
Under The Sea Documentation

Release 1.1.9

Vu Anh

Oct 29, 2020

1	Underthesea - Vietnamese NLP Toolkit	3
2	AUTHORS	9
3	History	11
4	word_tokenize	15
5	pos_tag	17
6	chunking	19
7	ner	21
8	classify	23
9	sentiment	25
10	Indices and tables	27

Vietnamese NLP Toolkit

Underthesea - Vietnamese NLP Toolkit



underthesea is a suite of open source Python modules, data sets and tutorials supporting research and development in Vietnamese Natural Language Processing.

Free software	GNU General Public License v3
Live demo	undertheseanlp.com
Colab notebooks	latest: v1.2.0 , stable: v1.1.17
Documentation	Underthesea Documentation
Facebook	Underthesea Page
Youtube	Underthesea NLP Channel

1.1 Installation

To install underthesea, simply:

```
$ pip install underthesea
```

Satisfaction, guaranteed.

1.2 Usage

- 1. Sentence Segmentation
- 2. Word Segmentation
- 3. POS Tagging
- 4. Chunking
- 5. Named Entity Recognition
- 6. Text Classification
- 7. Sentiment Analysis
- 8. Vietnamese NLP Resources

1.2.1 1. Sentence Segmentation

Usage

```
>>> # -*- coding: utf-8 -*-
>>> from underthesea import sent_tokenize
>>> text = 'Taylor cho bit lúc đầu cô cảm thấy ngại vì cô bạn thân Amanda nhng ri mi th
↳trôi qua nhanh chóng. Amanda cũng thôi mái vì mi quan h này.'

>>> sent_tokenize(text)
[
    "Taylor cho bit lúc đầu cô cảm thấy ngại vì cô bạn thân Amanda nhng ri mi th trôi qua
↳nhanh chóng.",
    "Amanda cũng thôi mái vì mi quan h này."
]
```

1.2.2 2. Word Segmentation

Usage

```
>>> # -*- coding: utf-8 -*-
>>> from underthesea import word_tokenize
>>> sentence = 'Chàng trai 9X Qung Tr khi nghip t nm sò'

>>> word_tokenize(sentence)
['Chàng trai', '9X', 'Qung Tr', 'khi nghip', 't', 'nm', 'sò']

>>> word_tokenize(sentence, format="text")
'Chàng_trai 9X Qung_Tr khi_nghip t nm sò'
```

1.2.3 3. POS Tagging

Usage

```
>>> # -*- coding: utf-8 -*-
>>> from underthesea import pos_tag
>>> pos_tag('Ch tht chó ni ting Sài Gòn b truy quét')
[('Ch', 'N'),
 ('tht', 'N'),
 ('chó', 'N'),
 ('ni ting', 'A'),
 ('', 'E'),
 ('Sài Gòn', 'Np'),
 ('b', 'V'),
 ('truy quét', 'V')]
```

1.2.4 4. Chunking

Usage

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

1.2.5 5. Named Entity Recognition

Usage

```
>>> # -*- coding: utf-8 -*-
>>> from underthesea import ner
>>> text = 'Cha tit l lch trình ti Vit Nam ca Tng thng M Donald Trump'
>>> ner(text)
[('Cha', 'R', 'O', 'O'),
 ('tit l', 'V', 'B-VP', 'O'),
 ('lch trình', 'V', 'B-VP', 'O'),
 ('ti', 'E', 'B-PP', 'O'),
 ('Vit Nam', 'Np', 'B-NP', 'B-LOC'),
 ('ca', 'E', 'B-PP', 'O'),
 ('Tng thng', 'N', 'B-NP', 'O'),
 ('M', 'Np', 'B-NP', 'B-LOC'),
 ('Donald', 'Np', 'B-NP', 'B-PER'),
 ('Trump', 'Np', 'B-NP', 'I-PER')]
```

1.2.6 6. Text Classification

Download models

```
$ underthesea download-model TC_GENERAL
$ underthesea download-model TC_BANK
```

Usage

```
>>> # -*- 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 gi thng quc t')
['Kinh doanh']

>> classify('Lãi sut t BIDV rt u đãi', domain='bank')
['INTEREST_RATE']
```

1.2.7 7. Sentiment Analysis

Download models

```
$ underthesea download-model SA_GENERAL
$ underthesea download-model SA_BANK
```

Usage

```
>>> # -*- coding: utf-8 -*-
>>> from underthesea import sentiment

>>> sentiment('hàng kém cht lg,chăn đp lên dnh lông lá khp ngi. tht vng')
negative
>>> sentiment('Sn phm hi nh so vi tng tng nhng cht lng tt, đóng gói cn thn.')
positive

>>> sentiment('Đky qua đng link bài vit này t th 6 mà gi cha thy ai lne ht', domain=
↳ 'bank')
```

