

# Introduction to Text-as-Data Software Prerequisites\*

Instructor: Joseph L. Sutherland<sup>†</sup>

August 23, 2019

---

\*Prepared for the DSNCCConf, Raleigh, NC, 2019.

<sup>†</sup>Head of Data Science, Search Discovery (<https://searchdiscovery.com>). Email: [joseph.sutherland@searchdiscovery.com](mailto:joseph.sutherland@searchdiscovery.com).

## Setup Instructions

1. Install Homebrew (<http://brew.sh/>) if you're on a Mac, or Chocolatey (<https://chocolatey.org/>) if you're on a PC, by following the instructions at their homepages.
2. Install the Python distribution. This will also install Python's package manager, pip.
  - Mac: `brew install python python3`
  - PC: `choco install python python3` (run as administrator)
3. Make sure pip is the latest version.
  - Mac: `pip install --upgrade pip`
  - PC: `pip install --upgrade pip` (run as administrator).
4. Use pip to install jupyter notebook.
  - Mac: `pip install jupyter`
  - PC: `pip install jupyter` (run as administrator).

**Important:** Make sure you are using **Python 3.x**, not Python 2.x (x can be any digit). These days, pip should redirect automatically to pip3. Use pip3 in case of emergency.

## Required Python Packages

It is necessary to have these packages installed and working before you arrive for class. After installing, make sure you can import the package by running `import PACKAGENAME` in the python environment, or even better, the jupyter notebook (see section below)!

## Data Management

```
pip install pandas
pip install numpy
pip install scipy
```

## Natural Language Processing

```
pip install nltk
pip install gensim
pip install unicodecode
```

### Install NLTK Data

After nltk is installed, open python3 by typing `python` or `python3` into your terminal. Type `import nltk`, then type `nltk.download()`. This should open a dialog box for extra NLTK packages. Double-click the "book" option to download all of the packages needed for the NLTK book (and for this course).

## Machine Learning

```
pip install scikit-learn
pip install gensim
```

## Statistics and Visualization

```
pip install statsmodels
pip install wordcloud
pip install seaborn
```

## Booting Jupyter Notebook

To boot the notebook, simply execute the following command: `jupyter notebook` (on PC it is sometimes `jupyter notepad`) from the command line.