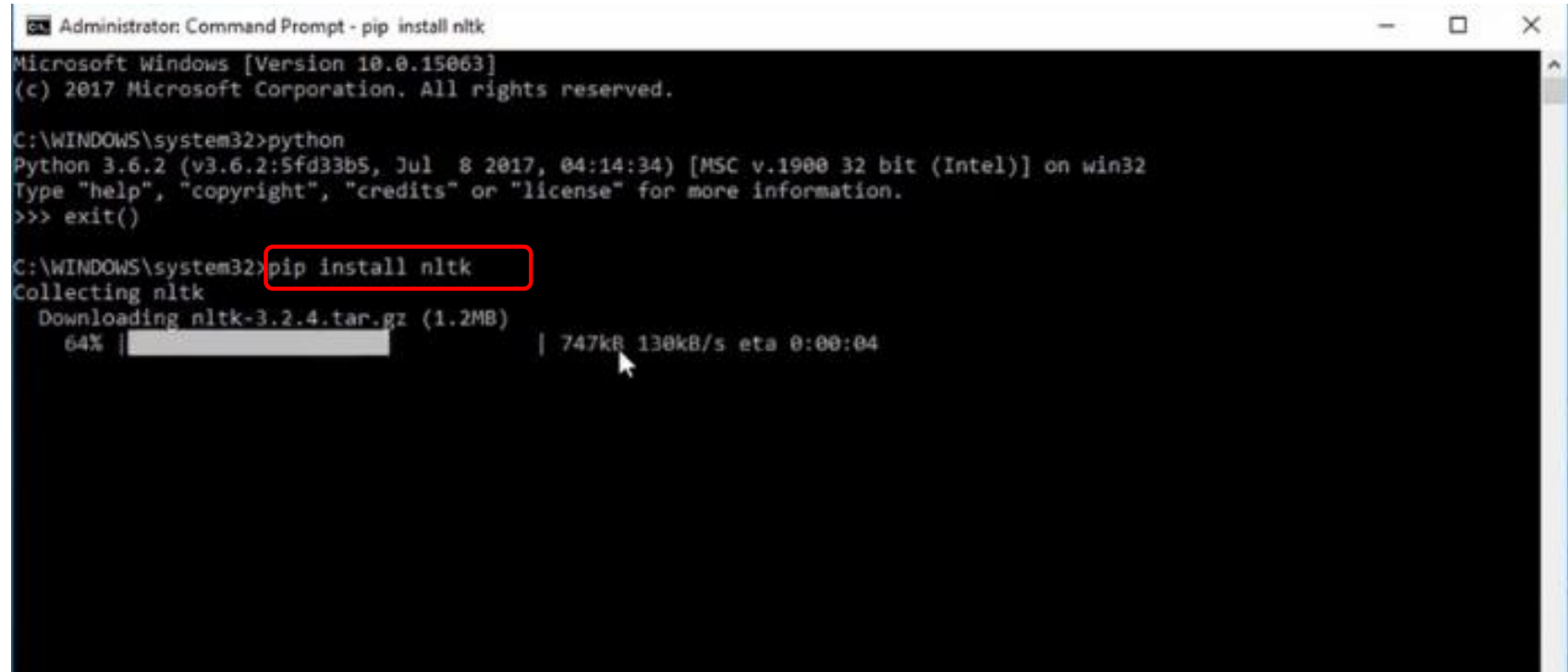
CSV File



Working On NLTK Corpora: Most Popular Library In Python For NLP

Setting The NLTK Environment

Open your cmd panel and install 'nltk' using the following command: `pip install nltk`

A screenshot of a Windows Command Prompt window titled "Administrator: Command Prompt - pip install nltk". The window shows the following text: "Microsoft Windows [Version 10.0.15063] (c) 2017 Microsoft Corporation. All rights reserved. C:\WINDOWS\system32>python Python 3.6.2 (v3.6.2:5fd33b5, Jul 8 2017, 04:14:34) [MSC v.1900 32 bit (Intel)] on win32 Type "help", "copyright", "credits" or "license" for more information. >>> exit() C:\WINDOWS\system32>pip install nltk Collecting nltk Downloading nltk-3.2.4.tar.gz (1.2MB) 64% | [progress bar] | 747kB 130kB/s eta 0:00:04". The command "pip install nltk" is highlighted with a red rectangle. The progress bar for the download is partially filled, and a mouse cursor is pointing at the download progress information.

```
Administrator: Command Prompt - pip install nltk
Microsoft Windows [Version 10.0.15063]
(c) 2017 Microsoft Corporation. All rights reserved.

C:\WINDOWS\system32>python
Python 3.6.2 (v3.6.2:5fd33b5, Jul 8 2017, 04:14:34) [MSC v.1900 32 bit (Intel)] on win32
Type "help", "copyright", "credits" or "license" for more information.
>>> exit()

C:\WINDOWS\system32>pip install nltk
Collecting nltk
  Downloading nltk-3.2.4.tar.gz (1.2MB)
    64% | [progress bar] | 747kB 130kB/s eta 0:00:04
```


