

NLP • 16 Min Read • Jul 24, 2020

Natural Language Processing (NLP) simplified : A step-by-step guide

By Dibyendu Banerjee

Highlights

The field of study that focuses on the interactions between human language and computers is called Natural Language Processing or NLP for short. It sits at the intersection of computer science, artificial intelligence, and computational linguistics.

Quick introduction – What is NLP?

The field of study that focuses on the interactions between human language and computers is called Natural Language Processing or NLP for short. It sits at the intersection of computer science, artificial intelligence, and computational linguistics (Wikipedia).

NLP is Artificial Intelligence or Machine Learning or a Deep Learning?

The answer is here. The question itself is not fully correct! Sometime people incorrectly use the terms AI, ML and DL. Why not we simplify those first and then come back.



Resources

Ecosystem

Explore

Government

Sectors

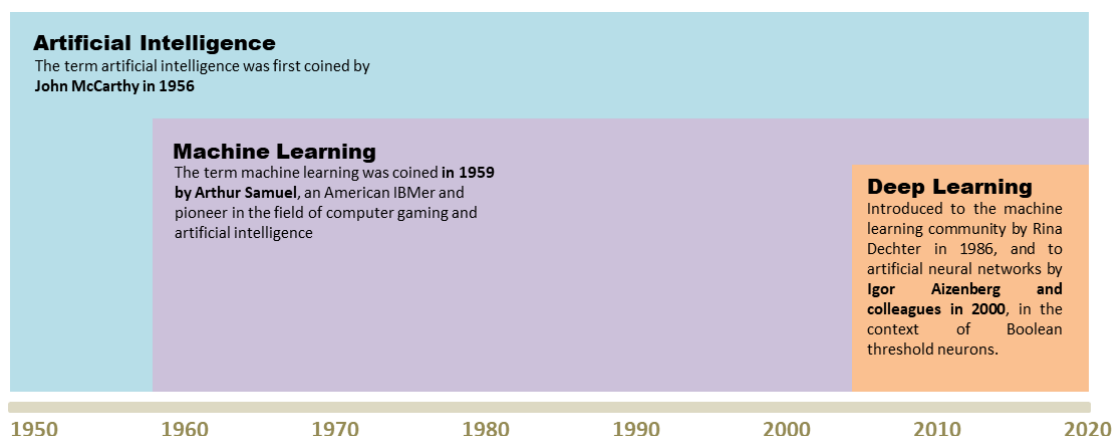
of India

to describe human thinking as a symbolic system. But the field of AI wasn't

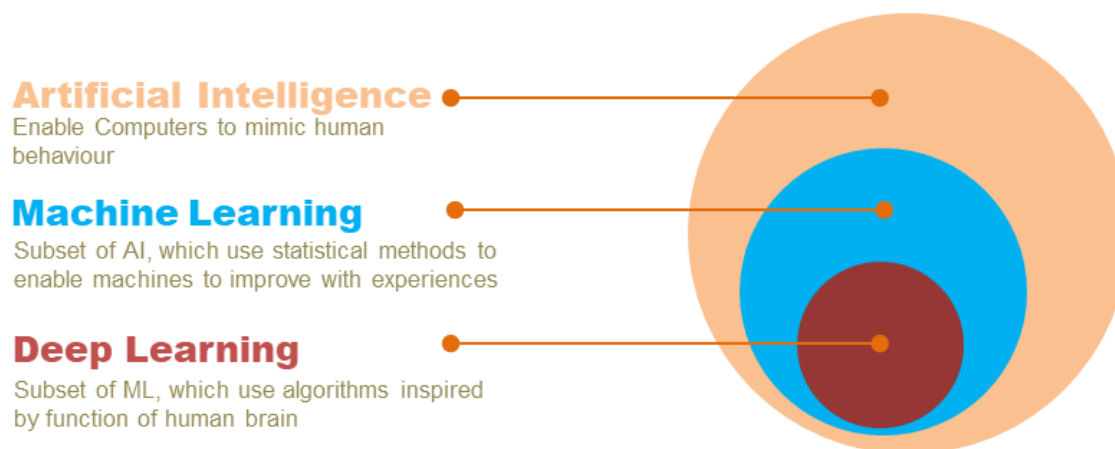
Lea

formally founded until 1956, at a conference at Dartmouth College, in Hanover, New Hampshire, where the term “artificial intelligence” was coined.

