# Natural Language Processing using NLTK

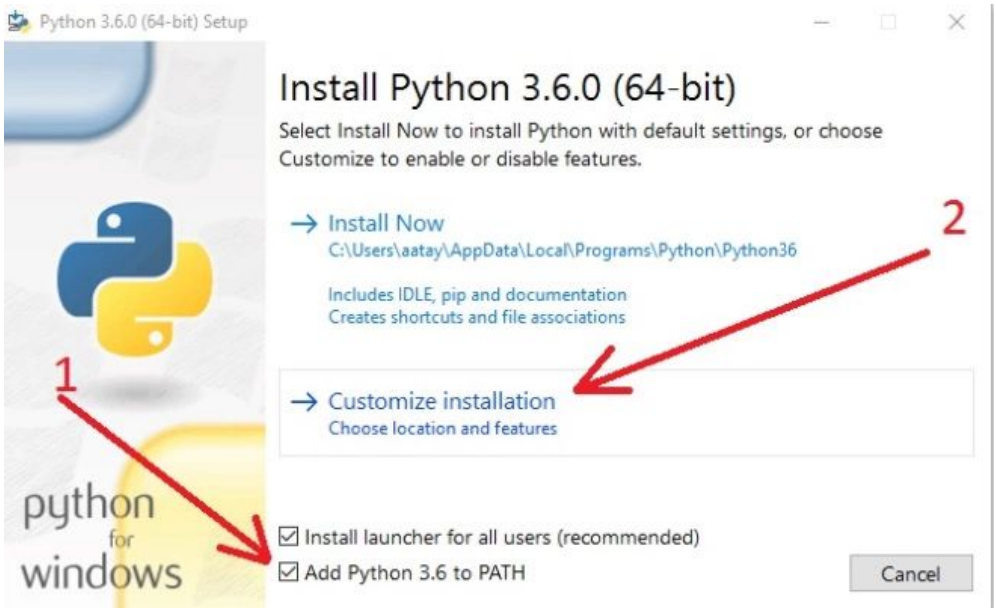
Please use this guide to install the important packages on your Windows system that would be needed for lab exercises.

## System-wide Packages

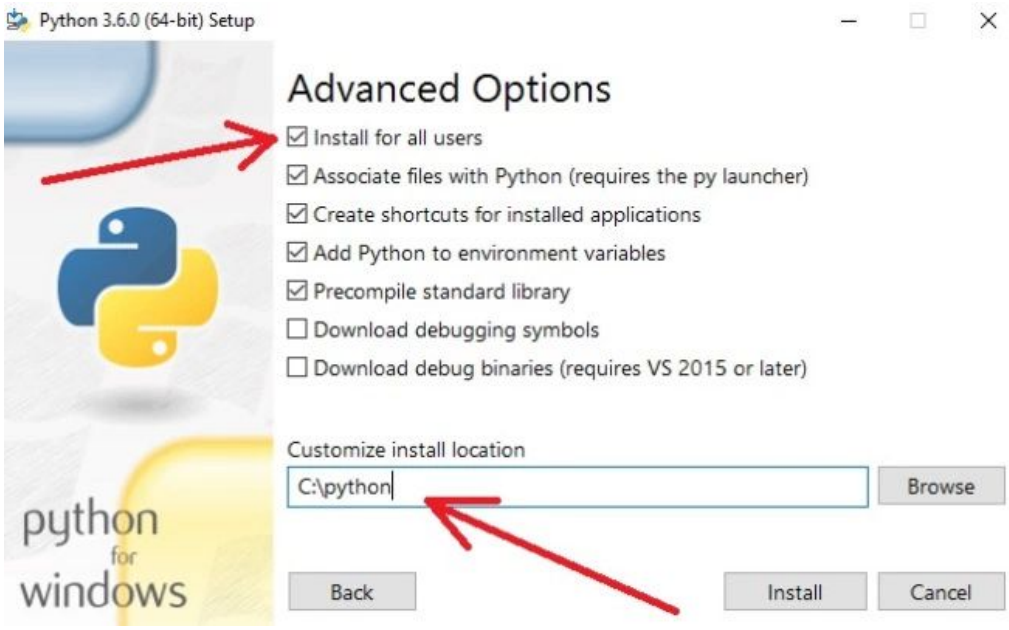
### Python 3

Let's start with the very first thing, Python!
We're going to use Python 3, specifically 3.6. Python can be installed by running the following [1]:

Download the latest Python 3.6 from https://www.python.org by going to Downloads->Windows and click on -> Latest Python 3 Release - Python 3.6.5 and at download the executable file acc. To your system configuration and run it. On running the .exe file :




After this, on the next window select all the Optional Features and click next.
After that do the changes as shown in the image.