Setting The NLTK Environment

Enter the python console and import the 'nltk' package:

```
Microsoft Windows [Version 10.0.15063]
(c) 2017 Microsoft Corporation. All rights reserved.

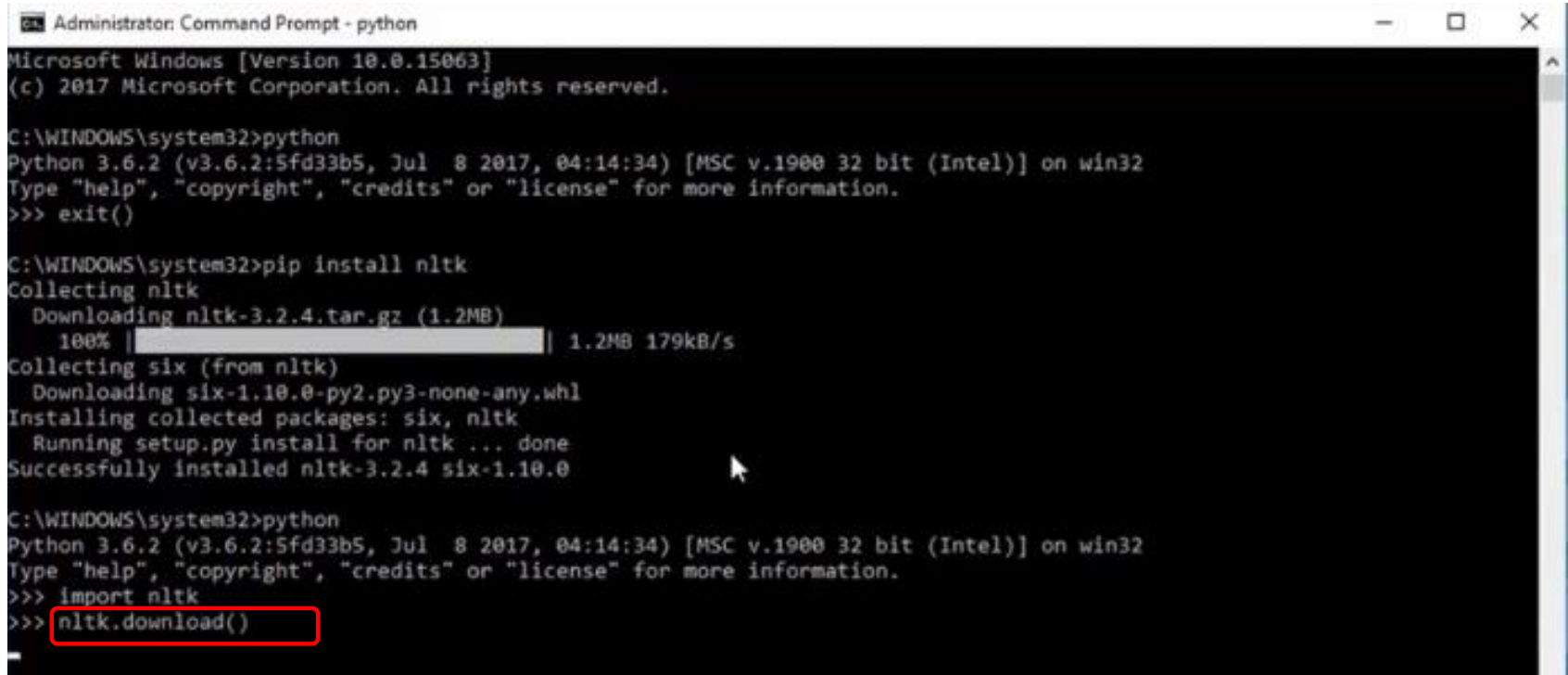
C:\WINDOWS\system32>python
Python 3.6.2 (v3.6.2:5fd33b5, Jul 8 2017, 04:14:34) [MSC v.1900 32 bit (Intel)] on win32
Type "help", "copyright", "credits" or "license" for more information.
>>> exit()

C:\WINDOWS\system32>pip install nltk
Collecting nltk
  Downloading nltk-3.2.4.tar.gz (1.2MB)
    100% |#####| 1.2MB 179kB/s
Collecting six (from nltk)
  Downloading six-1.10.0-py2.py3-none-any.whl
Installing collected packages: six, nltk
  Running setup.py install for nltk ... done
Successfully installed nltk-3.2.4 six-1.10.0

C:\WINDOWS\system32>python
Python 3.6.2 (v3.6.2:5fd33b5, Jul 8 2017, 04:14:34) [MSC v.1900 32 bit (Intel)] on win32
Type "help", "copyright", "credits" or "license" for more information.
>>> import nltk
```