Timeline view about when these jargons were first introduced...



Now, let us take a look what exactly AI, ML and Deep Learning is, in a very concise way. The relationship of AL, ML and DL can be treated as below.



NLP: How Does NLP Fit into the AI World?

With basic understanding of Artificial Intelligence, Machine Learning and Deep Learning, let's revisit our very first query NLP is Artificial Intelligence or Machine Learning or a Deep Learning?

The words AI, NLP, and ML (machine learning) are sometimes used almost interchangeably. However, there is an order to the madness of their relationship.



A MEITY, NEGD & NASSCOM INITIATIVE

Resources

Ecosystem

Explore

Government

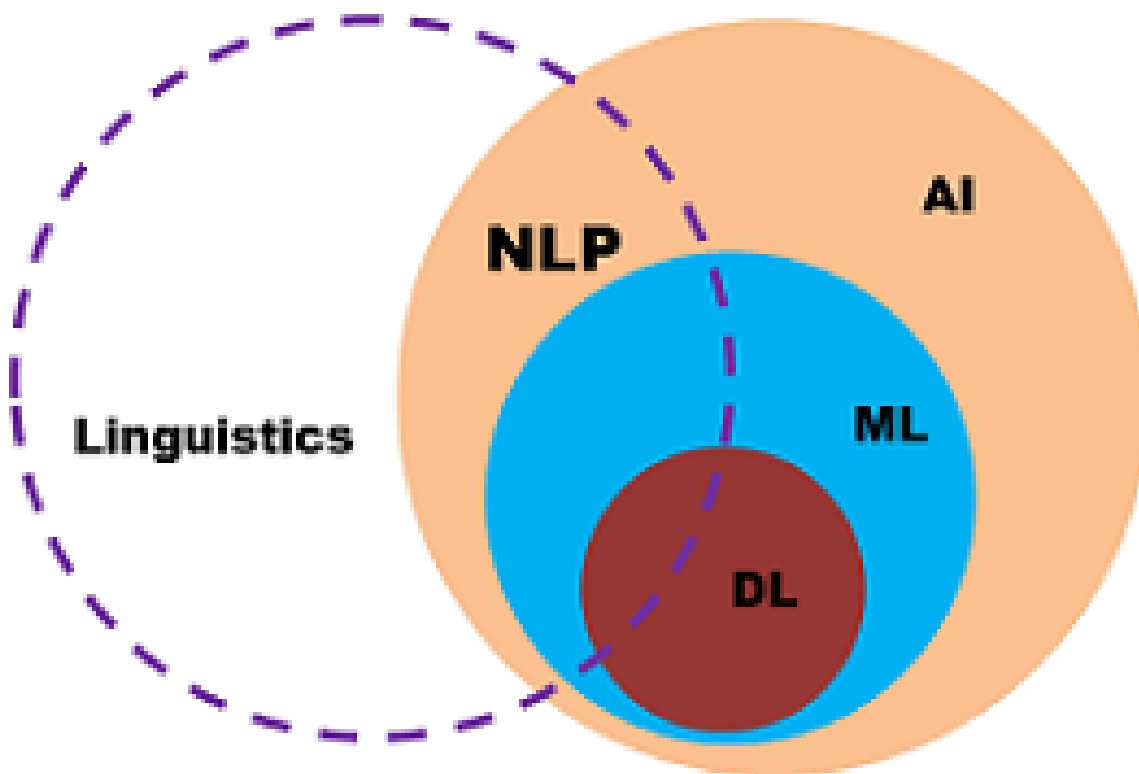
Linguistics so that computers and humans can talk seamlessly.

Sectors

of India



NLP endeavours to bridge the divide between machines and people by enabling a computer to analyse what a user said (input speech recognition) and process what the user meant. This task has proven quite complex.



To converse with humans, a program must understand syntax (grammar), semantics (word meaning), and morphology (tense), pragmatics (conversation). The number of rules to track can seem overwhelming and explains why earlier attempts at NLP initially led to disappointing results.

With a different system in place, NLP slowly improved moving from a cumbersome-rule based to a pattern learning based computer programming methodology. Siri appeared on the iPhone in 2011. In 2012, the new discovery of use of graphical processing units (GPU) improved digital neural networks and NLP.

