
Under The Sea Documentation

Release 1.1.9

Vu Anh

Jul 14, 2023

1	Underthesea - Vietnamese NLP Toolkit	3
2	AUTHORS	9
3	History	11
4	The API Documentation / Guide	19
5	Indices and tables	23
	Python Module Index	25
	Index	27

Vietnamese NLP Toolkit

CHAPTER 1

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.

Version 1.3.0 out now! [Underthesea meet deep learning!](#)

Free software	GNU General Public License v3
Live demo	undertheseanlp.com
Colab notebooks	latest / stable
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. Dependency Parsing
- 6. Named Entity Recognition
- 7. Text Classification
- 8. Sentiment Analysis
- 9. 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 cm thý ngi vi cô bn thân Amanda nhng ri mi th_
↳trôi qua nhanh chóng. Amanda cũng thoi mái vi mi quan h này.'

>>> sent_tokenize(text)
[
    "Taylor cho bit lúc đầu cô cm thý ngi vi cô bn thân Amanda nhng ri mi th trôi qua_
↳nhanh chóng.",
    "Amanda cũng thoi mái vi 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 gì có th thn nhiên báo tin bnh nhân b ung th?'
>>> chunk(text)
[('Bác sĩ', 'N', 'B-NP'),
 ('bây gì', '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. Dependency Parsing

Usage

```
>>> # -*- coding: utf-8 -*-
>>> from underthesea import dependency_parse
>>> text = 'Ti 29/11, Vit Nam thêm 2 ca mc Covid-19'
>>> dependency_parse(text)
[('Ti', 5, 'obl:tmod'),
 ('29/11', 1, 'flat:date'),
 (',', 1, 'punct'),
 ('Vit Nam', 5, 'nsubj'),
 ('thêm', 0, 'root'),
 ('2', 7, 'nummod'),
 ('ca', 5, 'obj'),
 ('mc', 7, 'nmod'),
 ('Covid-19', 8, 'nummod')]
```

1.2.6 6. Named Entity Recognition

Usage

```
>>> # -*- coding: utf-8 -*-
>>> from underthesea import ner
>>> text = 'Cha tit 1 lch trình ti Vit Nam ca Tng thng M Donald Trump'
>>> ner(text)
[('Cha', 'R', 'O', 'O'),
 ('tit 1', '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')]
```

1.2.7 7. 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']
```

1.2.8 8. 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
```

(continues on next page)

(continued from previous page)

```
positive

>>> sentiment('Đky qua đng link bài vit này t th 6 mà gi cha thy ai lhe ht', domain=
↳ 'bank')
['CUSTOMER_SUPPORT#negative']
>>> sentiment('Xem li vn thy xúc đng và t hào v BIDV ca mình', domain='bank')
['TRADEMARK#positive']
```

1.2.9 9. 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

- Machine Translation
- Text to Speech
- Automatic Speech Recognition

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

- Nguyen Dang Duc Tai <tainguyen7595@gmail.com>
- Bui Nhat Anh <buinhatanh1208@gmail.com>
- Vuong Quoc Binh <binh@hauai.vn>
- 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 6.4.0 (2023-07-14)

- GH-686: Fix fixed_words regex

3.2 6.3.0 (2023-06-28)

- GH-685: Support MacOS ARM

3.3 6.2.0 (2023-03-04)

- GH-173: Add Text to Speech API by @rain1024 in <https://github.com/undertheseanlp/underthesea/pull/668>
- GH-502: Provide training script for word segmentation and pos tagging and named entity recognition by @rain1024 in <https://github.com/undertheseanlp/underthesea/pull/666>
- GH-622: Create UTS_Dictionary v1.0 datasets by @rain1024 in <https://github.com/undertheseanlp/underthesea/pull/663>

3.4 6.1.4 (2023-02-26)

- GH-588: Support underthesea_core with python 3.11 by @rain1024 in <https://github.com/undertheseanlp/underthesea/pull/659>
- GH-588: update underthesea_core version by @rain1024 in <https://github.com/undertheseanlp/underthesea/pull/660>

3.5 6.1.3 (2023-02-25)

- Bump django from 4.1.6 to 4.1.7 in /apps/languages/backend by @dependabot in <https://github.com/undertheseanlp/underthesea/pull/652>
- Bump django from 3.2.17 to 3.2.18 in /apps/service by @dependabot in <https://github.com/undertheseanlp/underthesea/pull/651>
- GH-502: Training example for word segmentation by @rain1024 in <https://github.com/undertheseanlp/underthesea/pull/654>
- Add two new datasets UTS_Text and UTS_WTK

3.6 6.1.2 (2023-02-15)

- GH-648: Add option fixed_words to tokenize and word_tokenize api by @rain1024 in <https://github.com/undertheseanlp/underthesea/pull/649>

3.7 6.1.1 (2023-02-10)

