



Text Analysis with NLTK

March 30-April 8, 2022

Week 2

Today's Topics

Questions from previous classes

Regular Expression practice

Cleaning messy text

Understanding corpora NLTK is built on

Questions

`similar("input_word")`

- Looks at words on immediate left and right of input word
- Finds other words that are immediately surrounded by the same words
- Example:

```
text2.similar("good")  
...the good opinion...
```

Understanding the similar() method:

```
t = "...the good opinion...oh my goodness...your good advice...how good for  
them...the strong opinion...your misguided advice...advise me, please..."
```

```
t = "...the good opinion...oh my goodness...your good advice...how good for  
them...the strong opinion...your misguided advice...advise me, please..."
```

```
t.similar("good")
```

```
>> strong misguided
```

Assignment

How did it go?

Questions?

Regular Expressions (RegEx)

Pattern matching for the string (`str`) data type

Documentation:

Intro: <https://docs.python.org/3/howto/regex.html#regex-howto>

Detailed docs: <https://docs.python.org/3/library/re.html>

For practice: pythex.org

Check out the cheat sheet!

Regular Expressions (RegEx)

To use in a Jupyter Notebook:

```
import re
```

To find patterns (2 ways):

```
re.findall("regex_pattern", "string_to_search")
```

```
>> ["list", "of", "all", "matches", "found"]
```

```
pattern = re.compile("regex_pattern")
```

```
p.findall("string_to_search")
```

```
>> ["list", "of", "all", "matches", "found"]
```

Cleaning Messy Text

Regular Expressions

`s.strip()` - remove leading & trailing whitespace or input characters in `s`

`s.replace('a', 'b')` - replace `a` with `b` in string `s`

Remember, digitization is imperfect

- Includes OCR (optical character recognition)
- Includes HWT or HRT (handwriting recognition)



Be careful not to spend all your time cleaning data! It may be useful to time-box this task so that you do not run out of time for the actual analysis work.

Alternatively, you could consider investing in or applying for funds to manually correct your text. This will yield more accurate results than programmatic methods.

LET'S CODE!

Understanding NLTK: Tokenization

Whitespace	Isn't	Ahab,	Ahab?	;))					
Treebank	Is	n't	Ahab	,	Ahab	?	;)	
Tweet	Isn't	Ahab	,	Ahab	?	;))			
TokTok (Dehdari, 2014)	Isn	'	t	Ahab	,	Ahab	?	;)

Figure 4.1: The output of four NLTK tokenizers, applied to the string *Isn't Ahab, Ahab? ;)*

Reference: Jacob Eisenstein (2018) Chapter 4, Natural Language Processing.

Understanding NLTK

Original	The	Williams	sisters	are	leaving	this	tennis	centre
Porter stemmer	the	william	sister	are	leav	thi	tenni	centr
Lancaster stemmer	the	william	sist	ar	leav	thi	ten	cent
WordNet lemmatizer	The	Williams	sister	are	leaving	this	tennis	centre

Figure 4.2: Sample outputs of the Porter (1980) and Lancaster (Paice, 1990) stemmers, and the WORDNET lemmatizer

Reference: Jacob Eisenstein (2018) Chapter 4, Natural Language Processing.

Understanding NLTK: WordNet

What is WordNet

People

News

Use Wordnet Online

Download

Citing WordNet

License and Commercial Use

Related Projects

Documentation

Publications


Frequently Asked Questions

What is WordNet?

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the creators of WordNet and do not necessarily reflect the views of any funding agency or Princeton University.

When writing a paper or producing a software application, tool, or interface based on WordNet, it is necessary to properly [cite the source](#). Citation figures are critical to WordNet funding.

About WordNet

WordNet® is a large lexical database of English. Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept. Synsets are interlinked by means of conceptual-semantic and lexical relations. The resulting network of meaningfully related words and concepts can be navigated with the [browser](#) . WordNet is also freely and publicly available for [download](#). WordNet's structure makes it a useful tool for computational linguistics and natural language processing.