NLP empowers computer programs to comprehend unstructured content by utilizing AI and machine learning to make derivations and give context to language, similarly as human brains do. It is a device for revealing and analysing the “signals” covered in unstructured information. Organizations would then be able to get a deeper comprehension of public perception around their products, services and brand, just as those of their rivals.

Now Google has released its own neural-net-based engine for eight language pairs,

Lea



Resources

Ecosystem

Explore

Governmer

Sectors

of India

impressive advances in fields such as image recognition and speech processing.

Their application to Natural Language Processing (NLP) was less impressive at first, but has now proven to make significant contributions, yielding state-of-the-art results for some common NLP tasks. Named entity recognition (NER), part of speech (POS) tagging or sentiment analysis are some of the problems where neural network models have outperformed traditional approaches. The progress in machine translation is perhaps the most remarkable among all.

NLP: Game changers in our daily life, examples for Businesses

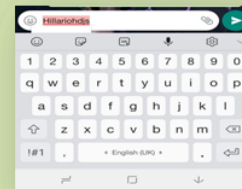


Smart Assistant

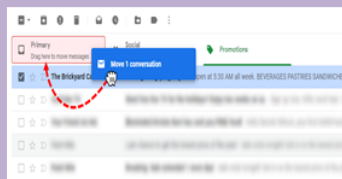
We've become used to the fact that we can say "Hey Siri," ask a question, and she understands what we said and responds with relevant answers based on context. And we're getting used to seeing Siri or Alexa pop up throughout our home and daily life as we have conversations with them through items like the thermostat, light switches, car, and more.

We now expect assistants like Alexa and Siri to understand contextual clues as they improve our lives and make certain activities easier like ordering items, and even appreciate when they respond humorously or answer questions about themselves. Our interactions will grow more personal as these assistants get to know more about us.

Things like autocorrect, autocomplete, and predictive text are so commonplace on our smartphones that we take them for granted. Autocomplete and predictive text are similar to search engines in that they predict things to say based on what you type, finishing the word or suggesting a relevant one. And autocorrect will sometimes even change words so that the overall message makes more sense.



Predictive text



Email Filters

One of the more prevalent, newer applications of NLP is found in Gmail's email classification. The system recognizes if emails belong in one of three categories (primary, social, or promotions) based on their contents. For all Gmail users, this keeps your inbox to a manageable size with important, relevant emails you wish to review and respond to quickly.

NLP is not Just About Creating Intelligent bots.



A MEITY, NEGD & NASSCOM INITIATIVE

Lea

Resources

Ecosystem

**Explore
Sectors**

**Governments
of India**

So, by using NLP, developers can organize and structure the mass of unstructured data to perform tasks such as intelligent:

Below are some of the widely used areas of NLPs.



Automatic Summarization

Intelligently shortening long pieces of text



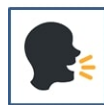
Named entity recognition

Locate and classify named entities pre-defined categories such as the organizations; person names; locations etc.



Sentiment analysis

To identify, for instance, positive, negative and neutral opinion from text or speech widely used to gain insights from social media comments, forums or survey responses



Speech recognition

Enables computers to recognize and transform spoken language into text - dictation - and, if programmed, act upon that recognition - e.g. in case of assistants like Google Assistant Cortana or Apple's Siri