- GH-641: Correct the error with the filename of the dataset in Windows by @rain1024 in <https://github.com/undertheseanlp/underthesea/pull/644>
- Bump django from 3.2.16 to 3.2.17 in /apps/service by @dependabot in <https://github.com/undertheseanlp/underthesea/pull/640>
- Bump django from 4.1.4 to 4.1.6 in /apps/languages/backend by @dependabot in <https://github.com/undertheseanlp/underthesea/pull/639>
- Bump ua-parser-js from 0.7.28 to 0.7.33 in /apps/directory/components/json_viewer/component/frontend by @dependabot in <https://github.com/undertheseanlp/underthesea/pull/636>
- Bump future from 0.16.0 to 0.18.3 in /apps/service by @dependabot in <https://github.com/undertheseanlp/underthesea/pull/645>

3.8 6.1.0 (2023-02-08)

- GH-641: fix issue filename of dataset is not correct by @rain1024 in <https://github.com/undertheseanlp/underthesea/pull/642>

3.9 6.0.3 (2023-01-25)

- GH-622: Initialize Dictionary page feature by @rain1024 in <https://github.com/undertheseanlp/underthesea/pull/633>
- GH-622: Add dictionary page by @rain1024 in <https://github.com/undertheseanlp/underthesea/pull/634>

3.10 6.0.2 (2023-01-17)

- GH-628: Create unittest for django API by @rain1024 in <https://github.com/undertheseanlp/underthesea/pull/629>
- GH-607: add test frontend with jest by @rain1024 in <https://github.com/undertheseanlp/underthesea/pull/630>

Full Changelog: <https://github.com/undertheseanlp/underthesea/compare/v6.0.1...v6.0.2>

3.11 6.0.1 (2023-01-08)

- GH-607: add Articles UI by @rain1024 in <https://github.com/undertheseanlp/underthesea/pull/620>
- GH-621: Corpus CP_Vietnamese_VLC_v2_2022 by @rain1024 in <https://github.com/undertheseanlp/underthesea/pull/624>

3.12 6.0.0 (2023-01-01)

- Happy New Year 2023! Let's bump up the version! (GH-616)

3.13 1.4.1 (2022-12-17)

- Create underthesea app (GH-607)
- Add viet2ipa module (GH-437)
- Training NER model with VLSP2016 dataset using BERT (GH-437)
- Remove unicode as a dependency (GH-569)

3.14 1.3.5 (2022-10-31)

- Add Text Normalization module (GH-534)
- Release underthesea_core version 0.0.5a2 (GH-550)
- Support GLIBC_2.17 (GH-530)
- Update resources path (GH-540)
- Fix function word_tokenize (GH-528)

3.15 1.3.4 (2022-01-08)

- Demo chatbot with rasa (GH-513)
- Lite version of underthesea (GH-505)
- Increase word_tokenize speed 1.5 times (GH-185)
- Add build for windows (GH-185)

3.16 1.3.3 (2021-09-02)

- Update torch and transformer dependency (GH-403)

3.17 1.3.2 (2021-08-04)

- Publish two ABSA open datasets (GH-417)
- Migrate from travis-ci to github actions (GH-410)
- Update ParserTrainer (GH-392)
- Add pipeline folder (GH-351)

3.18 1.3.1 (2021-01-11)

- Compatible with newer version of scikit-learn (GH-313)
- Retrain classification and sentiment models with latest version of scikit-learn (GH-381)
- Add ClassifierTrainer (from languageflow) (GH-381)
- Add 3 new datasets (GH-351)
- [Funny Update] Change underthesea's avatar (GH-371)
- [CI] Add Stale App: Automatically close stale Issues and Pull Requests that tend to accumulate during a project (GH-351)

3.19 1.3.0 (2020-12-11)

- Remove languageflow dependency (GH-364)
- Remove tabulate dependency (GH-364)
- Remove scores in text classification and sentiment section (GH-351)
- Add information of dependency_parse module in info function (GH-351)
- Try to use Github Actions (GH-353)
- Dependency Parsing (GH-157)

3.20 1.2.3 (2020-11-28)

- Refactor config for resources (GH-300)
- Thêm API x lý d liu (GH-299)

3.21 1.2.2 (2020-11-04)

- Remove nltk strict version (GH-308)
- Add word_hyphen rule (GH-290)
- Sanity check python version (GH-320)
- Handle exception case in sentiment module (GH-321)
- Cập nhật quản lý resources từ languageflow (GH-295)
- Lỗi về languageflow trong quá trình cài đặt (GH-295)
- Cập nhật phiên bản fasttext (GH-304)

3.22 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 liên 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)
- Liệt kê các câu sentiment demo (GH-236)
- Thống nhất cách đặt tên và quản lý model (GH-225)

3.23 1.1.12 (2019-03-13)

- Add sentence segmentation feature

3.24 1.1.9 (2019-01-01)

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

3.25 1.1.8 (2018-06-20)

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

3.26 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

3.27 1.1.6 (2017-12-26)

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

3.28 1.1.5 (2017-10-12)

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

3.29 1.1.4 (2017-09-12)

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

3.30 1.1.3 (2017-08-30)

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

3.31 1.1.2 (2017-08-22)

- Add dictionary

3.32 1.1.1 (2017-07-05)

- Support Python 3
- Refactor feature_engineering code

3.33 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

3.34 1.0.0 (2017-03-01)

- First release on PyPI.
- First release on Readthedocs

If you are looking for information on a specific function, class, or method, this part of the documentation is for you.

