NLTK corpus readers

In [3]:
```
import nltk
```

NLTK (I installed with sudo pip install nltk) can download some corpora for you. This seems to be a mixture of whole corpora, partial corpora and stubs for corpora.

- http://nltk.org
- http://nltk.org/nltk_data/
- http://freecode.com/articles/processing-corpora-with-python-and-the-natural-language-toolkit

In [11]:
```
nltk.download('brown')
```

```
[nltk_data] Downloading package 'brown' to /Users/alex/nltk_data...
[nltk_data]   Package brown is already up-to-date!
```

Out[11]:
```
True
```

Some example stuff to do with NLTK on a downloaded corpus.

In [18]:
```
from nltk.corpus import brown
brown_news_tagged = brown.tagged_words(categories='news', simplify_tags=Tru
tag_fd = nltk.FreqDist(tag for (word, tag) in brown_news_tagged)
print(tag_fd.keys())
```

```
['N', 'DET', 'P', 'NP', 'V', 'ADJ', ',', '.', 'CNJ', 'PRO', 'ADV', 'VD',
'NUM', 'VN', 'VG', 'TO', 'WH', 'MOD', '``', "''", 'VBZ', '', '*', ')',
'(', 'EX', ':', 'FW', "'", 'UH', 'VB+PPO']
```

NKTK PropBank 1 corpus

In [17]:
```
nltk.download('propbank')
```

```
[nltk_data] Downloading package 'propbank' to /Users/alex/nltk_data...
[nltk_data]   Package propbank is already up-to-date!
```

Out[17]:
```
True
```

In [23]:
```
from nltk.corpus import propbank
print propbank.readme()
```

```
Proposition Bank Corpus 1.0

Martha Palmer   http://verbs.colorado.edu/~mpalmer/projects/ace.html

Distributed with Permission
```

```
This directory contains the data of the UPenn Propbank.  This data is
collected
as an additional layer of annotation on the Penn Treebank, representing
the predicate
argument structure of verbs.  Below is a list of each file and a
description of
its contents.


File                        Description
------------------------------------------------------------------------
--------------
prop.txt                    The annotated data, file format described
below.
```

In [ ]: