Tagging Words
Classification by Part-of-Speech Categories

Behrang QasemiZadeh

me@atmykitchen.info

Goal

- Introduction to the task of **part-of-speech tagging**:
  - What are parts-of-speech, or lexical, categories?
  - How to manipulate words and their part-of-speech in Python ?
  - How to do automatic part-of-speech tagging?

Goal

- Introduction to the fundamental techniques for automatic part-of-speech tagging:
  - Sequence labelling
  - N-gram models
  - Backoff methods
  - Evaluation

Parts-of-Speech

- A part-of-speech (also known as a lexical category, lexical class or word category) is a linguistic category of words:
  - **Nouns**: John, Mary, Behrang, Passau.
  - **Verbs**: read, understand, write, speak.
  - **Adjectives**: good, bad, ugly.
  - ...
- Parts-of-speech are *defined* by the morphosyntactic behaviour of the word in question:
  - In English most nouns are inflected for number with the inflectional plural affix –s.

Part-of-Speech Tagset

- The collection of tags used for the classification of words into lexical categories is known as a tagset.
- Well known tag sets for English are:
  - Part-of-speech tags used in the Penn Treebank Project;
  - Part-of-speech tags used in the CLAWS (the Constituent Likelihood Automatic Word-tagging System) project.

Part-of-Speech Tagset

- Part-of-speech tags used in the Penn Treebank Project:
  - 36 tags.
  - For instance, Penn tag set has 4 different tags for distinguishing nouns and 6 tags for distinguishing verbs.
  - https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html

NN	Noun, singular or mass
NNS	Noun, plural
NNP	Proper noun, singular
NNPS	Proper noun, plural

VB	Verb, base form
VBD	Verb, past tense
VBG	Verb, gerund or present participle
VBN	Verb, past participle
VBP	Verb, non-3rd person singular present
VBZ	Verb, 3rd person singular present

Part-of-Speech Tagset

- Part-of-speech tags used in the recent version of CLAWS:
  - http://ucrel.lancs.ac.uk/claws7tags.html
  - CLAWS7 Tagset has 137 tags.
  - For instance, CLAWS7 tagset has 22 different tags for distinguishing nouns and many more for distinguishing verbs.

ND1	singular noun of direction (e.g. north, southeast)
NN	common noun, neutral for number (e.g. sheep, cod, headquarters)
NN1	singular common noun (e.g. book, girl)
NN2	plural common noun (e.g. books, girls)
NNA	following noun of title (e.g. M.A.)
NNB	preceding noun of title (e.g. Mr., Prof.)
NNL1	singular locative noun (e.g. Island, Street)
NNL2	plural locative noun (e.g. Islands, Streets)
NNO	numeral noun, neutral for number (e.g. dozen, hundred)
NNO2	numeral noun, plural (e.g. hundreds, thousands)
NNT1	temporal noun, singular (e.g. day, week, year)
NNT2	temporal noun, plural (e.g. days, weeks, years)
NNU	unit of measurement, neutral for number (e.g. in, cc)
NNU1	singular unit of measurement (e.g. inch, centimetre)
NNU2	plural unit of measurement (e.g. ins., feet)
NP	proper noun, neutral for number (e.g. IBM, Andes)
NP1	singular proper noun (e.g. London, Jane, Frederick)
NP2	plural proper noun (e.g. Browns, Reagans, Koreas)
NPD1	singular weekday noun (e.g. Sunday)
NPD2	plural weekday noun (e.g. Sundays)
NPM1	singular month noun (e.g. October)
NPM2	plural month noun (e.g. Octobers)

Part-of-Speech Tagging

- **Part-of-speech tagging**, **PoS tagging**, or simply **tagging** is the process of classifying words into **parts-of-speech.**
- PoS tagging thus is the process of labelling words using a predefined tagset.
- A **part-of-speech tagger**, or **PoS tagger**, processes a sequence of words, and attaches a part of speech tag to each word.

PoS tagging in Python

```
>>> import nltk
>>> text = nltk.word_tokenize("And now for something completely different")
```

PoS tagging in Python

```
>>> import nltk
>>> text = nltk.word_tokenize("And now for something completely different")
>>> text
['And', 'now', 'for', 'something', 'completely', 'different']
```

PoS tagging in Python

```
>>> import nltk
>>> text = nltk.word_tokenize("And now for something completely different")
>>> text
['And', 'now', 'for', 'something', 'completely', 'different']
>>> nltk.pos_tag(text)
```

PoS tagging in Python

```
>>> import nltk
>>> text = nltk.word_tokenize("And now for something completely different")
>>> text
['And', 'now', 'for', 'something', 'completely', 'different']
>>> nltk.pos_tag(text)
[('And', 'CC'), ('now', 'RB'), ('for', 'IN'), ('something', 'NN'),
('completely', 'RB'), ('different', 'JJ')]
>>>
```

PoS tagging in Python

```
>>> import nltk
>>> text = nltk.word_tokenize("And now for something completely different")
>>> text
['And', 'now', 'for', 'something', 'completely', 'different']
>>> nltk.pos_tag(text)
[('And', 'CC'), ('now', 'RB'), ('for', 'IN'), ('something', 'NN'),
('completely', 'RB'), ('different', 'JJ')]
>>>
```

PoS tagging in Python

```
>>> import nltk
>>> text = nltk.word_tokenize("And now for something completely different")
>>> text
['And', 'now', 'for', 'something', 'completely', 'different']
>>> nltk.pos_tag(text)
[('And', 'CC'), ('now', 'RB'), ('for', 'IN'), ('something', 'NN'),
('completely', 'RB'), ('different', 'JJ')]
>>>
```

CC	Coordinating conjunction
RB	Adverb
IN	Preposition
NN	Noun, singular or mass
JJ	Adjective

Distinguishing homographs by Parts-of-Speech

- Part-of-Speech tags can be used to distinguish words spelled the same but not necessarily pronounced the same and having different meanings:

Distinguishing homographs by Parts-of-Speech

- Part-of-Speech tags can be used to distinguish words spelled the same but not necessarily pronounced the same and having different meanings:

```
>>> text = nltk.word_tokenize("They refuse to permit us to obtain the refuse permit")
>>> text ['They', 'refuse', 'to', 'permit', 'us', 'to', 'obtain', 'the', 'refuse',
 'permit']
>>> nltk.pos_tag(text)
[('They', 'PRP'), ('refuse', 'VBP'), ('to', 'TO'), ('permit', 'VB'), ('us', 'PRP'),
('to', 'TO'),('obtain', 'VB'), ('the', 'DT'), ('refuse', 'NN'), ('permit', 'NN')]
>>>
```

Distinguishing homographs by Parts-of-Speech

- Part-of-Speech tags can be used to distinguish words spelled the same but not necessarily pronounced the same and having different meanings:

```
>>> text = nltk.word_tokenize("They refuse to permit us to obtain the refuse permit")
>>> text ['They', 'refuse', 'to', 'permit', 'us', 'to', 'obtain', 'the', 'refuse',
 'permit']
>>> nltk.pos_tag(text)
[('They', 'PRP'), ('refuse', 'VBP'), ('to', 'TO'), ('permit', 'VB'), ('us', 'PRP'),
('to', 'TO'),('obtain', 'VB'), ('the', 'DT'), ('refuse', 'NN'), ('permit', 'NN')]
>>>
```

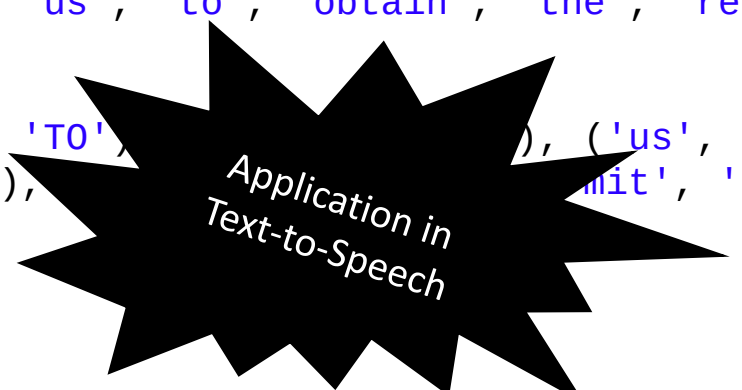
verb \ri-ˈfyüz\

noun \ˈre-ˌfyüs, -ˌfyüz\

Distinguishing homographs by Parts-of-Speech

- Part-of-Speech tags can be used to distinguish words spelled the same but not necessarily pronounced the same and having different meanings:

```
>>> text = nltk.word_tokenize("They refuse to permit us to obtain the refuse permit")
>>> text ['They', 'refuse', 'to', 'permit', 'us', 'to', 'obtain', 'the', 'refuse',
 'permit']
>>> nltk.pos_tag(text)
[('They', 'PRP'), ('refuse', 'VBP'), ('to', 'TO'), ('us', 'PRP'),
('to', 'TO'),('obtain', 'VB'), ('the', 'DT'), ('refuse', 'NN')]
>>>
```



Representing Tagged Tokens

- What data structure do you suggest for storing part-of-speech tagged text?

Representing Tagged Tokens

- What data structure do you suggest for storing part-of-speech tagged text?
  - A tuple consisting of the token and the tag.

```
>>> tagged_token = nltk.tag.str2tuple('fly/NN')
>>> tagged_token
('fly', 'NN')
>>> tagged_token[0]
 'fly'
>>> tagged_token[1]
'NN'
>>>
```

Reading Tagged Corpora

- Several of the corpora included with NLTK have been tagged for their part-of-speech.

Reading Tagged Corpora

- Several of the corpora included with NLTK have been tagged for their part-of-speech.
- These corpora use a variety of formats for storing part-of-speech tags.

Reading Tagged Corpora

- Several of the corpora included with NLTK have been tagged for their part-of-speech.
- These corpora use a variety of formats for storing part-of-speech tags.
- NLTK, however, provides a uniform interface to tagged corpora.

Reading Tagged Corpora

- Several of the corpora included with NLTK have been tagged for their part-of-speech.
- These corpora use a variety of formats for storing part-of-speech tags.
- NLTK, however, provides a uniform interface to tagged corpora.

```
>>> nltk.corpus.brown.tagged_words()
[(u'The', u'AT'), (u'Fulton', u'NP-TL'), ...]
>>> nltk.corpus.conll2000.tagged_words()
[(u'Confidence', u'NN'), (u'in', u'IN'), ...]
>>> nltk.corpus.nps_chat.tagged_words()
[(u'now', 'RB'), (u'im', 'PRP'), (u'left', 'VBD'), ...]
>>>
```

Reading Tagged Corpora

- Several of the corpora included with NLTK have been tagged for their part-of-speech.
- These corpora use a variety of formats for storing part-of-speech tags.
- NLTK, however, provides a uniform interface to tagged corpora.

```
>>> nltk.corpus.brown.tagged_words()
[(u'The', u'AT'), (u'Fulton', u'NP-TL'), ...]
>>> nltk.corpus.conll2000.tagged_words()
[(u'Confidence', u'NN'), (u'in', u'IN'), ...]
>>> nltk.corpus.nps_chat.tagged_words()
[(u'now', 'RB'), (u'im', 'PRP'), (u'left', 'VBD'), ...]
>>>
```

Mapping Tagsets Using Python

```
>>> nltk.corpus.conll2000.tagged_words()
[(u'Confidence', u'NN'), (u'in', u'IN'), ...]
>>> nltk.corpus.conll2000.tagged_words(tagset='universal')
[(u'Confidence', u'NOUN'), (u'in', u'ADP'), ...]
>>>
```

Mapping Tagsets Using Python

```
>>> nltk.corpus.conll2000.tagged_words()
[(u'Confidence', u'NN'), (u'in', u'IN'), ...]
>>> nltk.corpus.conll2000.tagged_words(tagset='universal')
[(u'Confidence', u'NOUN'), (u'in', u'ADP'), ...]
>>>
```

```
>>> nltk.corpus.brown.tagged_words()
[(u'The', u'AT'), (u'Fulton', u'NP-TL'), ...]
>>> nltk.corpus.brown.tagged_words(tagset='universal')
[(u'The', u'DET'), (u'Fulton', u'NOUN'), ...]
```

