Problem Solving with NLTK

MSE 2400 EaLiCaRA

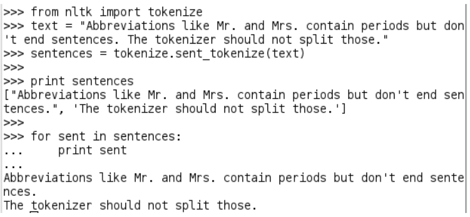
Dr. Tom Way

NLTK

- Natural Language Toolkit (NLTK) is a large collection of Python modules to facilitate natural language processing
- It also includes a large amount of optional data, such as annotated text corpora and WordNet
- NLTK: http://www.nltk.org/
- Free NLTK Book: http://www.nltk.org/book

Detecting Sentence Boundaries

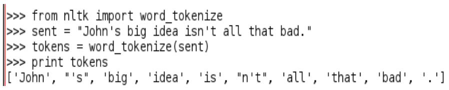
- sent_tokenize is NLTK’s current recommended method to tokenize sentences



```
>>> from nltk import tokenize
>>> text = "Abbreviations like Mr. and Mrs. contain periods but don
't end sentences. The tokenizer should not split those."
>>> sentences = tokenize.sent_tokenize(text)
>>>
>>> print sentences
["Abbreviations like Mr. and Mrs. contain periods but don't end sen
tences.", 'The tokenizer should not split those.']
>>>
>>> for sent in sentences:
...     print sent
...
Abbreviations like Mr. and Mrs. contain periods but don't end sente
nces.
The tokenizer should not split those.
```

Tokenizing Words

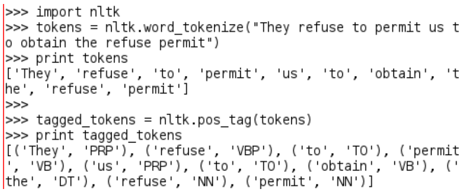
- word_tokenize is NLTK’s current recommended method to tokenize words



```
>>> from nltk import word_tokenize
>>> sent = "John's big idea isn't all that bad."
>>> tokens = word_tokenize(sent)
>>> print tokens
['John', "'s", 'big', 'idea', 'is', "n't", 'all', 'that', 'bad', '.']
```

Parts of Speech Tagging

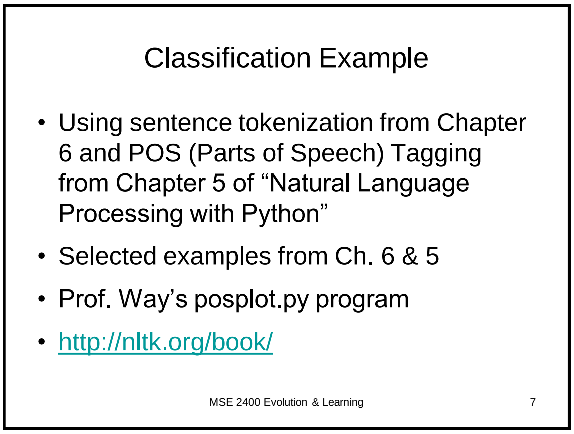
- Identifying the part of speech for each word in a text document



```
>>> import nltk
>>> tokens = nltk.word_tokenize("They refuse to permit us t
o obtain the refuse permit")
>>> print tokens
['They', 'refuse', 'to', 'permit', 'us', 'to', 'obtain', 't
he', 'refuse', 'permit']
>>>
>>> tagged_tokens = nltk.pos_tag(tokens)
>>> print tagged_tokens
[('They', 'PRP'), ('refuse', 'VBP'), ('to', 'TO'), ('permit
', 'VB'), ('us', 'PRP'), ('to', 'TO'), ('obtain', 'VB'), ('
the', 'DT'), ('refuse', 'NN'), ('permit', 'NN')]
```

Classification

- nltk.classify includes decision tree, maximum entropy and naive bayes classifiers



Classification Example

- Using sentence tokenization from Chapter 6 and POS (Parts of Speech) Tagging from Chapter 5 of “Natural Language Processing with Python”
- Selected examples from Ch. 6 & 5
- Prof. Way’s posplot.py program
- http://nltk.org/book/