NLTK Introduction

with brief NLP tools comparison

By Rui Mao

NetId: ruim6

Introduction

NLTK is an open-source free python toolkit for natural language processing.

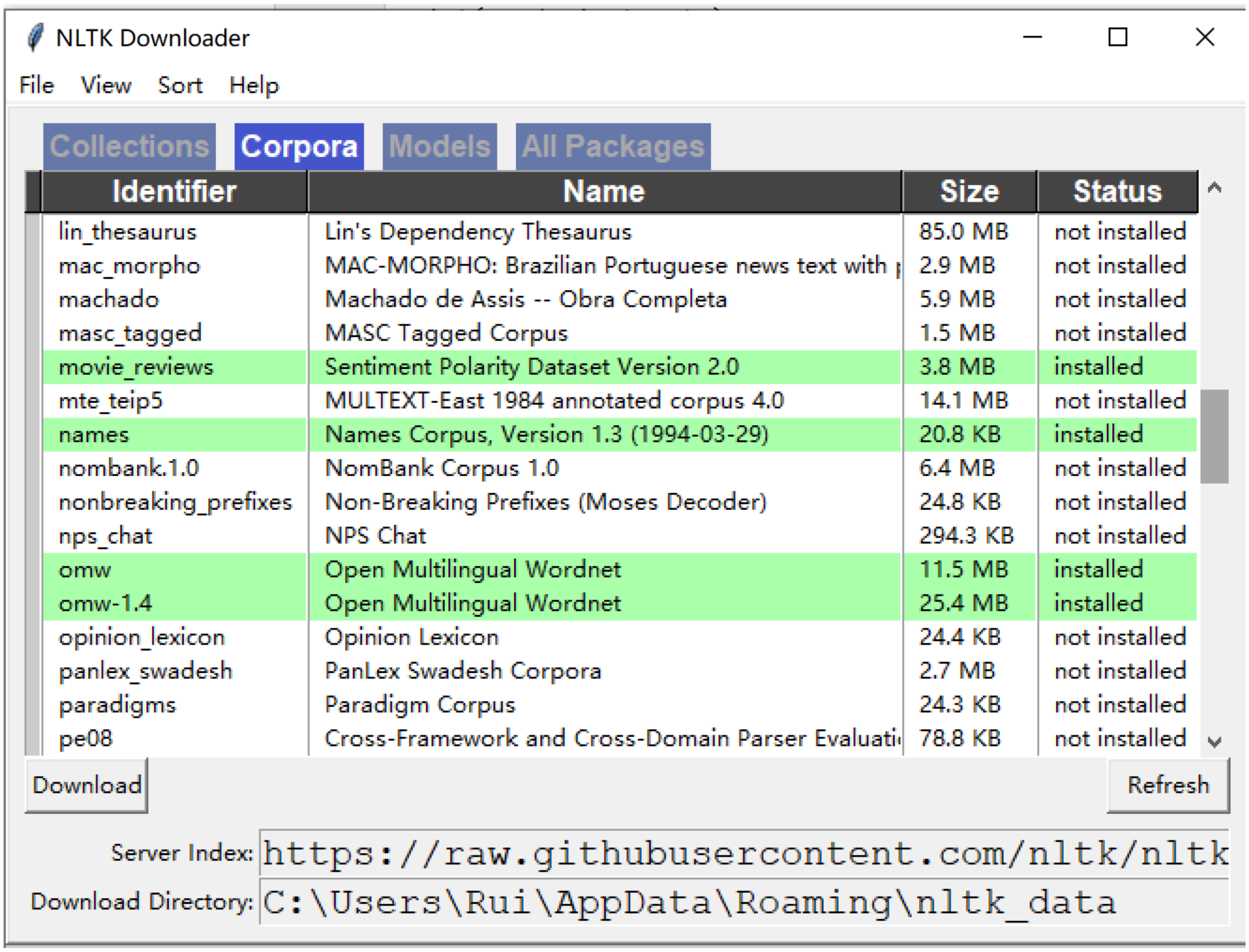
The full name of the NLTK is Natural Language Toolkit. It was developed by Steven Bird and Edward Loper of Univ of Pennsylvania.[3] NLTK was initially released in 2001 [2], and it was written in Python. It is mainly used for the English language. (Although it contains packages for other languages like Japanese and Indian)

NLTK provides over 50 corpora, it also has a lot of essential text-processing libraries for tagging, stemming, tokenization, parsing, and classification. [6]

The latest version of the NLTK upon writing this article is NLTK 3.7, released on Feb 2022. This version does not support python 3.6 and adds support for python 3.10. [7]