Download All The Packages Using NLTK Downloader

Enter the `nltk.download()` command



The screenshot shows a Windows Command Prompt window titled "Administrator: Command Prompt - python". The window displays the following commands and output:

```
Microsoft Windows [Version 10.0.15063]
(c) 2017 Microsoft Corporation. All rights reserved.

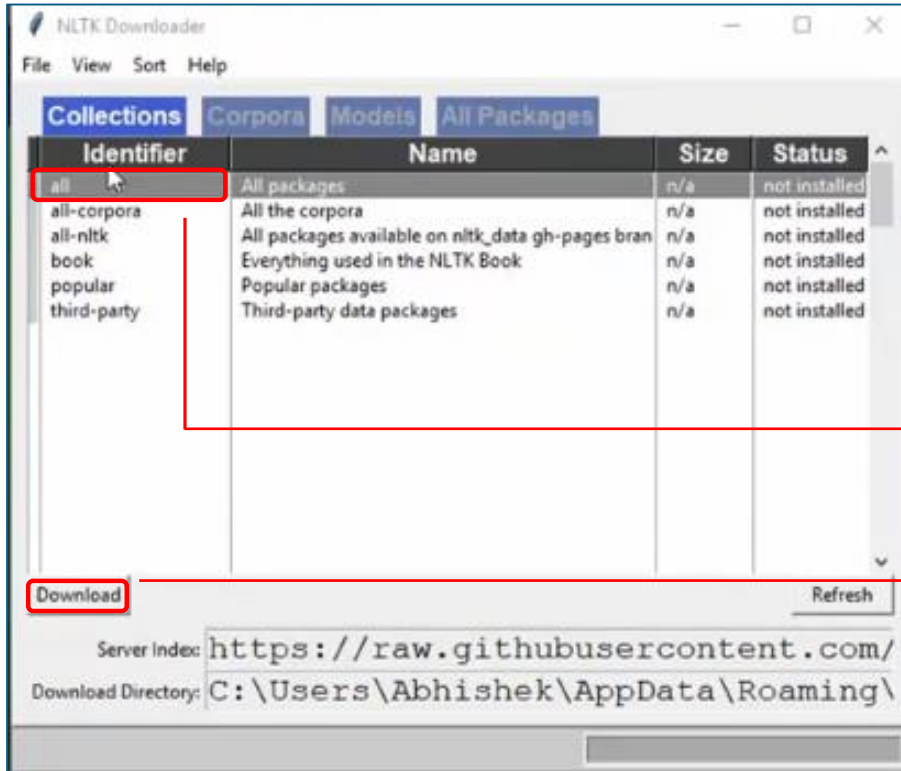
C:\WINDOWS\system32>python
Python 3.6.2 (v3.6.2:5fd33b5, Jul 8 2017, 04:14:34) [MSC v.1900 32 bit (Intel)] on win32
Type "help", "copyright", "credits" or "license" for more information.
>>> exit()

C:\WINDOWS\system32>pip install nltk
Collecting nltk
  Downloading nltk-3.2.4.tar.gz (1.2MB)
    100% |#####| 1.2MB 179kB/s
Collecting six (from nltk)
  Downloading six-1.10.0-py2.py3-none-any.whl
Installing collected packages: six, nltk
Running setup.py install for nltk ... done
Successfully installed nltk-3.2.4 six-1.10.0

C:\WINDOWS\system32>python
Python 3.6.2 (v3.6.2:5fd33b5, Jul 8 2017, 04:14:34) [MSC v.1900 32 bit (Intel)] on win32
Type "help", "copyright", "credits" or "license" for more information.
>>> import nltk
>>> nltk.download()
```

The `nltk.download()` command is highlighted with a red rectangle in the original image.

