Hauptseminar: Natural Language Processing Tools

NLTK

Daniel Nagel

23. November 2017

POS-Tagging

Part-of-Speech Tagging

In corpus linguistics part-of-speech tagging is the process of marking up a word in a text (corpus) as corresponding to a particular part of speech, based on both its definition and its context—i.e., its relationship with adjacent and related words in a phrase, sentence, or paragraph

Tags

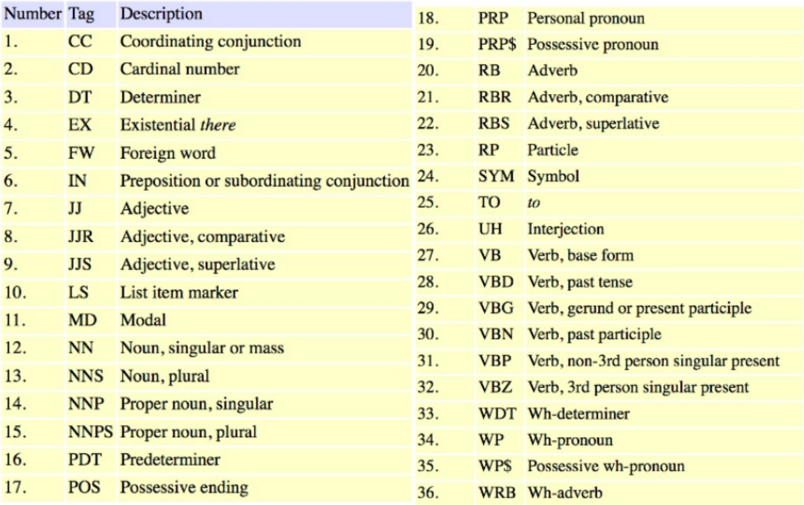
What parsed text corpus is used in NLTK to annotate the tokens?

Tags

What parsed text corpus is used in NLTK to annotate the tokens?