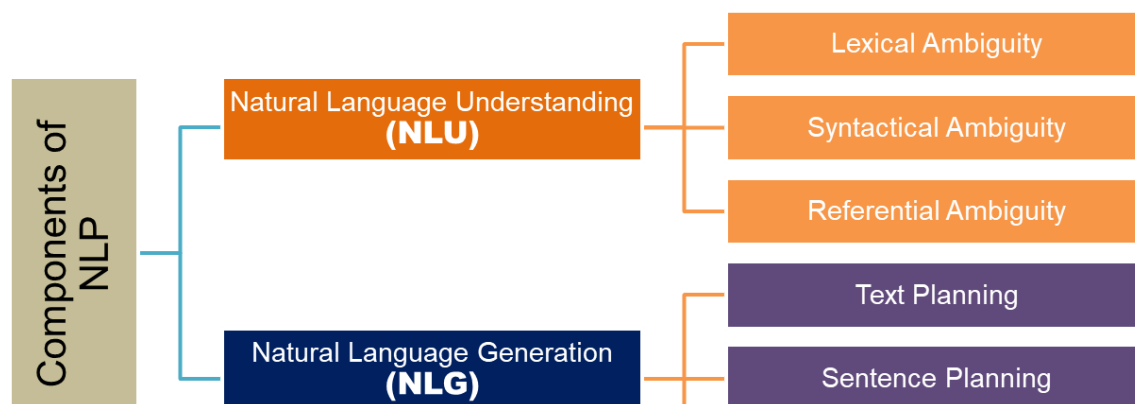
Topic segmentation

Automatically divides written texts, speech or recordings into shorter, topically coherent segments and is used in improving information retrieval or speech recognition

Components of NLP

NLP can be divided into two basic components.

- Natural Language Understanding
- Natural Language Generation



Lea



A MEITY, NEGD & NASSCOM INITIATIVE

Resources

Ecosystem

Explore
Sectors

Governmer
of India

NLU is naturally harder than NLG tasks. Really? Let's see what are all challenges faced by a machine while understanding.

There are lot of ambiguity while learning or trying to interpret a language.

He is looking for a **match**

What do you understand by 'match'?

Partner

Or Cricket/Football Match

Lexical Ambiguity can occur when a word carries different sense, i.e. having more than one meaning and the sentence in which it is contained can be interpreted differently depending on its correct sense. Lexical ambiguity can be resolved to some extent using parts-of-speech tagging techniques.

The chicken **is ready** to eat.

Is the chicken ready to eat his food or the chicken is ready for someone else to it? You never know.

Syntactical Ambiguity means when we see more than one meaning in a sequence of words. It is also termed as grammatical ambiguity.

Feluda met Topse and Jotayu. **They** went to restaurant

They refer to Topse and Jotayu or all?

Referential Ambiguity. Very often a text mentions as entity (something/someone), and then refers to it again, possibly in a different sentence, using another word. Pronoun causing ambiguity when it is not clear which noun it is referring to

Natural Language Generation (NLG)

This is the process of generating meaningful sentences in the form of

Lea



सत्यमेव जयते



A MEITY, NEGD & NASSCOM INITIATIVE

Resources

Ecosystem

Explore

Governmer

Sectors

of India

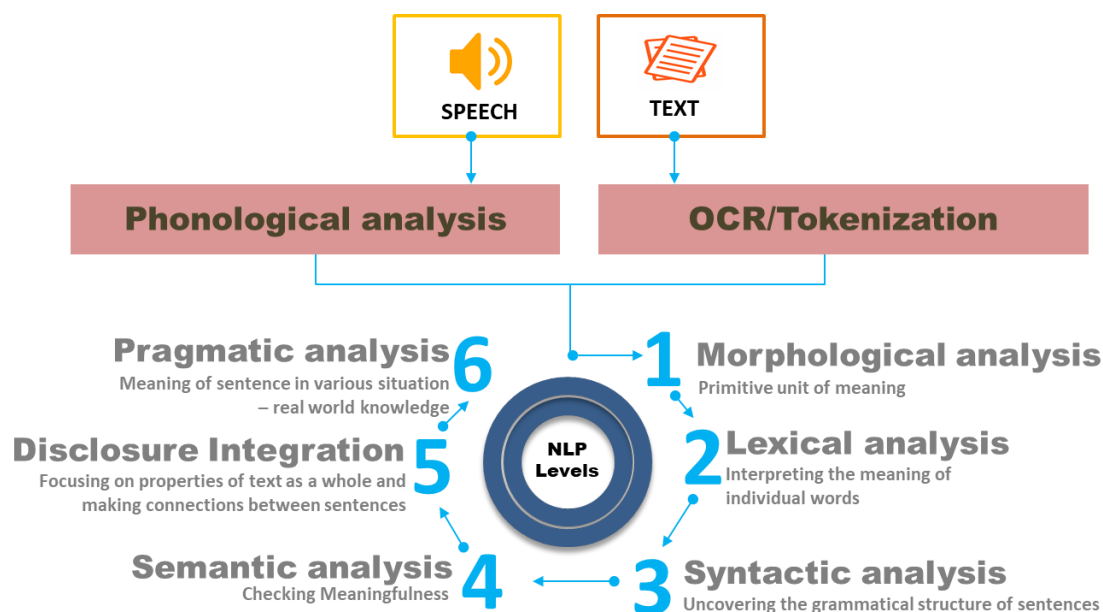
knowledge base.

