LAKIREDDY BALI REDDY COLLEGE OF ENGINEERING

(AUTONOMOUS)
Accredited by NAAC with 'A' Grade & NBA (Under Tier - I),
An ISO 21001:2018,14001:2015,50001:2018 Certified Institution
Approved by AICTE, New Delhi and Affiliated to JNTUK, Kakinada
L.B. REDDY NAGAR, MYLAVARAM, NTR DIST., A.P.-521 230.
http://lbrce.ac.in/it/index.php, hodit@lbrce.ac.in , Phone: 08659-222933, Fax: 08659-222931

DEPARTMENT OF INFORMATION TECHNOLOGY

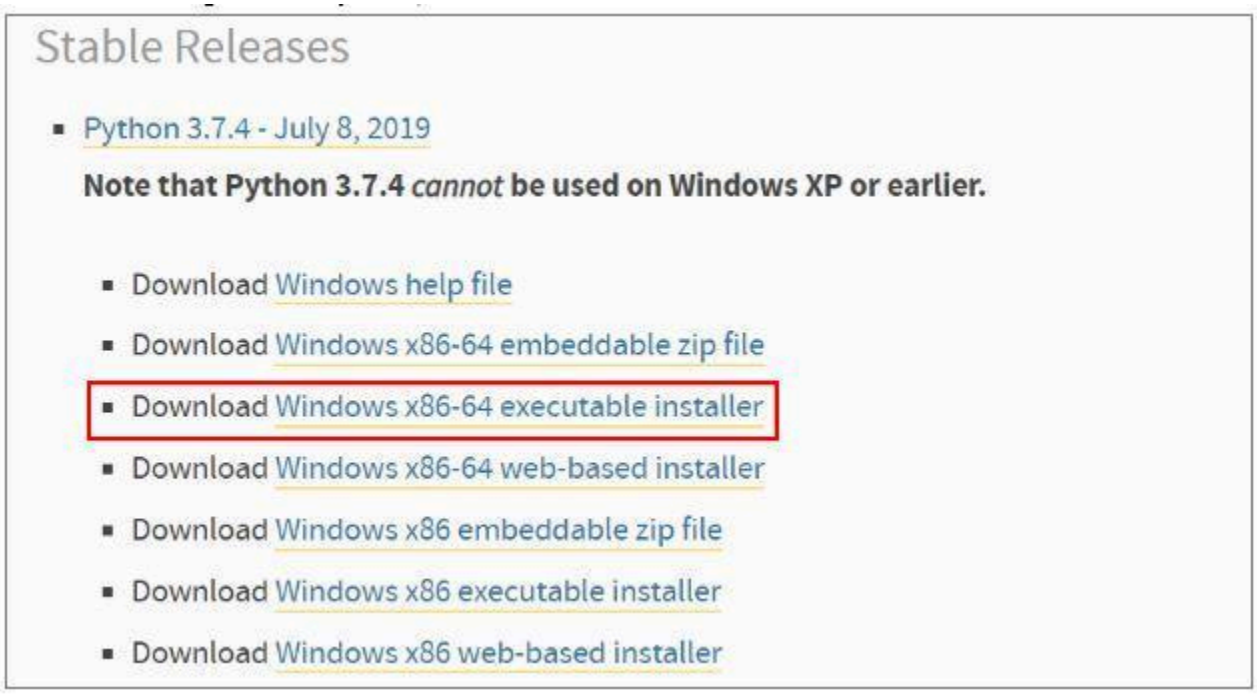
EXPERIMENT – 1: Install NLP TOOLKIT

Process::

To download Python, you need to visit www.python.org, which is the official Python website.

Click on the Downloads tab and then select the Windows option.

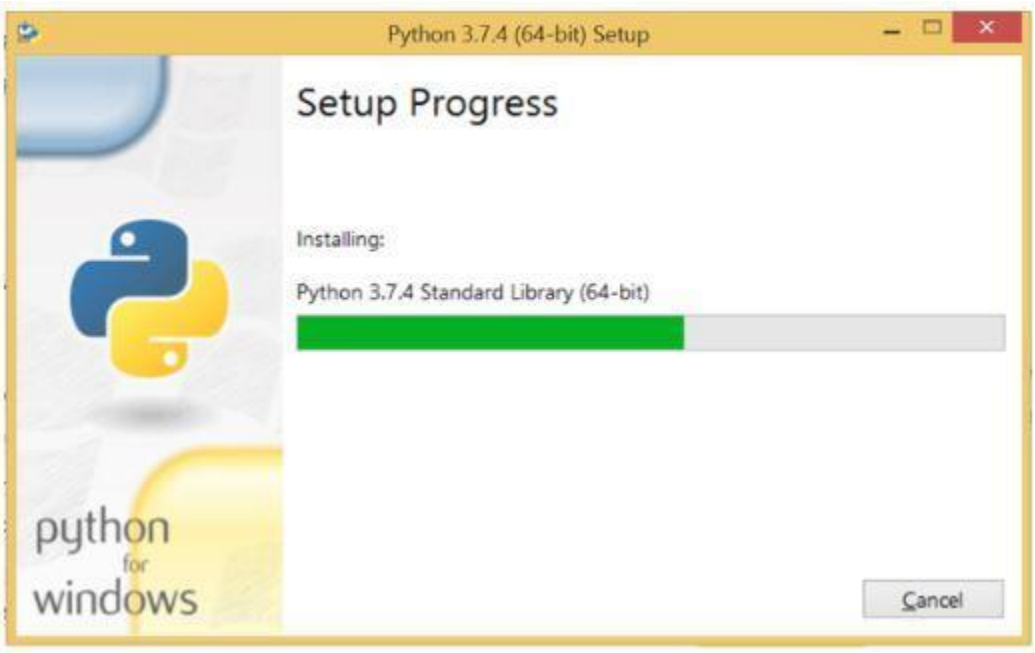
This will take you to the page where the different Python releases for Windows can be found. Since I am using a 64bit system, I'll select "Windows x86-64 executable installer".



