Under The Sea Documentation

Release 1.1.4

Vu Anh

Dec 11, 2017

Contents

Contents:

CHAPTER 1

Installation

1.1 Stable release

To install Under The Sea, run this command in your terminal:

```
$ pip install underthesea
```

This is the preferred method to install Under The Sea, as it will always install the most recent stable release.

If you don't have pip installed, this Python installation guide can guide you through the process.

1.2 From sources

The sources for Under The Sea can be downloaded from the Github repo.

You can either clone the public repository:

```
$ git clone git://github.com/magizbox/underthesea
```

Or download the tarball:

```
$ curl  -OL https://github.com/magizbox/underthesea/tarball/master
```

Once you have a copy of the source, you can install it with:

```
$ python setup.py install
```

CHAPTER 2

Usage

To use underthesea in your project:

```
import underthesea
```

2.1 Word Segmentation

```
# -*- coding: utf-8 -*-
>>> from underthesea import word_sent
>>> sentence = u"Chúng ta thng nói đn Rau sch , Rau an toàn đ phân bit vi các rau␣
↪bình thng bán ngoài ch."

>>> word_sent(sentence)
[u"Chúng ta", u"thng", u"nói", u"đn", u"Rau sch", u",", u"Rau", u"an toàn", u"đ", u
↪"phân bit", u"vi",
u"các", u"rau", u"bình thng", u"bán", u"ngoài", u"ch", u"."]

>>> word_sent(sentence, format="text")
u'Chúng_ta thng nói đn Rau_sch , Rau an_toàn đ phân_bit vi các rau bình_thng bán␣
↪ngoài ch .'
```

2.2 POS Tagging

```
# -*- coding: utf-8 -*-
>>> from underthesea import pos_tag
>>> text = u"Ch tht chó ni ting  TP H Chí Minh b truy quét"
>>> pos_tag(text)
[(u'Ch', 'N'),
 (u'tht', 'N'),
 (u'chó', 'N'),
```

```
 (u'ni ting', 'A'),
 (u'', 'E'),
 (u'TP HCM', 'Np'),
 (u'b', 'V'),
 (u'truy quét', 'V')]
```

2.3 Chunking

```
>>> # -*- coding: utf-8 -*-
>>> from underthesea import chunk
>>> text = u"Bác sĩ bây gi có th thn nhiên báo tin bnh nhân b ung th?"
>>> chunk(text)
[(u'Bác sĩ', 'N', 'B-NP'),
 (u'bây gi', 'P', 'I-NP'),
 (u'có th', 'R', 'B-VP'),
 (u'thn nhiên', 'V', 'I-VP'),
 (u'báo tin', 'N', 'B-NP'),
 (u'bnh nhân', 'N', 'I-NP'),
 (u'b', 'V', 'B-VP'),
 (u'ung th', 'N', 'I-VP'),
 (u'?', 'CH', 'O')]
```

2.4 Text Classification

```
>>> # -*- coding: utf-8 -*-
>>> from underthesea import classify
>>> classify("HLV đu tiên  Premier League b sa thi sau 4 vòng đu")
['The thao']
>>> classify("Hi đng t vn kinh doanh Asean vinh danh gii thng quc t")
['Kinh doanh']
>>> classify("Đánh giá “rp hát ti gia” Samsung Soundbar Sound+ MS750")
['Vi tinh']
```

CHAPTER 3

Contributing to underthesea

Contributions are welcome, and they are greatly appreciated! Every little bit helps, and credit will always be given.

You can contribute in many ways:

3.1 Types of Contributions

3.1.1 Report Bugs

Report bugs at https://github.com/magizbox/underthesea/issues.

If you are reporting a bug, please include:

- Your operating system name and version.
- Any details about your local setup that might be helpful in troubleshooting.
- Detailed steps to reproduce the bug.

3.1.2 Fix Bugs

Look through the GitHub issues for bugs. Anything tagged with "bug" and "help wanted" is open to whoever wants to implement it.

3.1.3 Implement Features

Look through the GitHub issues for features. Anything tagged with "enhancement" and "help wanted" is open to whoever wants to implement it.

3.1.4 Write Documentation

Under The Sea could always use more documentation, whether as part of the official Under The Sea docs, in docstrings, or even on the web in blog posts, articles, and such.

3.1.5 Submit Feedback

The best way to send feedback is to file an issue at https://github.com/magizbox/underthesea/issues.

If you are proposing a feature:

- Explain in detail how it would work.
- Keep the scope as narrow as possible, to make it easier to implement.
- Remember that this is a volunteer-driven project, and that contributions are welcome :)

3.2 Get Started!

Ready to contribute? Here's how to set up *underthesea* for local development.

1. Fork the *underthesea* repo on GitHub.
2. Clone your fork locally:

```
$ git clone git@github.com:your_name_here/underthesea.git
```

3. Install your local copy into a virtualenv. Assuming you have virtualenvwrapper installed, this is how you set up your fork for local development:

```
$ mkvirtualenv underthesea
$ cd underthesea/
$ python setup.py develop
```

4. Create a branch for local development:

```
$ git checkout -b name-of-your-bugfix-or-feature
```

   Now you can make your changes locally.

5. When you're done making changes, check that your changes pass flake8 and the tests, including testing other Python versions with tox:

```
$ flake8 underthesea tests
$ python setup.py test or py.test
$ tox
```

   To get flake8 and tox, just pip install them into your virtualenv.