- Sentence planning - It includes choosing required words, forming meaningful phrases, setting tone of the sentence.
- Text Realization - It is mapping sentence plan into sentence structure.

Levels of NLP

In the previous sections we have discussed different problem associated to NLP. Now let us see what are all typical steps involved while performing NLP tasks. We should keep in mind that the below section describes some standard workflow, it may however differ drastically as we do real life implementations basis on our problem statement or requirements.

The source of Natural Language could be speech (sound) or Text.



Phonological Analysis: This level is applied only if the text origin is a speech. It deals with the interpretation of speech sounds within and across words. Speech sound might give a big hint about the meaning of a word or a sentence.

It is study of organizing sound systematically. This requires a broad discussion and is out of scope of our current note.

Morphological Analysis: Deals with understanding distinct words according to their morphemes (the smallest units of meanings) . Taking, for example, the word: “unhappiness ” It can be broken down into three morphemes (prefix, stem, and

Lea



A MEITY, NEGD & NASSCOM INITIATIVE

Resources

Ecosystem

Explore
Sectors

Governments
of India

Lexical Analysis: It involves identifying and analysing the structure of words. Lexicon of a language means the collection of words and phrases in a language. Lexical analysis is dividing the whole chunk of text into paragraphs, sentences, and words. In order to deal with lexical analysis, we often need to perform **Lexicon Normalization**.

The most common lexicon normalization practices are Stemming:

- **Stemming:** Stemming is a rudimentary rule-based process of stripping the suffixes ("ing", "ly", "es", "s" etc) from a word.
- **Lemmatization:** Lemmatization, on the other hand, is an organized & step by step procedure of obtaining the root form of the word, it makes use of vocabulary (dictionary importance of words) and morphological analysis (word structure and grammar relations).

Syntactic Analysis: Deals with analysing the words of a sentence so as to uncover the grammatical structure of the sentence. E.g.. "Colourless green idea." This would be rejected by the Symantec analysis as colourless here; green doesn't make any sense.

Syntactical parsing involves the analysis of words in the sentence for grammar and their arrangement in a manner that shows the relationships among the words. Dependency Grammar and Part of Speech tags are the important attributes of text syntactics.

Semantic Analysis: Determines the possible meanings of a sentence by focusing on the interactions among word-level meanings in the sentence. Some people may think it's the level which determines the meaning, but actually all the level do. The semantic analyser disregards sentence such as "hot ice-cream".

Discourse Integration: Focuses on the properties of the text as a whole that convey meaning by making connections between component sentences. It means a sense of the context. The meaning of any single sentence which depends upon that sentences. It also considers the meaning of the following sentence. For example, the word "that" in the sentence "He wanted that" depends upon the prior discourse context.

Pragmatic Analysis: Explains how extra meaning is read into texts without actually being encoded in them. This requires much world knowledge, including the understanding of intentions, plans, and goals. Consider the following two sentences:



Lea

Resources

Ecosystem





Explore
SectorsGovernmer
of India

The meaning of "they" in the 2 sentences is different. In order to figure out the difference, world knowledge in knowledge bases and inference modules should be utilized.

Pragmatic analysis helps users to discover this intended effect by applying a set of rules that characterize cooperative dialogues. E.g., "close the window?" should be interpreted as a request instead of an order.

Widely used NLP Libraries

There are many libraries, packages, tools available in market. Each of them has its own pros and cons. As a market trend Python is the language which has most compatible libraries. Below table will give a summarised view of features of some of the widely used libraries. Most of them provide the basic NLP features which we discussed earlier. Each NLP libraries were built with certain objectives, hence it is quite obvious that a single library might not provide solutions for everything, it is the developer who need to use those and that is where experience and knowledge matters when and where to use what.

