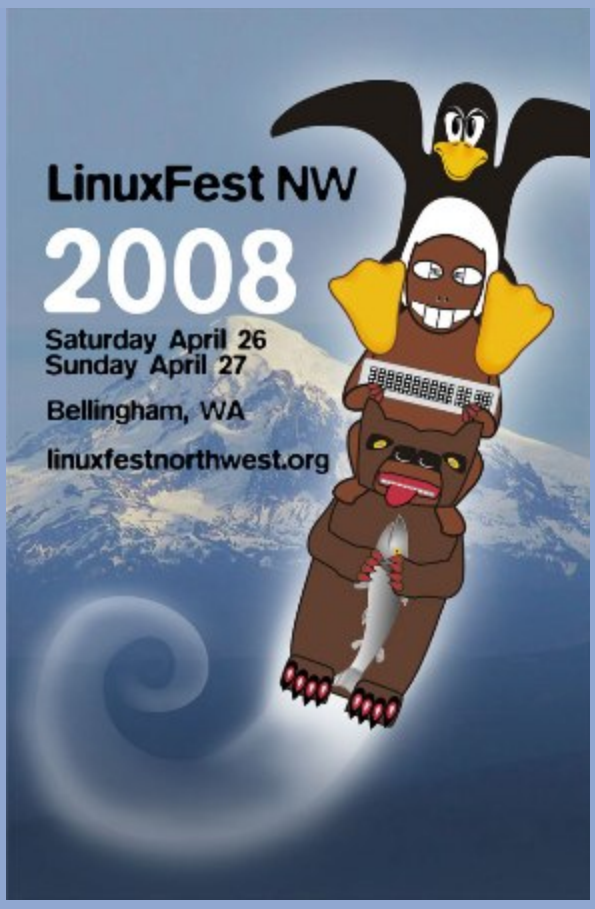
Natural Language Processing in Python using NLTK

Sean Boisen,



<[myfirstname]@logos.com>

LinuxFest Northwest 2008, April 26

Slides at
http://semanticbible.org/other/talks/2008/nltk/nltk.html

But First, a Word from our Sponsors …

- 

- My website: http://www.SemanticBible.com/
- My blog: http://www.SemanticBible.com/Blogos/
- Slidy, a standards- and browser-based XHTML presentation framework
- And the letters N, L, and P

Goals

Intersection of an entire technical field, a sophisticated programming language, and a complex toolkit: i can only do so much.

- Introduce the field of Natural Language Processing (NLP)
- Introduce NLTK
- Illustrate some simple applications for processing language
- Convince you it's simple enough to download and try it out!
- Non Goals:
  - *Teaching Python*
  - *Comprehensive discussion of NLP*
  - *Theoretical justification*
  - *Depth*

Intended Audience

- Pythonistas (or the ability to follow along) who:
  - *want to go beyond string processing and regexps*
  - *deal with large quantities of prose text*
  - *are looking for new tools and techniques*
- People who want to explore NLP technology

Outline

- Overviews: me, you, Python, NLP
- Introducing the Natural Language Toolkit (NLTK)
- Corpora
- Words and Word Frequencies
- Lexicons
- Part-of-speech analysis
- Word sense disambiguation

Overview: Why Me?

- MA in Linguistics from UCLA
- 19 years as a researcher and manager with Natural Language Processing group at BBN Technologies
- Senior Information Architect with Logos Bible Software, working on:
  - *Semantic Web technologies*
  - *Text processing and content extraction*
  - *Information visualization*
  - *General datamongering*
- Let's also hear a little about you.

Overview: Why Python?

- Clean, elegant, object-oriented language
- Rich set of library modules
- Active open-source community
- Interactive evaluation supports learning and rapid development
- Broadly used in scientific research

Overview: What is Natural Language Processing?

- A variety of technologies for analyzing, manipulating, and generating human language data
- Related fields of study:
  - *Linguistics and Cognitive Science*
  - *Artificial Intelligence*
  - *Computer Science*
  - *Probability and Statistics, Machine Learning*

Overview: What is Natural Language Processing? (2)

- Over the last 2 decades, approaches in the field have shifted toward:
  - *Computational*
  - *Statistical*
  - *Corpus-driven*

Why we need NLP

- Words are ambiguous: "stock" meaning
  - *share of a company*
  - *broth*
  - *goods stored in a warehouse*
  - *unmodified*
- Much information in language is implicit:
  - *"Edward Nance, 34, doesn't think ....*
    Nance, who is also a paid consultant to ABC News, said ..."
  - *My wife loves pickles, but I don't.*

