Natural Language Toolkit

NLTK is a leading platform for building Python programs to work with human language data. It provides easy-to-use interfaces to over 50 corpora and lexical resources such as WordNet, along with a suite of text processing libraries for classification, tokenization, stemming, tagging, parsing, and semantic reasoning, wrappers for industrial-strength NLP libraries, and an active discussion forum.

Thanks to a hands-on guide introducing programming fundamentals alongside topics in computational linguistics, plus comprehensive API documentation, NLTK is suitable for linguists, engineers, students, educators, researchers, and industry users alike. NLTK is available for Windows, Mac OS X, and Linux. Best of all, NLTK is a free, open source, community-driven project.

NLTK has been called "a wonderful tool for teaching, and working in, computational linguistics using Python," and "an amazing library to play with natural language."

Natural Language Processing with Python provides a practical introduction to programming for language processing. Written by the creators of NLTK, it guides the reader through the fundamentals of writing Python programs, working with corpora, categorizing text, analyzing linguistic structure, and more. The book is being updated for Python 3 and NLTK 3. (The original Python 2 version is still available at http://nltk.org/book_1ed.)
Platforms

• Today I'll go through installation for:
  1. MacOS (and Linux)
  2. Windows 10

• Your homework assignment:
  • install nltk, and
  • check to see if it works (by next time)

• In the next few lectures:
NLTK 3.2.5 Install

• See http://www.nltk.org/install.html
• Use pip3 (for python3) to install packages from the Python Package Index (PyPI)
• sudo pip3 install –U nltk

![Installation Command Output]

- Updated my nltk 3.2.4 to 3.2.5
NLTK Data Install

- See [http://www.nltk.org/data.html](http://www.nltk.org/data.html)
- `python3`

Python 3.5.2 (v3.5.2:4def2a2901a5, Jun 26 2016, 10:47:25)
[ GCC 4.2.1 (Apple Inc. build 5666) (dot 3) ] on darwin
Type "help", "copyright", "credits" or "license" for more.

```python
>>> import nltk
>>> nltk.download()
showing info https://raw.githubusercontent.com/nltk/nltk_data/master/packages/index.xml
```

- If you get an SSL certificate error message, run:
  `/Applications/Python 3.6/Install Certificates.command`
Windows 10: setup

- Environment variable PATH should be set correctly to point to Python 3 install directory

- Type in search:
  - Edit environment variables for your account
Windows 10: install nltk

- On the command line:
  - `pip3 install pyyaml nltk`
  - Package pyyaml must be used somewhere in nltk ...

- Source: [http://www.pitt.edu/~naraehan/python2/faq.html](http://www.pitt.edu/~naraehan/python2/faq.html)
Windows 10: install numpy and test nltk

• On the command line:
  • `pip3 install numpy`
  • (the chunking algorithm uses it)

• Let's test nltk:
  • `.word_tokenize()` converts a string into words
  • `.pos_tag()` does part-of-speech tagging
  • `.ne_chunk()` does named entity recognition
Windows 10: test nltk

- `.draw()` takes a Tree object and draws it in a pop-up window.
Windows 10: install nltk data

- Install corpus data (from inside Python) using
  - `nltk.download()`
There is a sample of the well-known Penn Treebank Wall Street Journal (WSJ) corpus included

- 3,914 parsed sentences
- 49,000+ parsed sentences in the full corpus
From the class exercise
nltk: where is it installed?