6. Commit your changes and push your branch to GitHub:

```
$ git add .
$ git commit -m "Your detailed description of your changes."
$ git push origin name-of-your-bugfix-or-feature
```

7. Submit a pull request through the GitHub website.

3.3 Pull Request Guidelines

Before you submit a pull request, check that it meets these guidelines:

1. The pull request should include tests.
2. If the pull request adds functionality, the docs should be updated. Put your new functionality into a function with a docstring, and add the feature to the list in README.rst.
3. The pull request should work for Python 2.6, 2.7, 3.3, 3.4 and 3.5, and for PyPy. Check https://travis-ci.org/magizbox/underthesea/pull_requests and make sure that the tests pass for all supported Python versions.

3.4 Tips

To run a subset of tests:

```
$ python -m unittest tests.test_underthesea
```

CHAPTER 4

API

4.1 underthesea Package

4.2 corpus Package

4.3 transformer Package

4.4 word_sent Package

underthesea.word_sent.**tokenize**(*sentence*)

tokenize a sentence

Parameters **text** – raw text input

Returns tokenize text

Return type unicodelstr

```
# -*- coding: utf-8 -*-
>>> from underthesea.word_sent.tokenize import tokenize
>>> text = u"Đám cháy bùng phát tra nay, 7/4, ti khu nhà tôn ngay gn tòa nhà Keangnam,
↪ đng Phm Hùng. Ngn la cùng khói đen bc lên d di làm đen kt mt góc không gian. Giao␣
↪thông quanh khu vc b nh hng, trong đó đng trên cao b tc mt đon khá dài..."

>>> tokenize(text)
u"Đám cháy bùng phát tra nay , 7 / 4 , ti khu nhà tôn ngay gn tòa nhà Keangnam , đng␣
↪Phm Hùng . Ngn la cùng khói đen bc lên d di làm đen kt mt góc không gian . Giao␣
↪thông quanh khu vc b nh hng , trong đó đng trên cao b tc mt đon khá dài ..."
```

underthesea.**word_sent**(*sentence*)

word segmentation

Parameters **sentence** (*unicode | str*) – raw sentence

Returns segmented sentence

Return type unicodelstr

```
>>> # -*- coding: utf-8 -*-
>>> from underthesea import word_sent
>>> sentence = u"Chúng ta thng nói đn Rau sch , Rau an toàn đ phân bit vi các rau␣
↪bình thng bán ngoài ch ."

>>> word_sent(sentence)
[u"Chúng ta", u"thng", u"nói", u"đn", u"Rau sch", u",", u"Rau", u"an toàn", u"đ", u
↪"phân bit", u"vi",
u"các", u"rau", u"bình thng", u"bán", u"ngoài", u"ch", u"."]

>>> word_sent(sentence, format="text")
u'Chúng_ta thng nói đn Rau_sch , Rau an_toàn đ phân_bit vi các rau bình_thng bán␣
↪ngoài ch .'
```

4.5 pos_tag Package

underthesea.**pos_tag**(*sentence*)

part of speech tagging

Parameters **sentence** (*unicode | str*) – raw sentence

Returns tagged sentence

Return type list

```
>>> # -*- coding: utf-8 -*-
>>> from underthesea import pos_tag
>>> text = u"Ch tht chó ni ting  TP H Chí Minh b truy quét"
>>> pos_tag(text)
[(u'Ch', 'N'),
 (u'tht', 'N'),
 (u'chó', 'N'),
 (u'ni ting', 'A'),
 (u'', 'E'),
 (u'TP HCM', 'Np'),
 (u'b', 'V'),
 (u'truy quét', 'V')]
```

4.6 chunking Package

underthesea.**chunk**(*sentence*)

chunk a sentence to phrases

Parameters **sentence** (*unicode*) – raw sentence

Returns list of tuple with word, pos tag, chunking tag

Return type list

```
>>> # -*- coding: utf-8 -*-
>>> from underthesea import chunk
>>> text = u"Bác sĩ bây gi có th thn nhiên báo tin bnh nhân b ung th?"
```

```
>>> chunk(text)
[(u'Bác sĩ', 'N', 'B-NP'),
 (u'bây gi', 'P', 'I-NP'),
 (u'có th', 'R', 'B-VP'),
 (u'thn nhiên', 'V', 'I-VP'),
 (u'báo tin', 'N', 'B-NP'),
 (u'bnh nhân', 'N', 'I-NP'),
 (u'b', 'V', 'B-VP'),
 (u'ung th', 'N', 'I-VP'),
 (u'?', 'CH', 'O')]
```

4.7 `classify` Package

underthesea.**classify**(*text*)

Text classification

Parameters **sentence** (*unicode*) – raw text

Returns list of labels

Return type list

```
>>> # -*- coding: utf-8 -*-
>>> from underthesea import classify
>>> classify("HLV đu tiên  Premier League b sa thi sau 4 vòng đu")
['The thao']
>>> classify("Hi đng t vn kinh doanh Asean vinh danh gii thng quc t")
['Kinh doanh']
>>> classify("Đánh giá “rp hát ti gia” Samsung Soundbar Sound+ MS750")
['Vi tinh']
```

CHAPTER 5

History

5.1 1.1.4 (2017-09-12)

- New feature: text classification
- [bug] Fix Text error
- [doc] Add facebook link

5.2 1.1.3 (2017-08-30)

- Add live demo: https://underthesea.herokuapp.com/

5.3 1.1.2 (2017-08-22)

- Add dictionary

5.4 1.1.1 (2017-07-05)

- Support Python 3
- Refactor feature_engineering code

5.5 1.1.0 (2017-05-30)

- Add chunking feature
- Add pos_tag feature

- Add word_sent feature, fix performance
- Add Corpus class
- Add Transformer classes
- Integrated with dictionary of Ho Ngoc Duc
- Add travis-CI, auto build with PyPI

5.6 1.0.0 (2017-03-01)

- First release on PyPI.
- First release on Readthedocs

CHAPTER 6

Indices and tables

- genindex
- modindex
- search

Index

U