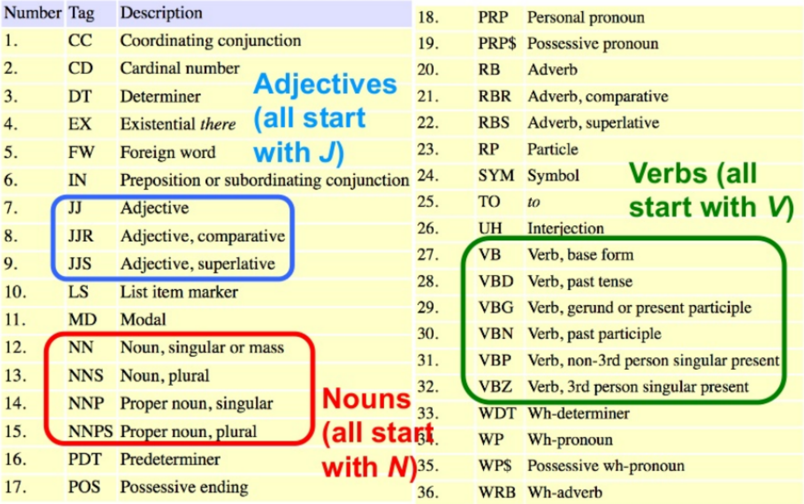
→ The Penn Treebank

The Penn Treebank Tags



Number	Tag	Description
1.	CC	Coordinating conjunction
2.	CD	Cardinal number
3.	DT	Determiner
4.	EX	Existential *there*
5.	FW	Foreign word
6.	IN	Preposition or subordinating conjunction
7.	JJ	Adjective
8.	JJR	Adjective, comparative
9.	JJS	Adjective, superlative
10.	LS	List item marker
11.	MD	Modal
12.	NN	Noun, singular or mass
13.	NNS	Noun, plural
14.	NNP	Proper noun, singular
15.	NNPS	Proper noun, plural
16.	PDT	Predeterminer
17.	POS	Possessive ending
18.	PRP	Personal pronoun
19.	PRP$	Possessive pronoun
20.	RB	Adverb
21.	RBR	Adverb, comparative
22.	RBS	Adverb, superlative
23.	RP	Particle
24.	SYM	Symbol
25.	TO	*to*
26.	UH	Interjection
27.	VB	Verb, base form
28.	VBD	Verb, past tense
29.	VBG	Verb, gerund or present participle
30.	VBN	Verb, past participle
31.	VBP	Verb, non-3rd person singular present
32.	VBZ	Verb, 3rd person singular present
33.	WDT	Wh-determiner
34.	WP	Wh-pronoun
35.	WP$	Possessive wh-pronoun
36.	WRB	Wh-adverb

The Penn Treebank Tags



Number	Tag	Description
1.	CC	Coordinating conjunction
2.	CD	Cardinal number
3.	DT	Determiner
4.	EX	Existential *there*
5.	FW	Foreign word
6.	IN	Preposition or subordinating conjunction
7.	JJ	Adjective
8.	JJR	Adjective, comparative
9.	JJS	Adjective, superlative
10.	LS	List item marker
11.	MD	Modal
12.	NN	Noun, singular or mass
13.	NNS	Noun, plural
14.	NNP	Proper noun, singular
15.	NNPS	Proper noun, plural
16.	PDT	Predeterminer
17.	POS	Possessive ending
18.	PRP	Personal pronoun
19.	PRP$	Possessive pronoun
20.	RB	Adverb
21.	RBR	Adverb, comparative
22.	RBS	Adverb, superlative
23.	RP	Particle
24.	SYM	Symbol
25.	TO	*to*
26.	UH	Interjection
27.	VB	Verb, base form
28.	VBD	Verb, past tense
29.	VBG	Verb, gerund or present participle
30.	VBN	Verb, past participle
31.	VBP	Verb, non-3rd person singular present
32.	VBZ	Verb, 3rd person singular present
33.	WDT	Wh-determiner
34.	WP	Wh-pronoun
35.	WP$	Possessive wh-pronoun
36.	WRB	Wh-adverb

Adjectives (all start with *J*)

Nouns (all start with *N*)

Verbs (all start with *V*)

Does this work?

Overall one can get over 90% correct tokens but only around 60% correctness on a sentences-level.

Getting started

```
import nltk
```

Getting started

```
from nltk import word_tokenize
```

Getting started

variable = word_tokenize(”a sentence”)

Getting started

```
nltk.pos_tag(text)
```

Getting started

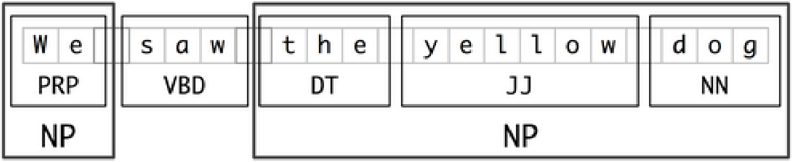
Let’s try it out!

Chunking

Chunking

Chunking (shallow parsing) is an analysis of a sentence which first identifies constituent parts of sentences (nouns, verbs, adjectives, etc.) and then links them to higher order units that have discrete grammatical meanings (noun groups or phrases, verb groups, etc.)

Chunking



Chunking

We will start with a common task in NLP: NP-Chunking

Getting started

```
import nltk
```

Getting started

```
alice = nltk.corpus.gutenberg.words(’carroll-alice.txt’)
```

Getting started

```
posTagged = nltk.pos_tag(alice)
```

Getting started

posTagged = nltk.pos_tag(alice)

Getting started

grammar = ”NP: {⟨DT⟩?⟨JJ⟩*⟨NN⟩}”

Getting started

cp = nltk.RegexpParser(grammar)

Getting started

```
result = cp.parse(posTagged)
```

Getting started

Let's try it out!