

Part-of-speech (POS) tagging

- What is POS tagging?
- Why POS-tagging?
- How to access POS-tagged data?
- How to POS-tag words automatically?
- How to POS-tag words manually?

Familiar and unfamiliar parts-of-speech

- Familiar parts-of-speech
 - Noun
 - Verb
 - Adjective
 - Preposition
 - Adverb
- What about “five”, “the”, “\$”?

Penn Treebank Tagsets

- CC - coordinating conjunction: and, but
- CD - cardinal number: one, two, three
- DT - determiner: a, the, this, that
- EX - existential there
- FW - foreign word
- IN - preposition or subordinate conjunction
- LS - list marker: firstly, secondly
- To - to
- UH - interjection, uh, oh

What is POS-tagging?

- What is the part-of-speech of:
 - order, rent, present, wonder, protest, wind
- POS-tagging: the process of assigning part-of-speech tags to words in a corpus (automatically or manually)
 - Unusually one part-of-speech per word
- Resolving lexical ambiguity.

What is the use of a POS-tagged corpus?

- For linguistic observations
- Serving as training and test corpus for automatic POS taggers

CC or DT

- Neither/?? he or/CC she likes skiing.
- Neither/?? men like skiing .
- Either/?? Jean or/CC Mary likes singing.
- Either/?? Girl likes singing.
- Both/?? Jack and/CC Tom hates singing .
- Both/?? men hates singing.

CC or DT

- Neither/**CC** he or/**CC** she likes skiing.
- Neither/**DT** men like skiing .
- Either/**CC** Jean or/**CC** Mary likes singing.
- Either/**DT** Girl likes singing.
- Both/**CC** Jack and/**CC** Tom hates singing .
- Both/**DT** men hates singing.

CD or NN

- One/?? of the best reasons
- The only one/?? Of its kind
- The only ones/?? of its kind

CD or NN

- One/**CD** of the best reasons
- The only one/**NN** Of its kind
- The only ones/**NN** of its kind

EX or RB

- There/?? was a party in progress.
- There/?? ensued a melee.
- There/?? , a party was in progress.
- There/?? , ensued a melee.

EX or RB

- There/**EX** was a party in progress.
- There/**EX** ensued a melee.
- There/**RB** , a party was in progress.
- There/**RB** , ensued a melee.

Accessing POS-tagged corpus with NLTK

```
>>> import nltk, re
>>> text = "Neither/CC he/PPS or/CC she/PPS
likes/VB skiing/VBN"
>>> [nltk.tag.str2tuple(t) for t in text.split()]
>>> nltk.corpus.brown.tagged_words()
```

Ex: How do you get a list of words (just words) for a tagged corpus?

Finding the most frequent verbs in a tagged corpus

```
>>> wt = []
>>> brown_a =
    nltk.corpus.brown.tagged_words(categories='a
    ')
>>> for (word, tag) in brown_a:
        if tag[:2] == 'VB':
            wt.append(word + '/' + tag)
>>> fd=nltk.FreqDist(wt)
>>> fd.sorted()[:20]
```

Finding most frequent nouns

```
>>> nouns = []
>>> for (w, t) in brown_a:
        if t[:2] == 'NN':
            nouns.append(w + '/' + t)
>>> nfd = nltk.FreqDist(nouns)
>>> nfd.sorted()[:20]
```

Most frequent tags

```
>>> tags = [tag for (word,tag) in brown_a]
>>> td = nltk.FreqDist(tags)
>>> td.sorted()[:20]
>>> td.max()
```

Tag words automatically

- How would you POS-tag a corpus automatically?

Default tagger: use the most frequent tag

```
>>> tokens = "Thieves leave young athletes in  
the dark".split();  
>>> default_tagger = nltk.DefaultTagger('NN')  
>>> default_tagger.tag(tokens)  
>>> nltk.tag.accuracy(default_tagger,  
nltk.corpus.brown.tagged_sents(categories =  
'a'))
```

Is a default tagger good for anything?