Mapping Tagsets Using Python

```
>>> nltk.corpus.conll2000.tagged_words()
[(u'Confidence', u'NN'), (u'in', u'IN'), ...]
>>> nltk.corpus.conll2000.tagged_words(tagset='universal')
[(u'Confidence', u'NOUN'), (u'in', u'ADP'), ...]
>>>
```

```
>>> nltk.corpus.brown.tagged_words()
[(u'The', u'AT'), (u'Fulton', u'NP-TL'), ...]
>>> nltk.corpus.brown.tagged_words(tagset='universal')
[(u'The', u'DET'), (u'Fulton', u'NOUN'), ...]
```

```
>>> nltk.corpus.sinica_treebank.tagged_words()
[(u'\u4e00', u'Neu'), (u'\u53cb\u60c5', u'Nad'), ...]
>>> nltk.corpus.sinica_treebank.tagged_words(tagset='universal')
```

Mapping Tagsets Using Python

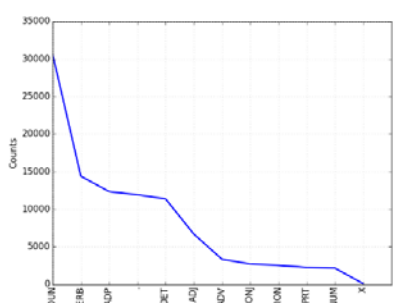
```
>>> nltk.corpus.conll2000.tagged_words()
[(u'Confidence', u'NN'), (u'in', u'IN'), ...]
>>> nltk.corpus.conll2000.tagged_words(tagset='universal')
[(u'Confidence', u'NOUN'), (u'in', u'ADP'), ...]
>>>
```

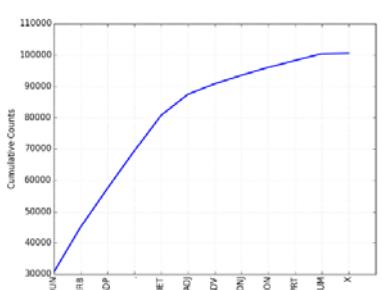
```
>>> nltk.corpus.brown.tagged_words()
[(u'The', u'AT'), (u'Fulton', u'NP-TL'), ...]
>>> nltk.corpus.brown.tagged_words(tagset='universal')
[(u'The', u'DET'), (u'Fulton', u'NOUN'), ...]
```

```
>>> nltk.corpus.sinica_treebank.tagged_words()
[(u'\u4e00', u'Neu'), (u'\u53cb\u60c5', u'Nad'), ...]
>>> nltk.corpus.sinica_treebank.tagged_words(tagset='universal')
```

Quiz/Exercise !

- Build a frequency profile of part-of-speech tags in Brown corpus and visualize it in a plot.
  - Hints:
    - Look page 184 of the book for accessing the “UNIVERSAL” part-of-speech tags frequencies.
    - Look page 168 for the visualization of frequencies in a bar chart.
    - Use structured programming so that we can pass other corpus than brown.





Quiz2/Exercise2 !

- Compare the frequencies of part-of-speech tags of words for different text genres in brown corpus.
  - Hint:
    - Remember that you can tell python what genre your are looking for using the categories=X!

```
>>> brown_news_tagged = nltk.corpus.brown.tagged_words(categories='news', tagset='universal')
>>> brown_news_tagged
[(u'The', u'DET'), (u'Fulton', u'NOUN'), ...]
>>> brown_fiction_tagged = nltk.corpus.brown.tagged_words(categories='fiction', tagset='universal')
>>> brown_fiction_tagged
[(u'Thirty-three', u'NUM'), (u'Scotty', u'NOUN'), ...]
>>>
```

A Closer look at part-of-speeches

- Nouns
  - Nouns generally refer to people, places, things, or concepts, e.g., woman, Scotland.
  - Nouns show certain grammatical patterns, e.g.:
    - they can appear after determiners and adjectives.
    - They can be the subject or object of the verb.

A Closer look at part-of-speeches

- Verbs
  - Verbs are lexical items that describe events and actions, e.g., fly and essen!
  - Verbs express relations involving the referents of one or more noun phrases.
  - Verbs can appear in several forms, e.g. gerund, past participle, etc.

