Theory: POS tagging

18 minutes 0 / 5 problems solved

§1. What is POS-tagging?

To start with, we need to define what a **part of speech** is. Let's go back to elementary school. We know that all words can be divided into different classes, such as, for example, adverbs, adjectives, nouns, and verbs. These categories are called **parts of speech** and it's a basic concept for natural language processing.

So, part-of-speech tagging is the process of classifying words into such lexical categories. Sometimes, it is also called **word-category disambiguation** or **grammatical tagging**.

POS-tagging is a very important NLP procedure. It is a step of text preprocessing, and its results may be used to improve the performance of various NLP tasks, such as lemmatization (a procedure that reduces word forms to one root form), semantic analysis of texts, machine translation, and many others.

§2. POS-tagging in NLTK

POS-tags are assigned to words with the help of a POS-tagger. But how does it work? There are many ways in different libraries to label words with a POS-tag, but in this topic, we will learn how to do it with NLTK for English.

> NLTK functionality for different languages varies, so when you work with languages other than English, it may be necessary to find some other library for POS-tagging.

NLTK uses a pre-trained model for POS-tagging. It works with a tokenized text only, so you need to split your text into tokens beforehand. You can see how NLTK tags one sentence below:

```
import nltk


nltk.download('averaged_perceptron_tagger')  # download the tagger


text = ['there', 'is', 'a', 'dwarf', 'looking', 'out', 'of', 'the', 'window', '!']
print(nltk.pos_tag(text))

# [('there', 'EX'), ('is', 'VBZ'), ('a', 'DT'), ('dwarf', 'NN'), ('looking', 'VBG'
), ('out', 'IN'), ('of', 'IN'), ('the', 'DT'), ('window', 'NN'), ('!', '.')]
```

Great! We converted the sentence into a list of tuples consisting of words and tags. Now, we need to read tags so let's learn more about it.

§3. POS-tags

A set of tags is called a **tagset**. The way we classify words into such categories may depend on the particular task at hand and a particular system of notation, so tagsets may differ. Bear in mind that tagsets may vary depending on the language you work with. Here you can see some examples of tagsets for the English language. In general, for English, NLTK uses Penn Treebank's POS tagset, one of the most popular. The tagsets are called based on the text resources they are used at; below is a list of them with brief descriptions:

- Penn Treebank – the first large-scale treebank (developed in the LINK Laboratory, University of Pennsylvania);
- Brown Corpus – The Brown University Standard Corpus of Present-Day American English (contains 500 samples of English-language text);

Current topic:

POS tagging

Topic depends on:

- ✓ Tuple
- ✓ For loop 12 Stage 1
- ✗ Tokenization

Topic is required for:

Text normalization

Table of contents:

- LOB Corpus – Lancaster-Oslo-Bergen Corpus (a million-word collection of British English texts);
- BNC – British National Corpus (a 100-million-word text corpus of samples of written and spoken English).

You can see how tagsets vary in these systems in the table below:

POS	Penn Treebank	Brown Corpus	LOB Corpus	BNC2	Examples
Adjective	JJ	JJ	JJ	AJ0	gorgeous, attractive
Adjective, comparative	JJR	JJR	JJR	AJC	better
Superlative adjective	JJS	JJS	JJT	AJS	best
The -ing form of the verb	VBG	VBG	VBG	VHG	yelling, playing
Reflexive pronoun	-	-	PPL\PPLS	PNX	herself, myself
The present tense forms of the verb	VBP\VBZ	VBP\VBZ	VBZ (3rd person singular)	VBB	see, goes
Verb, the base form	VB	VB	VB	VDB	drive, get

Look at the first example in our sentence. There are some elements such as a gerund '*looking*' or a 3rd person singular present verb '*is*' which are verbs, but they have their tags. Such tags are called *unsimplified,* most of them have suffixes that represent different variants of usual tags, for example, *VBZ* for 3rd person singular present or *VBG* for present participle.

You can find more information about tags from the Penn Treebank that are used in NLTK by typing `nltk.help.upenn_tagset()`: a complete list of tags with their explanations and examples will be shown. You can also get information about a particular tag by specifying it in the parentheses:

```
nltk.help.upenn_tagset('VBG')
# VBG: verb, present participle or gerund
#    telegraphing stirring focusing angering judging stalling lactating
#    hankerin alleging veering capping approaching traveling besieging
#    encrypting interrupting erasing wincing ...
```

§4. POS-disambiguation

POS-tagging is not that easy all the time. Sometimes, the same form of a word, for example, '*play*', may be a noun or a verb, depending on the context. This is called **ambiguity**, meaning that the tag we should assign to the word is ambiguous in such cases. The process of resolving the ambiguity, understanding which tag is correct, is called **disambiguation**.

Have a look at this example:

```

sentence = ['My', 'peer', 'will', 'peer', 'through', 'the', 'window', 'hole', '.']
  print(nltk.pos_tag(sentence))

# [('My', 'PRP$'), ('peer', 'NN'), ('will', 'MD'), ('peer', 'VB'), ('through', 'IN
'), ('the', 'DT'), ('window', 'NN'), ('hole', 'NN'), ('.', '.')]
```

As you can see, *'peer'* comes up first as a noun and then as a verb. Both these words are the same, but NLTK labels them differently. How does this happen?

The NLTK tagger deals with ambiguity with the help of the context: trained on a huge tagged corpus, it is capable of analyzing the distribution of words in a text, so now, if you put '*to*' before '*peer*', the tagger will label it as a verb, and similarly with '*the*' before nouns.

However, the result of POS-tagging in NLTK is not always perfect. Let's find out what happens if we give the tagger a weird but possible sentence *'The old man the boats.'*:

```
[('The', 'DT'), ('old', 'JJ'), ('man', 'NN'), ('the', 'DT'), ('boats', 'NNS'),
('.', '.')]
```

As expected, NLTK defined '*man*' as a noun and it seems like we don't have a verb here. However, the word *'man'* here is a verb meaning to take one's place for service, and *'old'* is used to mean a group of old people.

In tricky sentences like this one, sentences with errors, or just ambiguous ones, NLTK's tagger can make some mistakes, and you need to bear that in mind.

§5. Summary

At this point, we have familiarized ourselves with one of the many possible ways to complete POS-tagging. There are many kinds of POS-tagging implemented in other libraries, but here we have learned the fundamental approach using NLTK.

You have learned about:

- tagging a text with NLTK;
- different parts of speech, their tags, and tagsets;
- POS-ambiguity and how NLTK deals with it.

Now it is time to practice!

Report a typo

14 users liked this theory. **0** didn't like it. **What about you?**

Start practicing

Comments (0) Hints (0) Useful links (0) Show discussion