Annoying Questions (Computational) Linguists Hear

- How many languages do you speak?
- How many words are in the English language?
- How many Eskimo words are there for snow?
  - *See The Great Eskimo Vocabulary Hoax by (linguist) Geoffrey Pullum*

Overview: The Natural Language Toolkit

- Available from http://nltk.org: "A suite of open source Python modules, data sets, and tutorials supporting research and development in NLP", including:
  - *Code: corpus readers, tokenizers, stemmers, taggers, parsers, WordNet, and more (50k lines of code)*
  - *Corpora: >30 annotated data sets widely used in natural language processing (>300Mb data)*
  - *Documentation: a 400-page book, articles, reviews, API documentation, licensed under Creative Commons (see individual license details)*
- Used for instruction in more than 3 dozen universities
- Project leaders: Steven Bird, Ewan Klein, Edward Loper, along with dozens of other developers and contributors

Overview: The Natural Language Toolkit (2)

- All open source and freely available
- Under active development by a responsive community
- Most modules have a demo() function that shows what they do
- Installation:
  - *Pre-requisite: Python 2.4 or later*
  - *Download and install the code (tarball, zip, Windows installer)*
  - *Download the data (corpora)*
  - *Optional modules:* `numpy`, `matplotlib`
  - *Optional: download the documentation*

So What Can You Do With NLTK?

- Work with a corpus
- Tokenize, stem, and count words
- Part-of-speech analysis ("tagging")
- Work with word semantics

So What *Else* Can You Do With NLTK?

(We won't have time to cover these today)

- More sophisticated kinds of part-of-speech tagging
- Simplified phrase analysis ("chunking") and sophisticated syntactic parsing
- Run a chatbot
- Feature-based statistical classification
- Draw trees, directed graphs, graphs, and other objects using Tkinter
- Work with other container objects
- Theorem proving and model building
- Create Word Finder style puzzles (and their solution)
- Represent and evaluate semantic structures for language as first-order logic formulas

NLTK Corpora

- NLTK includes more than 40 corpora and corpus samples (750Mb), along with readers and processing methods
- Various types and stages of analysis
- Text: News, presidential addresses
- Tagged: Brown Corpus
- Parsed: Penn Treebank Sample, other languages
- Categorized: Reuters-21578
- Lexicons: CMU Pronouncing Dictionary, stopwords in multiple languages
- Shakespeare XML corpus
- ... and many others

Useful Utilities: `nltk.probability`

- frequency distributions for counting things
- probability distributions
- conditional distributions

Useful Utilities: `nltk.evaluate`

- Confusion matrices
- Levenshtein distance
- Evaluation metrics

Useful Utilities: nltk.evaluate (2)

```
c:\Python24\Lib\site-packages\nltk>evaluate.py
--------------------------------------------------------------
--------------
Reference = ['DET', 'NN', 'VB', 'DET', 'JJ', 'NN', 'NN',
'IN', 'DET', 'NN']
Test      = ['DET', 'VB', 'VB', 'DET', 'NN', 'NN', 'NN',
'IN', 'DET', 'NN']
Confusion matrix:
    | D         |
    | E I J N V |
    | T N J N B |
----+-----------+
DET | 3 0 0 0 0 |
 IN | 0 1 0 0 0 |
 JJ | 0 0 0 1 0 |
 NN | 0 0 0 3 1 |
 VB | 0 0 0 0 1 |
----+-----------+
(row = reference; col = test)
Accuracy: 0.8
--------------------------------------------------------------
--------------
Reference = set(['VB', 'DET', 'JJ', 'NN', 'IN'])
Test = set(['VB', 'DET', 'NN', 'IN'])
Precision: 1.0
Recall: 0.8
F-Measure: 0.888888888889
```

Chunking

- In `nltk.chunk`
- Simplified parsing to identify phrases, without complete syntactic analysis
- Uses:
  - *First step in key phrase extraction (e.g. Amazon's "statistically improbable phrases")*

Chatbots

- In `nltk.chat`
- Eliza-style patterns of question/response with fuzzy matching

Other NLP Appplications (not in NLTK)

- Machine Translation
- Information Extraction
- Summarization
- Speech recognition
- Audio transcription and segmentation (EveryZing)
- Dialogue Systems (spoken or written)

Resources

- WordNet
- Learn more about NLTK
  - *Getting Started on Natural Language Processing with Python, Nitin Madnani. ACM Crossroads, Volume 13, Issue 4, 2007. (NB this version revised slightly from ACM version.)*
- MIT Open Courseware