Once the executable file download is complete, you can open it to install Python.

Click on Run, which will start the installation process.

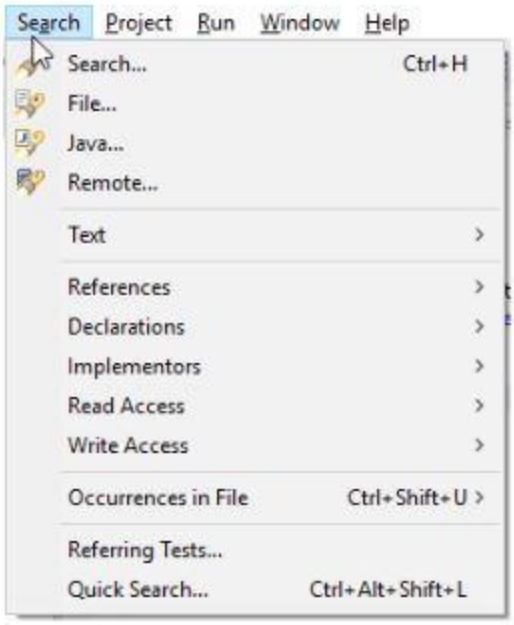
If you want to save the installation file in a different location, click on Customize installation; otherwise, continue with Install Now. Also, select the checkbox at the bottom to Add Python 3.7 to PATH.



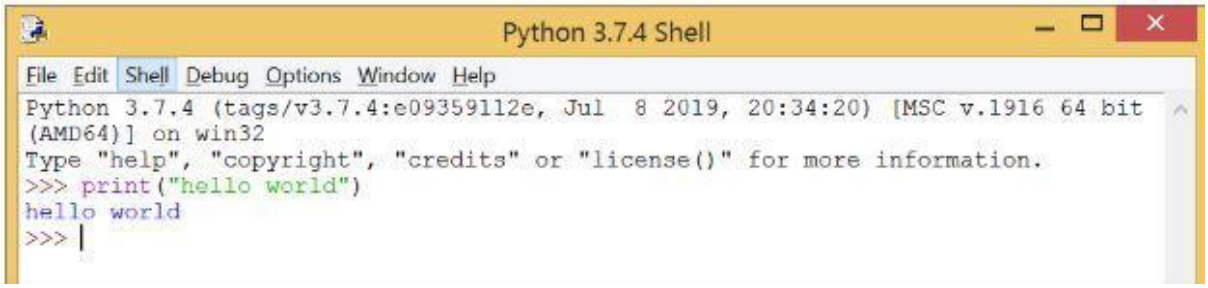
Once the installation is complete, the below pop-up box will appear: Setup was successful.

Now that the installation is complete, you need to verify that everything is working fine.

Go to Start and search for Python.

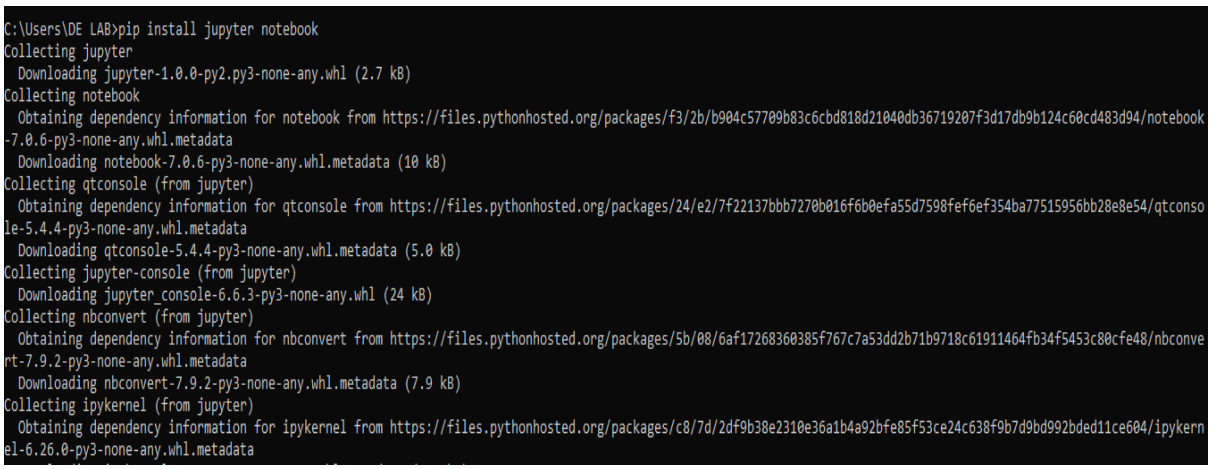


You can see Python 3.7 (64-bit) and IDLE. Let's open IDLE, which is the short form for Integrated Development Environment, and run a simple print statement.



