**What Is NLTK?**

The [Natural Language Toolkit (NLTK)](http://www.nltk.org/) is a platform used for building programs for text analysis. The platform was originally released by Steven Bird and Edward Loper in conjunction with a computational linguistics course at the University of Pennsylvania in 2001. There is an accompanying book for the platform called [Natural Language Processing with Python](http://www.nltk.org/book/).

*Corpus, plural corpora; A collection of linguistic data, either compiled as written texts or as a transcription of recorded speech. The main purpose of a corpus is to verify a hypothesis about language - for example, to determine how the usage of a particular sound, word, or syntactic construction varies. Corpus linguistics deals with the principles and practice of using corpora in language study. A computer corpus is a large body of machine-readable texts.*

### Stop Words

Sometimes we need to filter out useless data to make the data more understandable by the computer. In natural language processing (NLP), such useless data (words) are called stop words. So, these words to us have no meaning, and we would like to remove them.

from nltk.corpus import stopwords

print(set(stopwords.words('English')))

### Searching

Let's say we have the following text file (download the text file from [Dropbox](https://www.dropbox.com/s/lztwlvo9myexd36/NLTK.txt?dl=1)). We would like to look for (search) the word language. We can simply do this using the NLTK platform as follows:

import nltk

file = open('NLTK.txt', 'r')

read\_file = file.read()

text = nltk.Text(nltk.word\_tokenize(read\_file))

match = text.concordance('language')

**The Gutenberg Corpus**

*Project Gutenberg (PG) is a volunteer effort to digitize and archive cultural works, to "encourage the creation and distribution of eBooks". It was founded in 1971 by Michael S. Hart and is the oldest digital library. Most of the items in its collection are the full texts of public domain books. The project tries to make these as free as possible, in long-lasting, open formats that can be used on almost any computer. As of 3 October 2015, Project Gutenberg reached 50,000 items in its collection.*

import nltk

gutenberg\_files = nltk.corpus.gutenberg.fileids()

print(gutenberg\_files)