103_pos_tagging

December 15, 2018

1 POS (Parts of Speech) tagging

POS tagging labels words by type of word, which may enhance the quality of information that may be extracted from a piece of text.

There are varying sets of tags, but the common universal set is:

ADJ: adjective ADP: adposition (preopositions and postpositions) ADV: adverb AUX: auxiliary CCONJ: coordinating conjunction DET: determiner INTJ: interjection NOUN: noun NUM: numeral PRT: particle or other function words PRON: pronoun VERB: verb X: other .: Punctuation

Other, more granular sets of tags include those included in the Brown Corpus (a coprpus of text with tags). NLTK can convert more granular data sets to tagged sets.

The two most commonly used tagged corpus datasets in NLTK are Penn Treebank and Brown Corpus. Both take text from a wide range of sources and tag words.

Details of the brown corpus and Penn treebank tags may be found here:

- https://en.wikipedia.org/wiki/Brown_Corpus
- http://clu.uni.no/icame/manuals/BROWN/INDEX.HTM
- https://www.sketchengine.eu/penn-treebank-tagset/

1.1 An example of tagging from the Brown corpus, and conversion to the universal tag set

```
In [1]: import nltk
        # Download the brown corpus if it has not previously been downloaded
        nltk.download('brown');
```

```
[nltk_data] Downloading package brown to /home/michael/nltk_data...
[nltk_data]   Package brown is already up-to-date!
```

```
In [2]: from nltk.corpus import brown
        # Show a set of tagged words from the Brown corpus
        print(brown.tagged_words()[20:40])
```

```
[('any', 'DTI'), ('irregularities', 'NNS'), ('took', 'VBD'), ('place', 'NN'), ('.', '.'), ('The
```

Convert more granular brown tagging to universal tagging.

```
In [3]: print(brown.tagged_words(tagset='universal')[20:40])
```

```
[('any', 'DET'), ('irregularities', 'NOUN'), ('took', 'VERB'), ('place', 'NOUN'), ('.', '.'),
```

Details of the brown corpus tags may be found here:
https://en.wikipedia.org/wiki/Brown_Corpus

In the above example the brown tags NNS (plural noun), NN (singlular noun) and NN-TL (singluar noun found in a title) are all converted to the universal tag NOUN.

1.2 Use of tagging to distinguish between different meanings of the same word

Consider the two uses of the word 'left' in the sentence below:

```
In [4]: text = "I left the hotel to go to the coffee shop which is on the left of the church"
```

Let's look at how 'left' is tagged in the two sentences:

```
In [5]: # Split text into words
        tokens = nltk.word_tokenize(text)

        print ('Word tags for text:', nltk.pos_tag(tokens, tagset="universal"))
```

```
Word tags for text: [('I', 'PRON'), ('left', 'VERB'), ('the', 'DET'), ('hotel', 'NOUN'), ('to'
```

'The first use of 'left' has been identified as a verb, and the second use a noun.

So POS-tagging may be used to enhance simple text-based methods, by providing additional information about words taking into account the context of the word.