Tools	Features
 NLTK	<ul style="list-style-type: none"> ▪ The most well-known and full NLP library ▪ Plenty of approaches to each NLP task ▪ Supports large number of languages ▪ No integrated Word Vectors
 spaCy	<ul style="list-style-type: none"> ▪ Fastest NLP framework ▪ Easy to learn as it has one single highly optimized tool for each task ▪ Supports neural networks for training some models ▪ Lesser Language support
 leap NLP toolkit	<ul style="list-style-type: none"> ▪ Most effective for Machine Learning implementation ▪ Good documentation available ▪ No neural network support for text processing
 gensim	<ul style="list-style-type: none"> ▪ Works with large datasets and processes data streams ▪ Supports Deep Learning ▪ Designed primarily of unsupervised text modeling

NLP Hands on Using Python NLTK (Simple Examples)

NLTK is a leading platform for building Python programs to work with human language data. It provides easy-to-use interfaces to over 50 corpora and lexical



Lea

Resources

Ecosystem

Explore
SectorsGovernmer
of India

NLTK comes with many corpora, toy grammars, trained models, etc. A complete list is posted at: http://nltk.org/nltk_data/.

Before we start doing experiments on some of the techniques which are widely used during Natural Language Processing task, let's first get hands on into the installation.

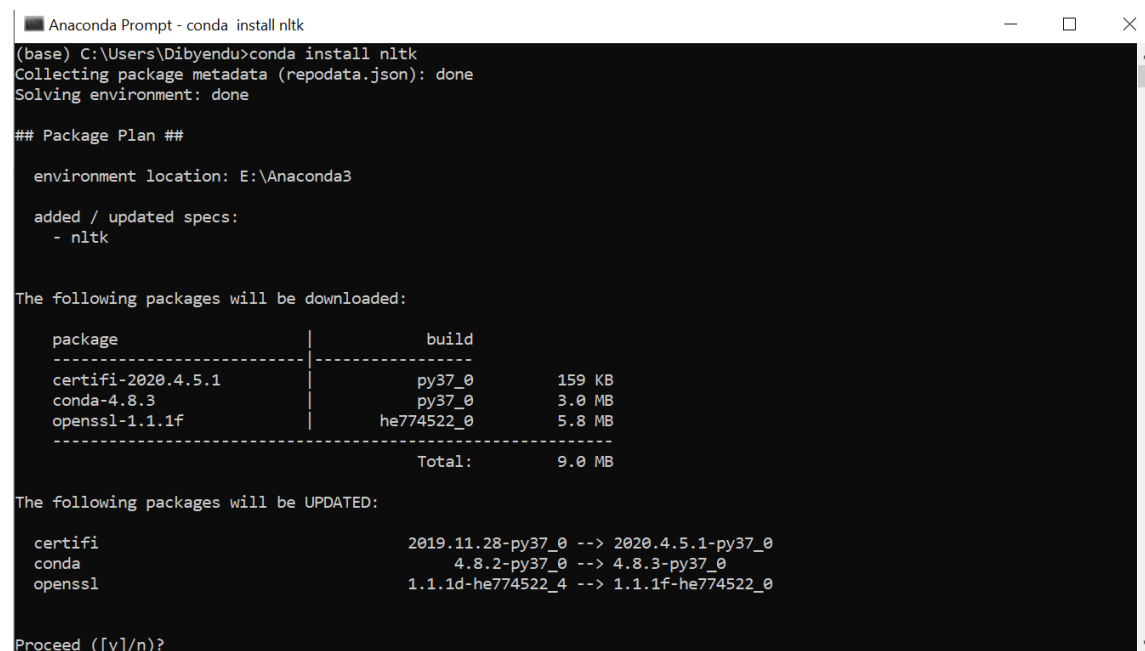
NLTK Installation

If you are using Windows or Linux or Mac, you can install NLTK using pip:

```
$ pip install nltk
```

Optionally you can also use Anaconda prompt.

```
$ conda install nltk
```



```

Anaconda Prompt - conda install nltk
(base) C:\Users\Dibyendu>conda install nltk
Collecting package metadata (repodata.json): done
Solving environment: done

## Package Plan ##

  environment location: E:\Anaconda3

  added / updated specs:
    - nltk

The following packages will be downloaded:

package | build | size
-----|-----|-----
certifi-2020.4.5.1 | py37_0 | 159 KB
conda-4.8.3 | py37_0 | 3.0 MB
openssl-1.1.1f | he774522_0 | 5.8 MB
-----|-----|-----
Total: | | 9.0 MB

The following packages will be UPDATED:

certifi 2019.11.28-py37_0 --> 2020.4.5.1-py37_0
conda 4.8.2-py37_0 --> 4.8.3-py37_0
openssl 1.1.1d-he774522_4 --> 1.1.1f-he774522_0

Proceed ([y]/n)?

