# **Prescript Instructions**

\*\*These instructions require an above average level of Python. They also assume you are using an Anaconda build of Python.

## **Step 1: Installing the Natural Language Toolkit Library**

The Natural Language Toolkit (NLTK) library is a “leading platform for building Python programs to work with human language data.” (Bird, Loper, & Klein, 2009) The library uses many text corpora to help with and refine natural language processing. For the NLTK library to work, you must have the packages installed, as well as the corpora that corresponds to the task being performed. For example, to exclude a list of the most common words in the English language, the NLTK.corpus package should be downloaded.

* To install NLTK open an Anaconda console and use the command conda install -c anaconda NLTK.
* Once you have NLTK installed, you have to install the corpora packages.
* Open a Python interpreter.
* Within the shell window, type import NLTK and press Enter.
* Type nltk.download(). There will be a GUI window that will open up where you can download the necessary corpora.
* Select “all packages” and click download. Once it completes, you can move on to the next step.

## **Step 2: Installing the PyPDF2 Library**

PyPDF2 is a library that is useful for working with PDF files capable of splitting pages, extracting text, and manipulating PDF files. (About PyPDF2, 2019)

* To install PyPDF2, type pip install PyPDF2 within the Anaconda console window.

## **Step 3: Installing the Ebooklib and Textract Library**

The Textract library is used to extract text from different file types including Microsoft Word documents and PDF files. In order to parse the text, you must have the individual packages installed. Most of these are installed with the textract library. However, textract has known problems that are there are easy to overcome. These require additional steps to install the textract library.

* Open an Anaconda shell.
* Type the command conda install swig.
* After you install swig, download the EbookLib 0.15 zip from the releases at <https://github.com/aerkalov/ebooklib/releases>
* Extract the zip file to a folder location.
* Before you proceed further, you need to modify the README file Intro heading, line 44, to ASCII text like “Hello World”.
* Save the file.
* Go back to the Anaconda shell
* Change the directory to the unzipped folder location by typing cd file\location.
* Type pip install .
* Finally type pip install textract.

## **Step 4: Installing the re Library**

The re library “provides regular expression matching operations.” (Friedl, 2009) In this case, re is used to strip numbers from the output text.

* To install re, type pip install re within the Anaconda console window.

## **Step 5: Installing the Pandas Library**

The Pandas library is a popular Python package for data science. (Willems, 2019)

* To install Pandas, type pip install Pandas.

# References

*About PyPDF2*. (2019). Retrieved from pyhtonhosted.org: https://pythonhosted.org/PyPDF2/About%20PyPDF2.html

Bird, S., Loper, E., & Klein, E. (2009). *Natural Language Processing with Python.* O’Reilly Media Inc.

Friedl, J. (2009). *Mastering Regular Expressions.* O’Reilly Media.