NLTK Tutorial

Pham Quang Nhat Minh

FPT Technology Research Institute (FTRI)
minhpqn@fpt.edu.vn

November 12, 2016

Table of contents

What is NLTK?

- Stand for Natural Language Toolkit (website: http://www.nltk.org/)
- A suite of libraries and programs for Natural Language Processing for English
- Developed by Steven Bird and Edward Loper in the Department of Computer and Information Science at the University of Pennsylvania.
- Provide easy-to-use interfaces to over 50 corpora and lexical resources and text processing libraries.

Goals of NLTK

- GOALS
  - Simplicity
  - Consistency
  - Extensibility
  - Modularity
- NON-GOALS
  - Encyclopedic coverage
  - Optimization/clever tricks

What NLTK can do?

- Accessing Text Corpora and Lexical Resources
- Processing Raw Text
    - Sentence segmentation
    - Tokenization
    - Word Stemming
    - Lemmatization
- Categorizing and Tagging Words
    - POS Tagging
    - Text Classification

What NLTK can do?

- Extracting Information from Text
  - Chunking
  - Named Entity Recognition
  - Relation Extraction
- Analyzing Sentence Structure
  - Phrasal structure analysis
  - Dependency Parsing
- Analyzing the Meaning of Sentences (Semantic Analysis)

(Some) NLTK Modules

Task	NLTK module	Some functionalities
Accessing corpora	nltk.corpus	Standardized interfaces to corpora and lexicons
String processing	nltk.tokenize	Sentence and word tokenizers
	nltk.stem	Stemmers
Part-of-speech tagging	nltk.tag	Various part-of-speech taggers
Classification	nltk.classify	Decision tree, maximum entropy
	nltk.cluster	K-means
Chunking	Nltk.chunk	Regular expressions, named entity tagging

Download and install nltk

- `nltk` is included in anaconda
    - Download and Install Anaconda on first.
- Test your install
    - Type `import nltk` in your Interactive Python terminal
- Install nltk data
    - Interactive installer
      ```
      >> import nltk
      >> nltk.download()
      ```
    - Command line installation
      ```
      >> python -m nltk.downloader all
      ```
- It takes for a while to download all nltk.data
    - You can copy from my laptop

Processing text with NLTK

We will learn how to use nltk to perform basic text processing tasks:

- Sentence segmentation
- Tokenization
- Lemmatization
- Word stemming
- POS tagging

Processing text with NLTK

- Information Extraction
  - Chunking
  - Named Entity Recognition
- Syntax Analysis
  - Phrasal structure analysis
  - Dependency Parsing
  - Extract predicate/argument structures

Processing text with NLTK

- See iPython notebook "NLTK Basic Tutorial" on:
  http://tinyurl.com/hqaug8g
- PDF version of the tutorial is upload to CMS.
- NLTK Practice exercises
  - NLTK exercises are uploaded to CMS

References

- Bird, Steven; Klein, Ewan; Loper, Edward (2009). *Natural Language Processing with Python*. O'Reilly Media Inc. http://www.nltk.org/book/
- Perkins, Jacob (2010). *Python Text Processing with NLTK 2.0 Cookbook*. Packt Publishing.
- The Python Tutorial https://docs.python.org/3/tutorial/