WordNet superficially resembles a thesaurus, in that it groups words together based on their meanings. However, there are some important distinctions. First, WordNet interlinks not just word forms—strings of letters—but specific senses of words. As a result, words that are found in close proximity to one another in the network are semantically disambiguated. Second, WordNet

Note

Due to funding and staffing issues, we are no longer able to accept comment and suggestions.

We get numerous questions regarding topics that are addressed on our [FAQ](#) page. If you have a problem or question regarding something you downloaded from the "[Related projects](#)" page, you must contact the developer directly.

Please note that any changes made to the database are not reflected until a new version of WordNet is publicly

NLTK Documentation & Book

NLTK

Documentation

Search

Natural Language Toolkit

NLTK Documentation

API Reference

Example Usage

Module Index

Wiki

FAQ

Open Issues

NLTK on GitHub

Installation

Installing NLTK

Installing NLTK Data

More

Release Notes

NLTK is a leading platform for building Python programs to work with human language data. It provides easy-to-use interfaces to **over 50 corpora and lexical resources** such as WordNet, along with a suite of text processing libraries for classification, tokenization, stemming, tagging, parsing, and semantic reasoning, wrappers for industrial-strength NLP libraries, and an active **discussion forum**.

Thanks to a hands-on guide introducing programming fundamentals alongside topics in computational linguistics, plus comprehensive API documentation, NLTK is suitable for linguists, engineers, students, educators, researchers, and industry users alike. NLTK is available for Windows, Mac OS X, and Linux. Best of all, NLTK is a free, open source, community-driven project.

NLTK has been called “a wonderful tool for teaching, and working in, computational linguistics using Python,” and “an amazing library to play with natural language.”

Natural Language Processing with Python provides a practical introduction to programming for language processing. Written by the creators of NLTK, it guides the reader through the fundamentals of writing Python programs, working with corpora, categorizing text, analyzing linguistic structure, and more. The online version of the book has been updated for Python 3 and NLTK 3. (The original Python 2 version is still available at [http://www.nltk.org/book/ch01.html](#).)

NLTK Documentation & Book

NLTK

Documentation

Search

Natural Language Toolkit

NLTK Documentation

API Reference

Example Usage

Module Index

Wiki

FAQ

Open Issues

NLTK on GitHub

Installation

Installing NLTK

Installing NLTK Data

More

Release Notes

NLTK is a leading platform for building Python programs to work with human language data. It provides easy-to-use interfaces to **over 50 corpora and lexical resources** such as WordNet, along with a suite of text processing libraries for classification, tokenization, stemming, tagging, parsing, and semantic reasoning, wrappers for industrial-strength NLP libraries, and an active **discussion forum**.

Thanks to a hands-on guide introducing programming fundamentals alongside topics in computational linguistics, plus comprehensive API documentation, NLTK is suitable for linguists, engineers, students, educators, researchers, and industry users alike. NLTK is available for Windows, Mac OS X, and Linux. Best of all, NLTK is a free, open source, community-driven project.

NLTK has been called “a wonderful tool for teaching, and working in, computational linguistics using Python,” and “an amazing library to play with natural language.”

Natural Language Processing with Python provides a practical introduction to programming for language processing. Written by the creators of NLTK, it guides the reader through the fundamentals of writing Python programs, working with corpora, categorizing text, analyzing linguistic structure, and more. The online version of the book has been updated for Python 3 and NLTK 3. (The original Python 2 version is still available at [http://www.nltk.org/book/ch01.html](#).)

Text Analysis: NLTK & Beyond

- CDCS Digital Method of the Month on Text Analysis
- Training Pathway for Text Analysis
- Sentiment Analysis Tutorial in Jupyter Notebooks
 - Includes cleaning up text!
 - Uses an NLTK tool called VADER for sentiment analysis
- Text Analysis with Constellate
 - Uses Jupyter Notebooks and libraries other than NLTK
- W3Schools - Python, RegEx, and much more!

Thanks everyone!

Please share your feedback with us in [this survey!](#)