And click on Install.
In the next screen, click "**Disable path length limit**" and click "Close" button to finish the installation procedure.

## Virtual Environment

Since Python comes in different flavours and versions, it is convenient to use a virtual environment to maintain Python packages independently for each project. You can install *virtualenv*, a tool to have isolated Python environments by following these commands[2]:

```
$ pip install virtualenv
```

Once the *virtualenv* is installed, you can create a project directory as:

```
$ mkdir NLPWorkshop
```

Now change the current directory to the new one we just made and initiate a virtual environment with name *venv*:

```
$ cd NLPWorkshop
$ virtualenv venv
```

To start using the virtual environment, we activate it using:

```
$ venv\Scripts\activate.bat
```

We install all our packages in this virtual environment.
Once done, we can deactivate the virtual environment by running:

```
$ deactivate
```

## Python Packages

We will need a certain Python packages throughout the lab sessions for the Summer School. These packages are listed below, with the commands to install them.

We recommend that you go through rest of this section to install the packages, however, if you don't want to install them individually, we provide a *win-requirements.txt* file here, that can be used to download all the dependencies. Once you download the *win-requirements.txt* file, run the following command, after activating the virtual environment to install the dependencies:

```
$ pip install -r win-requirements.txt
```

**Preprocessor**
Preprocessor is a preprocessing library for tweet data written in Python[13].
Download the zip file from here.
Run the pip command as :

```
$ pip install <path to the file>\tweet-preprocessor-0.4.0.zip
```

**Jupyter Notebook**
Jupyter Notebook is widely used in academia and industry for interactable Python development environment. Although, you can work with plain Python scripts as well, we highly recommend that you install it[3].

```
$ pip install jupyter
```

Once installed, you can run the notebook by:

```
$ jupyter notebook
```

**Tweepy**
We need a Python wrapper for Twitter API in order to write code in Python. There are number of Python wrappers available, but we recommend using *tweepy*[4].

```
$ pip install tweepy
```

**Matplotlib**
Matplotlib is a popular library used to create visualisations and plots[5].

```
$ pip install matplotlib
```

**NetworkX**
NetworkX is used to create, manipulate, and study of the structure and functions of complex networks[6].

```
$ pip install networkx
```

**NLTK**
Natural Language Toolkit is used for basic NLP tasks[7]:

```
$ pip install nltk
```

**NumPy**
NumPy is majorly used to run computation on high-dimensional arrays and has defined methods for high-level mathematical functions[8].

```
$ pip install numpy
```

**Pandas**
Python Data Analysis Library is a common data analysis library[9].

```
$ pip install pandas
```

**SpaCy**
SpaCy is another NLP library[10].

```
$ pip install spacy
```

Troubleshooting: Install spacy by using the binaries provided here.
To install a binary via .whl file, run:

```
$ python -m pip install <binary.whl>
```

**Gensim**

Gensim is used for topic modelling on text[11].

```
$ pip install gensim
```

**Scikit-learn**

It is a machine learning library for Python[12].

```
$ pip install scikit-learn
```

**jsonpickle**

jsonpickle is a library used for serialization and deserialization of complex Python objects to and from JSON[15].

```
$ pip install jsonpickle
```

**word_cloud**

A little word cloud generator in Python[16].

```
$ pip install wordcloud
```

# What Will We Learn?

1. A high-level introduction to what NLTK is, the documentation and its real world applications.
2. Exercise on regular expressions
3. Hands-On (Student can access data from library)
    - Requesting twitter data from Twitter API
    - Exploring Twitter Data and building basic graphs
4. Accessing data from the nltk corpus, exploring on various features of NLTK like:
    - Tokenization
    - Various Text Normalization techniques: Stemming and Lemmatization, advantages of one over the other.
    - POS tagging of words
    - Types of grammar and N-grams
5. Hands-On (Dataset will be provided):
    - Text cleaning (removal of noise, stop words, convert words to lowercase)
    - Generate word frequency
    - Exploratory Data Analysis (EDA) using word cloud
6. Popular methods to derive inferences from raw text:
    - Bag of Words
    - TF-IDF
    - Word2Vec
    - Cosine Similarity and Jaccard Similarity
7. Hackathon