(continues on next page)

(continued from previous page)

```
[ 'CUSTOMER_SUPPORT#negative' ]
>>> sentiment('Xem li vn thy xúc đng và t hào v BIDV ca mình', domain='bank')
[ 'TRADEMARK#positive' ]
```

1.2.8 8. Vietnamese NLP Resources

List resources

```
$ underthesea list-data
| Name          | Type          | License | Year | Directory          |
|-----+-----+-----+-----+-----+
| UTS2017-BANK  | Categorized   | Open    | 2017 | datasets/UTS2017-BANK |
| VNESES        | Plaintext     | Open    | 2012 | datasets/LTA         |
| VNTQ_BIG      | Plaintext     | Open    | 2012 | datasets/LTA         |
| VNTQ_SMALL    | Plaintext     | Open    | 2012 | datasets/LTA         |
| VNTC          | Categorized   | Open    | 2007 | datasets/VNTC        |

$ underthesea list-data --all
```

Download resources

```
$ underthesea download-data VNTC
100%|| 74846806/74846806 [00:09<00:00, 8243779.16B/s]
Resource VNTC is downloaded in ~/.underthesea/datasets/VNTC folder
```

1.3 Up Coming Features

- Text to Speech
- Automatic Speech Recognition
- Machine Translation
- Dependency Parsing

1.4 Contributing

Do you want to contribute with underthesea development? Great! Please read more details at [CONTRIBUTING.rst](#).

2.1 Original Authors

- Vu Anh <anhv.ict91@gmail.com>

2.2 Awesome Contributors

- Bui Nhat Anh <buinhatanh1208@gmail.com>
- Doan Viet Dung <doanvietdung273@gmail.com>

2.3 Thanks

Thanks to all the wonderful folks who have contributed to schedule over the years

- Nhu Bao Vu <nhubaovu@gmail.com>
- Hoai-Thu Vuong <thuvh87@gmail.com>

3.1 1.2.1 (2020-10-28)

- Sanity check python version (GH-320)
- Handle exception case in sentiment module (GH-321)
- Cập nhật quản lý resources từ languageflow (GH-295)
- Lỗi b languageflow trong quá trình cài đặt (GH-295)
- Cập nhật phiên bản fasttext (GH-304)

3.2 1.1.16 (2019-06-15)

- Bumping up version of the languageflow dependency (GH-231)
- Update phiên bản scikit-learn 0.20.2 (GH-229)
- Cập nhật li các dependencies (GH-241)
- Cập nhật mô hình trên bộ dữ liệu VNTC (GH-246)
- Cập nhật mô hình trên bộ dữ liệu UTS2017_BANK_TC (GH-243)
- Cập nhật mô hình trên bộ dữ liệu UTS2017_BANK_SA (GH-244)
- Li ví các câu sentiment demo (GH-236)
- Thống nhất cách đặt tên và quản lý model (GH-225)

3.3 1.1.12 (2019-03-13)

- Add sentence segmentation feature

3.4 1.1.9 (2019-01-01)

- Improve speed of word_tokenize function
- Only support python 3.6+
- Use flake8 for style guide enforcement

3.5 1.1.8 (2018-06-20)

- Fix word_tokenize error when text contains tab (t) character
- Fix regex_tokenize with url

3.6 1.1.7 (2018-04-12)

- Rename word_sent function to word_tokenize
- Refactor version control in setup.py file and __init__.py file
- Update documentation badge url

3.7 1.1.6 (2017-12-26)

- New feature: aspect sentiment analysis
- Integrate with languageflow 1.1.6
- Fix bug tokenize string with '=' (#159)

3.8 1.1.5 (2017-10-12)

- New feature: named entity recognition
- Refactor and update model for word_sent, pos_tag, chunking

3.9 1.1.4 (2017-09-12)

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

3.10 1.1.3 (2017-08-30)

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

3.11 1.1.2 (2017-08-22)

- Add dictionary

3.12 1.1.1 (2017-07-05)

- Support Python 3
- Refactor feature_engineering code

3.13 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

3.14 1.0.0 (2017-03-01)

- First release on PyPI.
- First release on Readthedocs

CHAPTER 4

word_tokenize

CHAPTER 5

pos_tag

CHAPTER 6

chunking

CHAPTER 7

ner

CHAPTER 8

classify

Install dependencies and download default model

```
$ pip install Cython
$ pip install future scipy numpy scikit-learn
$ pip install -U fasttext --no-cache-dir --no-deps --force-reinstall
$ underthesea data
```


CHAPTER 9

sentiment

Install dependencies

```
$ pip install future scipy numpy scikit-learn==0.19.2 joblib
```


CHAPTER 10

Indices and tables

- `genindex`
- `modindex`
- `search`