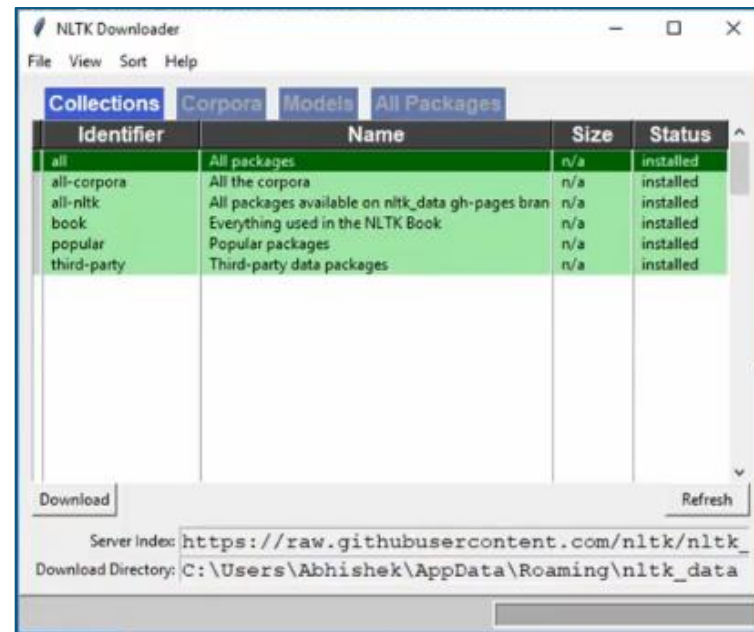
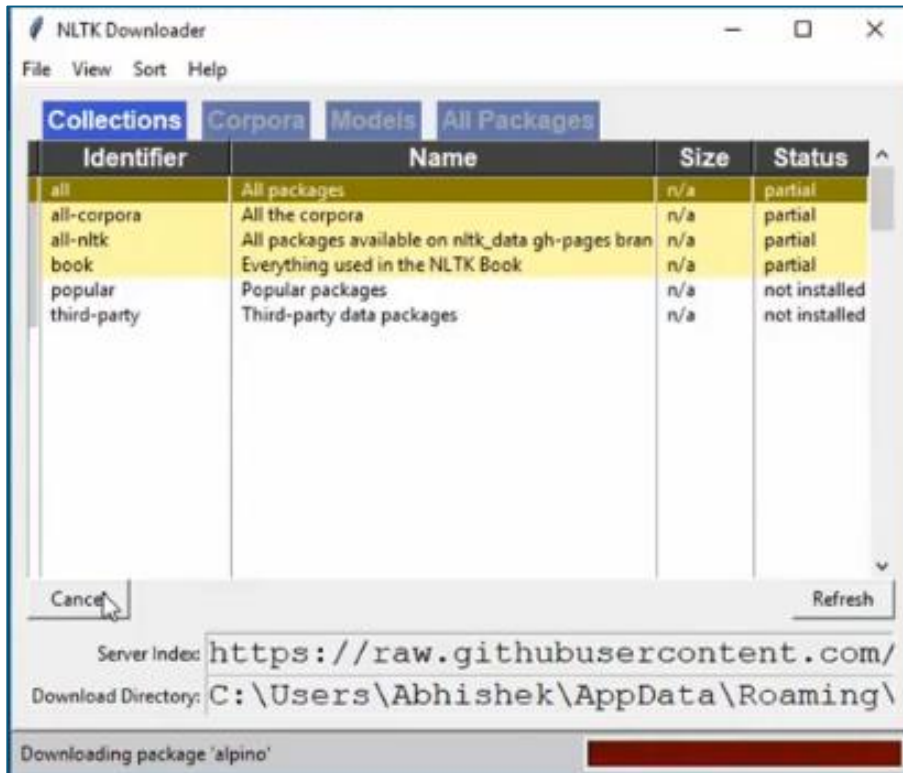
Download All The Packages Using NLTK Downloader



1 Click the 'all' tab within the downloader

2 Click the 'download' tab

Download All The Packages Using NLTK Downloader



The download will start.....

