Introduction to NLP & Spacy

ChennaiPy | May Meetup

About myself

vishalgupta.me

- 3rd Year CSE Undergrad @SSN
- Google Summer of Code 2018 intern @ Debian
- Student Researcher @ Microsoft Research India

Definitions

Natural Language

An ordinary language such as English, Tamil, etc. that humans speak

NLP

Parsing and understanding Natural language to mine information

Libraries

NLTK & Spacy

NLTK

Natural Language ToolKit

Created to support education.
For demo purposes, to help students explore ideas

- Provides
  - Corpora
  - String processing and Tokenizing
  - POS Tagging
  - Chunking (n-grams)
  - Machine Learning
  - Probability and Estimation
- NLP with Python (O'Reilly book) http://www.nltk.org/book/
- Only for English
- pip install nltk
- ```
  >>> import nltk
  >>> nltk.download()
  ```
  ```

spaCy

Written to help you get things done.
"Industrial-Strength Natural Language Processing"

- Fastest
- Named entity recognition
- Support for 28+ languages
- Pre-trained word vectors
- Easy deep learning integration
- POS tagging and dependency parsing
- Syntax-driven sentence segmentation
- Visualizers for syntax and NER
- Easy model packaging and deployment
- State-of-the-art speed
- pip install spacy
  python -m spacy download en
- ```
  >>> import spacy
  >>> nlp = spacy.load('en')
  ```
  ```

NLTK vs Spacy

	ABSOLUTE (MS PER DOC)			RELATIVE (TO SPACY)		
SYSTEM	TOKENIZE	TAG	PARSE	TOKENIZE	TAG	PARSE
spaCy	**0.2ms**	**1ms**	**19ms**	1x	1x	1x
CoreNLP	2ms	10ms	49ms	10x	10x	2.6x
ZPar	1ms	8ms	850ms	5x	8x	44.7x
NLTK	4ms	443ms	*n/a*	20x	443x	*n/a*

Basics of NLP

Corpus

Large bodies of linguistic data

Variety of domains and authors

NLTK has built-in support for dozens of corpora and trained models

Links

- http://www.nltk.org/nltk_data/
- https://www.nltk.org/book/ch02.html

Basic Pre-processing

Eliminating noise and processing text

Lowercase

Regex

Special Character elimination

Regularised Encoding
(Unicode issues)

Tokenization
Sentence / Word

Stopword elimination

https://www.nltk.org/book/ch03.html

Language-based Pre-processing

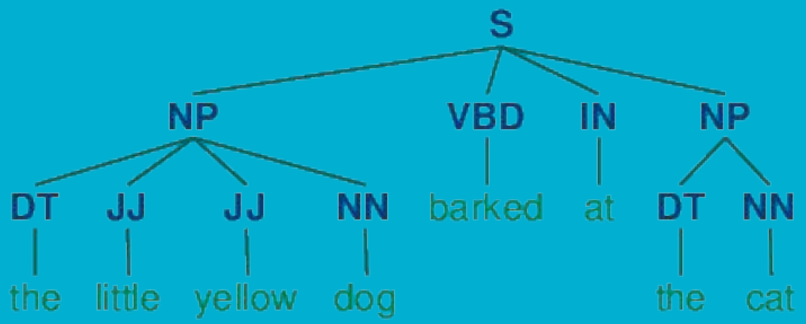
Processing text using some properties of the Natural Language being used

POS Tag
(POS = Parts of Speech)

Stemming

Lemmatization

http://textminingonline.com


S
NP
VBD
IN
NP
DT
JJ
JJ
NN
barked
at
DT
NN
the
little
yellow
dog
the
cat

POS Tagging

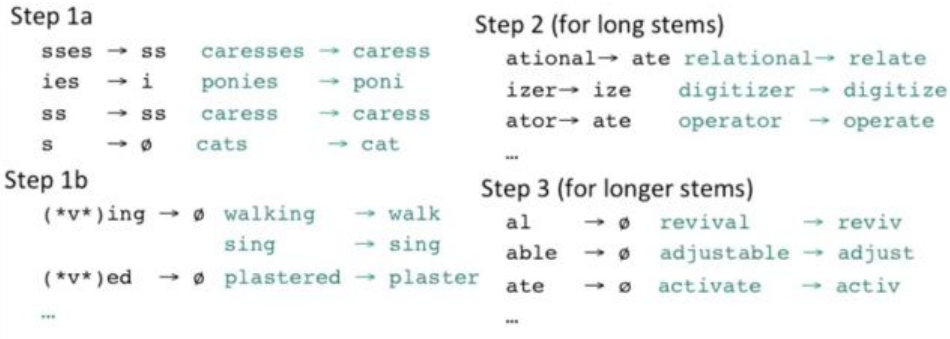
Tag	Description
CC	Coordinating conjunction
CD	Cardinal number
DT	Determiner
EX	Existential there
FW	Foreign word
IN	Preposition or subordinating conjunction
JJ	Adjective
JJR	Adjective, comparative
JJS	Adjective, superlative
LS	List item marker
MD	Modal
NN	Noun, singular or mass
NNS	Noun, plural
NNP	Proper noun, singular
NNPS	Proper noun, plural
PDT	Predeterminer
POS	Possessive ending
PRP	Personal pronoun

Tag	Description
PRP$	Possessive pronoun
RB	Adverb
RBR	Adverb, comparative
RBS	Adverb, superlative
RP	Particle
SYM	Symbol
TO	to
UH	Interjection
VB	Verb, base form
VBD	Verb, past tense
VBG	Verb, gerund or present participle
VBN	Verb, past participle
VBP	Verb, non3rd person singular present
VBZ	Verb, 3rd person singular present
WDT	Whdeterminer
WP	Whpronoun
WP$	Possessive whpronoun
WRB	Whadverb

Stemming vs. Lemmatization

- To reduce inflectional forms and sometimes derivationally related forms of a word to a common base form
- **Stemmer** : Derive the stem a word (branch) is derived from with language-specific production rules.
  Exact stemmed form does not matter, only the equivalence classes it forms.

  `car, cars, car's, cars'  => car`



- **Lemmatizers** : derive lemma of a word using a complete vocabulary and morphological analysis.

  `am, are, is  => be`

  `drive, drives, drove, driven  => drive`

Code Samples

tiny.cc/chpyNLP