README FOR PYTHON & NLTK

Important resources:

- Installing Python: https://docs.python.org/3/
- Installing Natural Language Tool Kit: https://www.nltk.org

Instructions from https://www.nltk.org:

NLTK 3.4.5 documentation

1. **Installing NLTK**

- NLTK requires Python versions 3.5, 3.6, or 3.7
- For Windows users, it is strongly recommended that you go through this guide to install Python 3 successfully:
    - https://docs.python-guide.org/starting/install3/win/#install3-windows

2. **Setting up a Python Environment (Mac/Unix/Windows)**

- Please go through this guide to learn how to manage your virtual environment managers before you install NLTK, https://docs.python-guide.org/dev/virtualenvs/
  Always use virtual environment!

- Alternatively, you can use the Anaconda distribution installer that comes "batteries included" https://www.anaconda.com/distribution/

3. **Mac/Unix**

1. Install and activate virtualenv
2. Install NLTK: run `pip install nltk`
3. Install Numpy (optional): run `pip install numpy`
4. Test installation:
    - run `python` then type `import nltk`

4. **Windows**

These instructions assume that you do not already have Python installed on your machine.

32-bit binary installation

1. Install Python 3.7: http://www.python.org/downloads/ (avoid the 64-bit versions)
2. Install Numpy (optional): https://www.scipy.org/scipylib/download.html
3. Install NLTK: http://pypi.python.org/pypi/nltk
4. Test installation:
    - `Start>Python37`, then type `import nltk`

NLTK Tokenizer Package

Tokenizers divide strings into lists of substrings. For example, tokenizers can be used to find the words and punctuation in a string. They can also operate at the level of sentences, using the sentence tokenizer directly. Please check https://www.nltk.org/api/nltk.tokenize.html

Example:

```
>>> from nltk.tokenize import sent_tokenize, word_tokenize
>>> s = '''Good muffins cost $3.88
... in New York. Please buy me
... two of them. Thanks. '''
>>>
>>> word_tokenize(s)
['Good', 'muffins', 'cost', '$', '3.88', 'in', 'New', 'York', '.', 'Please', 'buy', 'me', 'two', 'of', 'them', '.', 'Thanks', '.']
>>>
>>> sent_tokenize(s)
['Good muffins cost $3.88\nin New York.', 'Please buy me \ntwo of them.', 'Thanks.']
>>>
```