NLTK Corpora

- **Corpora:**

- Body of text containing collection of similar kind of documents
- Plural is 'corpora'

- **Document:**

- Newspaper article, novel, patent, scientific paper
- Blog post, comment, status update, tweet



Reading The NLTK Corpora

Reading NLTK Corpora – Example 1

You can read the NLTK corpora data(stored within your local) with the following code:

```
import os
import nltk
import nltk.corpus
print(os.listdir(nltk.data.find("corpora")))
```

```
['abc', 'abc.zip', 'alpino', 'alpino.zip', 'biocreative_ppi', 'biocreative_ppi.zip', 'brown', 'brown.zip', 'brown_tei', 'brown_tei.zip', 'cess_cat', 'cess_cat.zip', 'cess_esp', 'cess_esp.zip', 'chat80', 'chat80.zip', 'city_database', 'city_database.zip', 'cmudict', 'cmudict.zip', 'comparative_sentences', 'comparative_sentences.zip', 'comtrans.zip', 'conll2000', 'conll2000.zip', 'conll2002', 'conll2002.zip', 'conll2007.zip', 'crubadan', 'crubadan.zip', 'dependency_treebank', 'dependency_treebank.zip', 'dolch', 'dolch.zip', 'europarl_raw', 'europarl_raw.zip', 'floresta', 'floresta.zip', 'framenet_v15', 'framenet_v15.zip', 'framenet_v17', 'framenet_v17.zip', 'gazetteers', 'gazetteers.zip', 'genesis', 'genesis.zip', 'gutenberg', 'gutenberg.zip', 'ieer', 'ieer.zip', 'inaugural', 'inaugural.zip', 'indian', 'indian.zip', 'jeita.zip', 'kimmo', 'kimmo.zip', 'knbc.zip', 'lin_thesaurus', 'lin_thesaurus.zip', 'machado.zip', 'mac_morpho', 'mac_morpho.zip', 'masc_tagged.zip', 'movie_reviews', 'movie_reviews.zip', 'mte_teip5', 'mte_teip5.zip', 'names', 'names.zip', 'nombank.1.0.zip', 'nonbreaking_prefixes', 'nonbreaking_prefixes.zip', 'nps_chat', 'nps_chat.zip', 'omw', 'omw.zip', 'opinion_lexicon', 'opinion_lexicon.zip', 'panlex_swadesh.zip', 'paradigms', 'paradigms.zip', 'pil', 'pil.zip', 'pl196x', 'pl196x.zip', 'ppattach', 'ppattach.zip', 'problem_reports', 'problem_reports.zip', 'product_reviews_1', 'product_reviews_1.zip', 'product_reviews_2', 'product_reviews_2.zip', 'proppbank.zip', 'pros_cons', 'pros_cons.zip', 'ptb', 'ptb.zip', 'qc', 'qc.zip', 'reuters.zip', 'rte', 'rte.zip', 'semcor.zip', 'senseval', 'senseval.zip', 'sentence_polarity', 'sentence_polarity.zip', 'sentiwordnet', 'sentiwordnet.zip', 'shakespeare', 'shakespeare.zip', 'sinica_treebank', 'sinica_treebank.zip', 'smultron', 'smultron.zip', 'state_union', 'state_union.zip', 'stopwords', 'stopwords.zip', 'subjectivity', 'subjectivity.zip', 'swadesh', 'swadesh.zip', 'switchboard', 'switchboard.zip', 'timit', 'timit.zip', 'toolbox', 'toolbox.zip', 'treebank', 'treebank.zip', 'twitter_samples', 'twitter_samples.zip', 'udhr', 'udhr.zip', 'udhr2', 'udhr2.zip', 'unicode_samples', 'unicode_samples.zip', 'universal_treebanks_v20.zip', 'verbnet', 'verbnet.zip', 'webtext', 'webtext.zip', 'wordnet', 'wordnet.zip', 'wordnet_ic', 'wordnet_ic.zip', 'words', 'words.zip', 'ycoe', 'ycoe.zip']
```

Collection of documents

Accessing A Document

Let's now try accessing a document from the corpora:

```
nltk.corpus.gutenberg.fileids()
```

```
['austen-emma.txt',  
'austen-persuasion.txt',  
'austen-sense.txt',  
'bible-kjv.txt',  
'blake-poems.txt',  
'bryant-stories.txt',  
'burgess-busterbrown.txt',  
'carroll-alice.txt',  
'chesterton-ball.txt',  
'chesterton-brown.txt',  
'chesterton-thursday.txt',  
'edgeworth-parents.txt',  
'melville-moby_dick.txt',  
'milton-paradise.txt',  
'shakespeare-caesar.txt',  
'shakespeare-hamlet.txt',  
'shakespeare-macbeth.txt',  
'whitman-leaves.txt']
```