NLTK Corpora

After installing and importing NLTK package, you can check the nltk downloader by calling nltk.download(). From there you can check and install the corpora like movie_reviews shown in Fig.1 below:



Identifier	Name	Size	Status
lin_thesaurus	Lin's Dependency Thesaurus	85.0 MB	not installed
mac_morpho	MAC-MORPHO: Brazilian Portuguese news text with	2.9 MB	not installed
machado	Machado de Assis -- Obra Completa	5.9 MB	not installed
masc_tagged	MASC Tagged Corpus	1.5 MB	not installed
movie_reviews	Sentiment Polarity Dataset Version 2.0	3.8 MB	installed
mte_teip5	MULTEXT-East 1984 annotated corpus 4.0	14.1 MB	not installed
names	Names Corpus, Version 1.3 (1994-03-29)	20.8 KB	installed
nombank.1.0	NomBank Corpus 1.0	6.4 MB	not installed
nonbreaking_prefixes	Non-Breaking Prefixes (Moses Decoder)	24.8 KB	not installed
nps_chat	NPS Chat	294.3 KB	not installed
omw	Open Multilingual Wordnet	11.5 MB	installed
omw-1.4	Open Multilingual Wordnet	25.4 MB	installed
opinion_lexicon	Opinion Lexicon	24.4 KB	not installed
panlex_swadesh	PanLex Swadesh Corpora	2.7 MB	not installed
paradigms	Paradigm Corpus	24.3 KB	not installed
pe08	Cross-Framework and Cross-Domain Parser Evaluati	78.8 KB	not installed

Server Index: https://raw.githubusercontent.com/nltk/nltk

Download Directory: C:\Users\Rui\AppData\Roaming\nltk_data

Fig 1. NLTK Downloader shows it's corpora collection

From the downloader, we can see NLTK actually has about 90 corpora now, almost double what they asserted on the official website, which means NLTK is still developing at a fast pace after 20 years of development. This vast collection of corpora makes it a very convenient NLP tool. From the corpora collection, I'm especially interested what movie review corpora they used, NLTK uses sentiment polarity dataset version 2.0 which is developed by Cornell university. This corpora involves 2000 movie reviews. [12] We can see it is a 3.8 MB database which is perfect for lightweight NLP tasks.

I also found that NLTK has corpora for web chat, like *webtext* and *nps_chat*, which are very useful for sentiment analysis as we know people tend to use slang for movie or product reviews nowdays.

NLTK basic sentiment analysis

NLTK is good for sentiment analysis; in the simplest case, we can use nltk.classify.apply_features to get the training set. After that, we can create a classifier like naïve bayes by using nltk.NaiveBayesClassifier.train() function**.**

One best part of NLTK is it involves a lot of statistical functions, for example, we can use classifier.show_most_informative_features to get statistically most informative features. [4]

Major difference with other NLP tools:

SpaCy:

SpaCy is generally faster than NLTK, it is usually for "Opinionated NLP tasks" [5]. Another major difference is that it has commercial usage.

SpaCy has build-in pretrained transformers.[8]

OpenNLP:

One major difference is that OpenNLP is in Java, or a sort of Java Version of NLTP.[5]

It is good for some basic NLP tasks and supports commercial usage.[9]

Scikit-learn:

Scikit-learn is also a widely used NLP tool, one advantage is that it has large support and is based on NumPy and SciPy, it is mainly used for predictive data analysis [10]. It is usually complex for beginners.

PyTorch:

For both research and commercial use, powerful machine learning tool for NLP tasks, it has distributed training backend so it is usually fast. [11]

Conclusion

This article briefly introduced NLTK, what it is, and a simple example of how to use it. Unlike other NLP tools, NLTK is mainly used for educational purposes. However, it has an extensive corpora collection and many tools for basic NLP tasks.

Most importantly, it is open source which makes it an excellent tool for teaching and learning natural language processing.

Reference

[1] https://www.nltk.org/book/ch00.html

[2] https://sourceforge.net/projects/nltk/

[3]https://en.wikipedia.org/wiki/Natural_Language_Toolkit#:~:text=It%20was%20developed%20by%20Steven,graphical%20demonstrations%20and%20sample%20data.

[4] https://juejin.cn/post/6844904149268561928

[5] https://www.langnerd.com/natural-language-processing-libraries/

[6] https://www.nltk.org/

[7] https://github.com/nltk/nltk

[8] https://spacy.io/

[9] https://opennlp.apache.org/

[10] https://scikit-learn.org/stable/

[11] https://pytorch.org/

[12] https://www.cs.cornell.edu/people/pabo/movie-review-data/