```

If everything goes fine, that means you've successfully installed NLTK library. Once you've installed NLTK, you should install the NLTK packages by running the following code:

Open your Jupyter Notebook and run the below commands.



Resources

Ecosystem

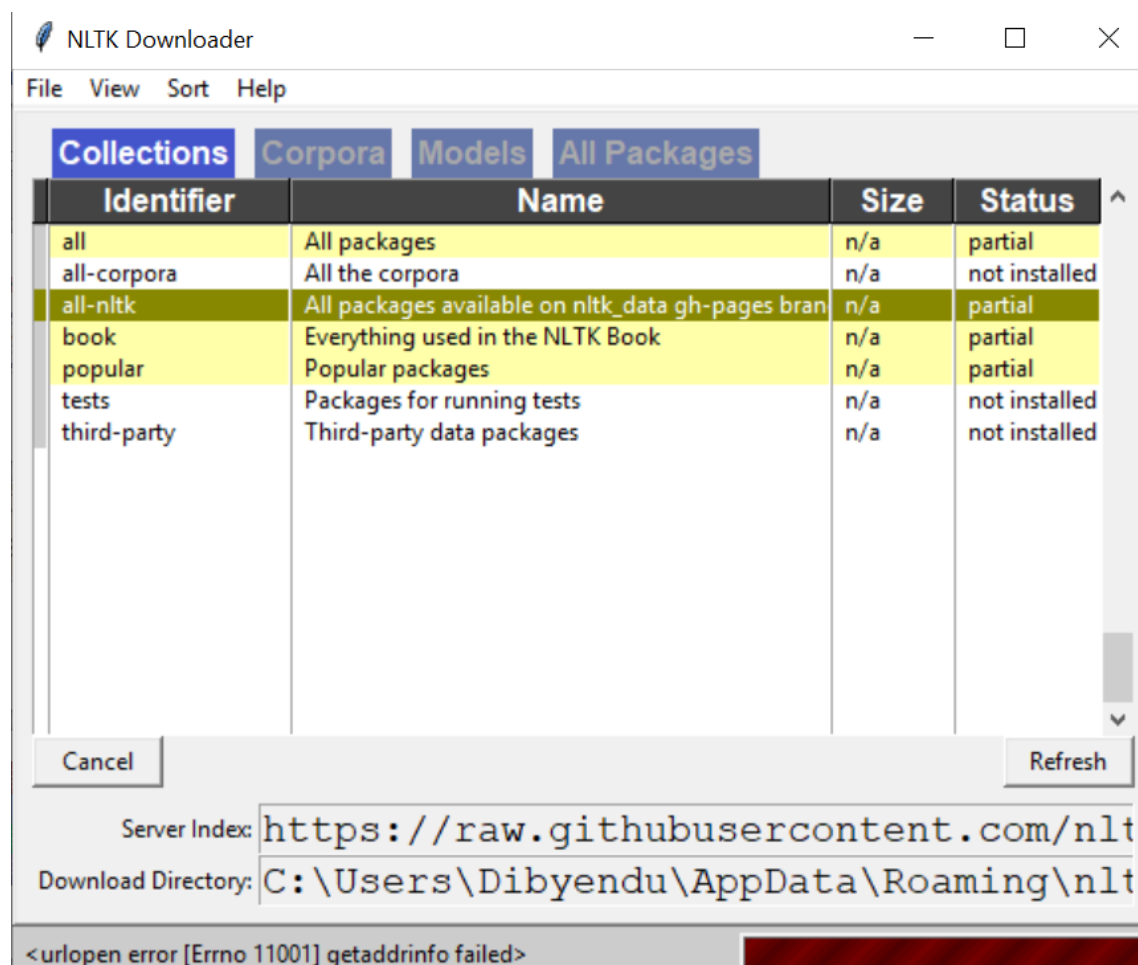
Explore
Sectors

Governor
of India

Lea

In []:

This will show the NLTK downloader to choose what packages need to be installed. You can install all packages since they have small sizes, so no problem. Now let's start the show.



सत्यमेव जयते



A MEITY, NEGD & NASSCOM INITIATIVE

Resources



Ecosystem

Explore
SectorsGovernance
of India

Lea