Here, we are accessing the gutenberg() document

List of files within the 'gutenberg' doc

Let's try accessing the file within the doc

Accessing The File

```
hamlet=nltk.corpus.gutenberg.words('shakespeare-hamlet.txt')
hamlet
```

```
['[', 'The', 'Tragedie', 'of', 'Hamlet', 'by', ...]
```



Note: The above list contains individual words segregated by commas. For a better clarity, you can convert the above set of words in the form of paragraph

```
for word in hamlet[:500]:
    print(word, sep=' ', end=' ')
```

```
[ The Tragedie of Hamlet by William Shakespeare 1599 ] Actus Primus . Scoena Prima . Enter Barnardo and Francisco two Centinels
. Barnardo . Who ' s there ? Fran . Nay answer me : Stand & vnfold your selfe Bar . Long liue the King Fran . Barnardo ? Bar .
He Fran . You come most carefully vpon your houre Bar . ' Tis now strook twelue , get thee to bed Francisco Fran . For this rel
eefe much thanks : ' Tis bitter cold , And I am sicke at heart Barn . Haue you had quiet Guard ? Fran . Not a Mouse stirring B
arn . Well , goodnight . If you do meet Horatio and Marcellus , the Riuals of my Watch , bid them make hast . Enter Horatio and
Marcellus . Fran . I thinke I heare them . Stand : who ' s there ? Hor . Friends to this ground Mar . And Leige - men to the Da
```

Reading NLTK Corpora – Example 2

Let's check another corpora containing movie reviews

```
from nltk.corpus import movie_reviews
print(movie_reviews.categories())
```

['neg', 'pos']



A trained corpus used to train a classifier using Machine Learning

Let's check the list of files under each category

```
print(len(movie_reviews.fileids('pos')))  
print(" ")  
print(movie_reviews.fileids('pos'))
```

1000

```
['pos/cv000_29590.txt', 'pos/cv001_18431.txt', 'pos/cv002_15918.txt', 'pos/cv003_11664.txt', 'pos/cv004_11636.txt', 'pos/cv005_29443.txt', 'pos/cv006_15448.txt', 'pos/cv007_4968.txt', 'pos/cv008_29435.txt', 'pos/cv009_29592.txt', 'pos/cv010_29198.txt', 'pos/cv011_12166.txt', 'pos/cv012_29576.txt', 'pos/cv013_10159.txt', 'pos/cv014_13924.txt', 'pos/cv015_29439.txt', 'pos/cv016_4659.txt', 'pos/cv017_22464.txt', 'pos/cv018_20137.txt', 'pos/cv019_14482.txt', 'pos/cv020_8825.txt', 'pos/cv021_15838.txt', 'pos/cv022_12864.txt', 'pos/cv023_12672.txt', 'pos/cv024_6778.txt', 'pos/cv025_3108.txt', 'pos/cv026_29325.txt', 'pos/cv027_25219.txt', 'pos/cv028_26746.txt', 'pos/cv029_18643.txt', 'pos/cv030_21593.txt', 'pos/cv031_18452.txt', 'pos/cv032_22550.txt']
```

Reading NLTK Corpora – Example 2

Let's check words in one of the files:

```
file = nltk.corpus.movie_reviews.words('pos/cv000_29590.txt')  
file
```

```
['films', 'adapted', 'from', 'comic', 'books', 'have', ...]
```



Convert the above set of words in the form of paragraph

```
for word in file:  
    print(word, sep=' ', end=' ')
```

```
films adapted from comic books have had plenty of success , whether they ' re about superheroes ( batman , superman , spawn ) ,  
or geared toward kids ( casper ) or the arthouse crowd ( ghost world ) , but there ' s never really been a comic book like from  
hell before . for starters , it was created by alan moore ( and eddie campbell ) , who brought the medium to a whole new level  
in the mid ' 80s with a 12 - part series called the watchmen . to say moore and campbell thoroughly researched the subject of j  
ack the ripper would be like saying michael jackson is starting to look a little odd . the book ( or " graphic novel , " if you
```

Summary

- Need of Text Mining
- NLTK Library
- Applications of NLP
- Text Reading Procedures



Questions



FEEDBACK



An illustration on a solid blue background. Two hands, depicted in a stylized manner with orange skin and blue sleeves, are holding a light blue rectangular banner. The banner is held taut between the hands, which are positioned at the bottom corners of the banner. The words "THANK YOU" are written in large, white, bold, sans-serif capital letters on the banner.

**THANK
YOU**

For more information please visit our website
www.edureka.co