A Closer look at part-of-speeches

- **Adjectives**
  - Adjectives describe nouns.
  - They can be used as modifiers, e.g. large in a large pizza.
  - They can be used as predicates, e.g. the pizza is large.
  - English adjectives can be derived using certain patterns (linguistically speaking through a morphological process)
    - fly+ing in the flying birds

A Closer look at part-of-speeches

- **Adverbs**
  - Adverbs modify verbs to specify manner, time, place. e.g., fast in those birds fly fast or quickly in the brown fox jumped very quickly.
  - Adverbs may also modify adjectives really in brown fox jumped really quickly.
  - You can find derivational morphology patterns also for adverbs:
    - In English adjectives/noun + ly gives an adverb, e.g. quick+ly = quickly.

A Closer look at part-of-speeches

- English has several other closed lexical categories:
  - **prepositions,** such as in, for, of.
  - **articles** (also called determiners) e.g., the, a.
  - **modals** e.g., should, may.
  - **Personal pronouns** e.g., she, they.
  - **…**
- Various classification for lexical categories can be found, e.g. CLAWS vs. Penn PoS tagset.

Quiz

- Tag the following sentence for parts-of-speech:

  The head of the UN mission charged with fighting Ebola tells the BBC he does not yet have the resources necessary to defeat the deadly disease.

Quiz

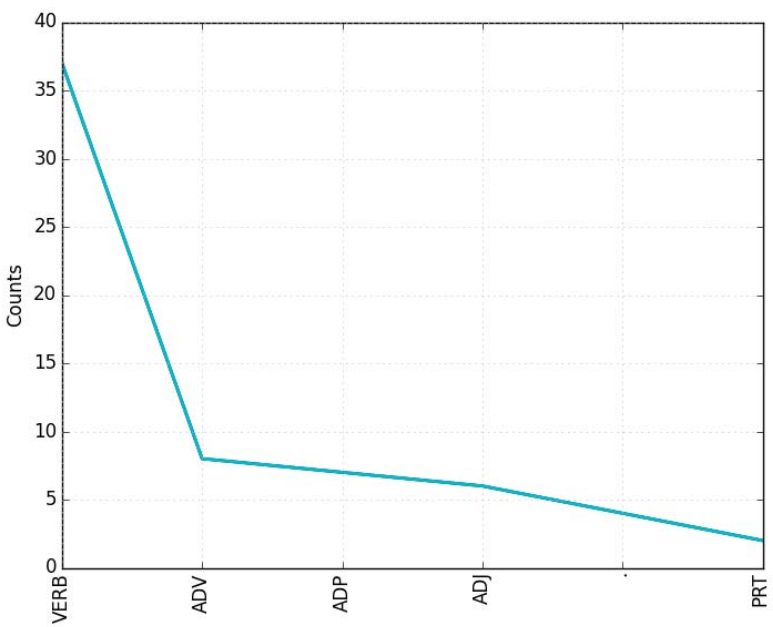
- Tag the following sentence for parts-of-speech:

  The head of the UN mission charged with fighting Ebola tells the BBC he does not yet have the resources necessary to defeat the deadly disease.

  DT/The NN/head IN/of DT/the NNP/UN NN/mission VBN/charged IN/with VBG/fighting NNP/Ebola VBZ/tells DT/the NNP/BBC PRP/he VBZ/does RB/not RB/yet VB/have DT/the NNS/resources JJ/necessary TO/to VB/defeat DT/the JJ/deadly NN/disease ./.

Exploring tagged corpora

```
>>> brown_lrnd_text = nltk.corpus.brown.tagged_words(
                  categories='learned‘, tagset='universal')
>>> tags = [b[1] for (a, b)
      in nltk.bigrams(brown_
>>> fd = nltk.FreqDist(tags)
>>> fd.tabulate()
VERB ADV ADP ADJ . PRT
37 8 7 6 4 2
>>> fd.plot()
```



Quiz – Exploring Word Context

- Use the `similar()` function (discussed in the previous session) to explore words of certain part of speech.

```
>>> text = nltk.Text(word.lower() for word in
      nltk.corpus.brown.words())
>>> text.similar('woman')
>>> text.similar('bought')
>>> text.similar('over’)
```

What do you recon from this experiment?