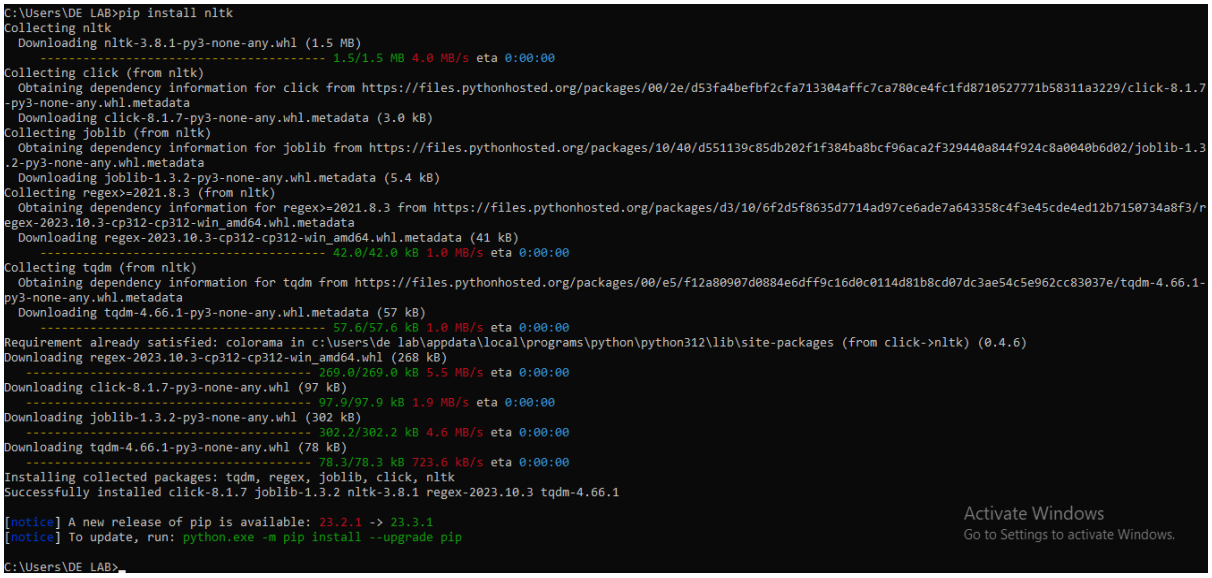
Now , Python is successfully installed on your windows

As the same install jupyter notebook by typing the following in Command Prompt



To install nltk type command in cmd as

Pip install nltk



EXPERIMENT: 2

NLTK Tokenize: Words and Sentences Tokenizer with Example

Tokenization is one of the first step in any NLP pipeline. Tokenization is nothing but splitting the raw text into small chunks of words or sentences, called tokens. If the text is split into words, then its called as 'Word Tokenization' and if it's split into sentences then its called as 'Sentence Tokenization'. Generally 'space' is used to perform the word tokenization and characters like 'periods, exclamation point and newline char are used for Sentence Tokenization. We have to choose the appropriate method as per the task in hand. While performing the tokenization few characters like spaces, punctuations are ignored and will not be the part of final list of tokens.

Types of Tokenization:

- **Sentence tokenization**
- **Word tokenization**

Aim: a) Write a python program to perform tokenization by word and sentence using nltk.

Program for sentence tokenization:

```
import nltk

nltk.download('punkt')  # Download the necessary tokenization models

 from nltk.tokenize import sent_tokenize

 def tokenize_sentences(text):

     sentences = sent_tokenize(text)

     return sentences
```

```
 # Example text
text = "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 resources such as WordNet, along with a
suite of text processing libraries for classification, tokenization, stemming, tagging, parsing, and semantic
reasoning, wrappers for industrial-strength NLP libraries, and an active discussion forum."

# Tokenize sentences
sentences = tokenize_sentences(text)

# Print tokenized sentences
for i, sentence in enumerate(sentences):
    print(f"Sentence {i+1}: {sentence}")
```

Output:

Program for word Tokenization:

```
 import nltk
nltk.download('punkt')  # Download the necessary tokenization models

from nltk.tokenize import word_tokenize

def tokenize_words(text):
    words = word_tokenize(text)
    return words

# Example text
text = "NLTK is a leading platform for building Python programs to work with human language data."

# Tokenize words
words = tokenize_words(text)
```

```
# Print tokenized words
print(words)
```

Output: