# Getting Started with Python

Anaconda website: [www.anaconda.com](http://www.anaconda.com) (choose Individual Edition, Python 3.7)

## Using Jupyter Notebook

### Keyboard shortcut

* Run a command use either Ctrl+Enter or Shift+Enter
* Toggle between edit and command mode with Esc and Enter, respectively.
* Once in command mode:
  + Scroll up and down your cells with your Up and Down keys.
  + Press A or B to insert a new cell above or below the active cell.
  + M will transform the active cell to a Markdown cell.
  + Y will set the active cell to a code cell.
  + D + D (D twice) will delete the active cell.
  + Z will undo cell deletion.
  + Hold Shift and press Up or Down to select multiple cells at once.
    - With multiple cells selected, Shift + M will merge your selection.
* Ctrl + Shift + -, in edit mode, will split the active cell at the cursor.
* You can also click and Shift + Click in the margin to the left of your cells to select them.

# Part 2

## Python Collections

Collections of *heterogeneous objects.*

* List
* Tuple
* Set
* Dictionary (will be covered in Practical 3)

Table 1: Comparisons of list, tuple and set

|  |  |  |  |
| --- | --- | --- | --- |
| **Characteristics** | **List** | **Tuple** | **Set** |
| **Syntax** | list = [1,2,3] | tuple = (1,2,3) | set1 = set(list)  set2 = set(tuple) |
| **Items can be edited?** | Mutable (can) | Immutable (cannot) | Mutable, but the items inside the set must be immutable type |
| **Items are ordered?** | Ordered. Can be accessed by sequential index start from 0 | Ordered | Unordered. Not able to access the items with index |
| **Allow duplication?** | duplicate item is allowed | duplicate item is allowed | No duplicate item is allowed |

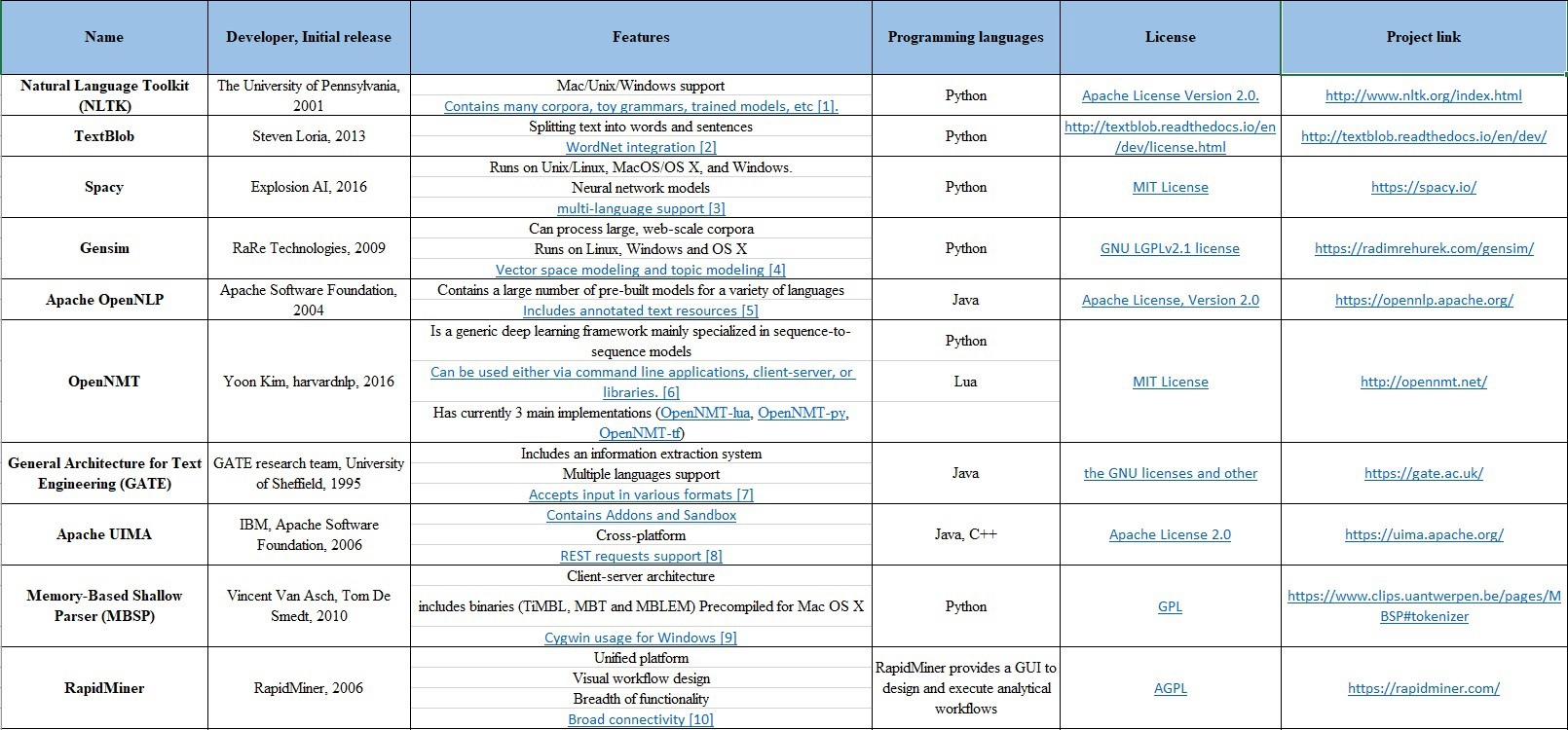
# Part 3

Text Processing is needed for transferring text from human language to machine-readable format for further processing. When a text is obtained, we start with text normalization. Text normalization includes:

* converting all letters to lower or upper case
* converting numbers into words or removing numbers
* removing punctuations, accent marks and other diacritics
* removing white spaces
* expanding abbreviations
* removing stop words, sparse terms, and particular words

## Tokenization

Tokenization is the process of splitting the given text into smaller pieces called tokens. Words, numbers, punctuation marks, and others can be considered as tokens.



## Remove Stop Words

“Stop words” are the most common words in a language like “the”, “a”, “on”, “is”, “all”. These words do not carry important meaning and are usually removed from texts. It is possible to remove stop words using Natural Language Toolkit (NLTK), a suite of libraries and programs for symbolic and statistical natural language processing.

## Stemming

Stemming is a process of reducing words to their word stem, base or root form (for example, books — book, looked — look).

## Lemmatization

The aim of lemmatization, like stemming, is to reduce inflectional forms to a common base form. As opposed to stemming, lemmatization does not simply chop off inflections. Instead it uses lexical knowledge bases to get the correct base forms of words.