>>> PYTHON TUTORIAL 4 NLP

Fall 2016

Sardar Hamidian

sardar@gwu.edu

Good to Know More about python!

– Python is a high-level, interpreted, interactive and object-oriented scripting language

Why Python?

– Easy to learn and read

– Broad Standard Libraries

– Database

Learn more at

http://www.tutorialspoint.com/python/python_overview.htm

Python Installation

- Anaconda
  - Why?
    - More than 100 python packages
    - Easier to setup a standard environment for large-scale setups within organizations
    - Dependencies

    Other platforms: Canopy
- Manual installation

  https://www.python.org/

Recommended IDE's

- Pycharm*
- Spyder*
- Komodo
- Pydev
- VIM
- IDLE

Outline

- Python
  - Version 3.5
- NLTK
  - Installation:
    - pip installation nltk, nlt.download()
    - conda install -c anaconda nltk=3.2.1

  Features:
  - 50 corpora and lexical resources
  - Stemming, Tokenization, Parsing, Semantic reasoning

- Please make a directory for this Tutorial.
- Code and Contents:

https://github.com/sardarr/Pyhthon_NLP_Tutorial.git

Exercise #1

- Reading and writing a document!
- Word and Sentences Tokenization
- Preprocessing
  - Stop words (a, the, and, …..)
  - Stemming
  - Lemmatization
    - Which one could help you more stemming or lemmatization?

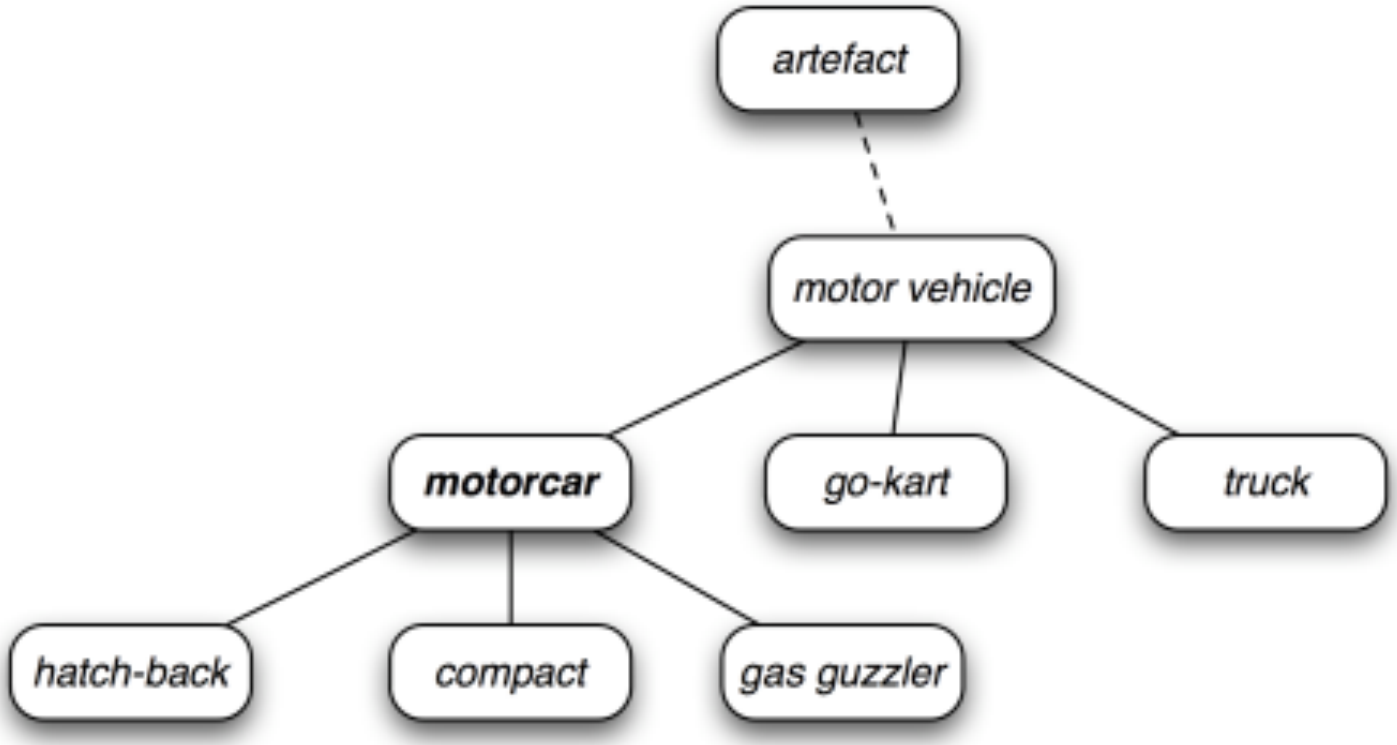
Exercise #2

- Part of Speech Tagging POS
- NER(***lots of FP)

Exercise #3

- Accessing different corpora in NLTK
- Popular corpora
- Wordnet(synonym, antonym, context,…)
- Semantic Similarity

Fragment of WordNet Concept Hierarchy: nodes correspond to synsets



Methods in Python

- Methods

  def name_of_function(inputs):

  .

  .

  .

  Return ()*

- Import and using methods

Exercise#4

- N-gram
  - Unigram, bigram, trigram
- Word count

Exercise#5

- Word list corpora in NLTK
  - Set of English word
  - Set of chat words
  - Set of unusual words?

Regular Expression

- What is Regular expression
- Finding a pattern inside text body
- Regex exercise
- http://regexr.com/ for practice

Useful tools and libraries

- Beautiful soup
- Scrapy python
- Scikit learn
- Weka

Review Classification using NLTK

- Code: Movie_review.py
- Import Naïve Bayes
- Feature Extraction
- Creating Train and Test Set

- You can find the code in my Github

>>>Thank you!!!