4.1 Developer Interface

4.1.1 word_tokenize

`underthesea.word_tokenize(sentence, format=None, use_token_normalize=True, fixed_words=[])`
Vietnamese word segmentation

Parameters

- **sentence** (*str*) – raw sentence
- **format** (*str*, *optional*) – format option. Defaults to None. use format='text' for text format
- **use_token_normalize** (*bool*) – True if use token_normalize
- **fixed_words** (*list*) – list of fixed words

Returns word tokens

Return type list of str

Examples

```
>>> # -*- coding: utf-8 -*-
>>> from underthesea import word_tokenize
>>> sentence = "Bác sĩ bây giờ có thể thn nhiên báo tin bệnh nhân bị ung thư"
```

```
>>> word_tokenize(sentence)
["Bác sĩ", "bây giờ", "có thể", "thn nhiên", "báo tin", "bệnh nhân", "bị", "ung thư"]
```

```
>>> word_tokenize(sentence, format="text")
"Bác_sĩ bây_gi có_th thn_nhiên báo_tin bnh_nhân b ung_th"
```

4.1.2 pos_tag

`underthesea.pos_tag(sentence, format=None, model=None)`

4.1.3 chunking

`underthesea.chunk(sentence, format=None)`

Vietnamese chunking

Parameters `sentence` (`{unicode, str}`) – raw sentence

Returns `tokens` – tagged sentence

Return type list of tuple with word, pos tag, chunking tag

Examples

```
>>> # -*- coding: utf-8 -*-
>>> from underthesea import chunk
>>> sentence = "Nghị vn 4 thi th Triu Tiên trôi dt b bin Nht Bn"
>>> chunk(sentence)
[('Nghị vn', 'N', 'B-NP'),
 ('4', 'M', 'B-NP'),
 ('thi th', 'N', 'B-NP'),
 ('Triu Tiên', 'Np', 'B-NP'),
 ('trôi dt', 'V', 'B-VP'),
 ('b bin', 'N', 'B-NP'),
 ('Nht Bn', 'Np', 'B-NP')]
```

4.1.4 ner

`underthesea.ner(sentence, format=None, deep=False)`

Location and classify named entities in text

Parameters `sentence` (`{unicode, str}`) – raw sentence

Returns `tokens`

Return type list of tuple with word, pos tag, chunking tag, ner tag tagged sentence

Examples

```
>>> # -*- coding: utf-8 -*-
>>> from underthesea import ner
>>> sentence = "Ông Putin ca ngi nhng thành tu vĩ đi ca Liên Xô"
>>> ner(sentence)
[('Ông', 'Nc', 'B-NP', 'O'),
 ('Putin', 'Np', 'B-NP', 'B-PER'),
 ('ca ngi', 'V', 'B-VP', 'O'),
```

(continues on next page)

(continued from previous page)

```
( 'nhng', 'L', 'B-NP', 'O'),
( 'thành tu', 'N', 'B-NP', 'O'),
( 'vĩ đi', 'A', 'B-AP', 'O'),
( 'ca', 'E', 'B-PP', 'O'),
( 'Liên Xô', 'Np', 'B-NP', 'B-LOC') ]
```

4.1.5 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
```

`underthesea.classify(X, domain=None)`

Text classification

Parameters

- **X** (*{unicode, str}*) – raw sentence
- **domain** (*{None, 'bank'}*) –
domain of text
 - None: general domain
 - bank: bank domain

Returns `tokens` – categories of sentence

Return type `list`

4.1.6 sentiment

Install dependencies

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

`underthesea.sentiment(X, domain='general')`

Sentiment Analysis

Parameters

- **X** (*str*) – raw sentence
- **domain** (*str*) – domain of text (bank or general). Default: *general*

Returns

- **Text** (*Text of input sentence*)
- **Labels** (*Sentiment of sentence*)

Examples

```
>>> from underthesea import sentiment
>>> sentence = "Chuyen tin k nhn Dc tien"
>>> sentiment(sentence, domain='bank')
[MONEY_TRANSFER#negative (1.0)]
```

4.1.7 viet2ipa

`underthesea.pipeline.ipa.viet2ipa` (*text: str, *args, **kwargs*)

Generate ipa of the syllable

Vietnamese syllabic structure (Anh & Trang 2022)

syllable = onset + rhyme + tone

rhyme = medial + nuclear vowel + (coda)

Parameters

- **text** (*str*) – represents syllable
- **dialect** (*str*) – Either the *string* “north” or “south”. Default: *north*
- **eight** (*boolean*) – If true, use eight tone format, else use six tone format. Default: *False*
- **tone** (*str*) – Either the *string* “ipa” or “number”. Default: *number*

Returns A *string*. Represents ipa of the syllable

Examples

```
>>> # -*- coding: utf-8 -*-
>>> from underthesea.pipeline.ipa import viet2ipa
>>> viet2ipa("trng")
to32
```

CHAPTER 5

Indices and tables

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

u

`underthesea`, [19](#)

C

`chunk()` (*in module underthesea*), [20](#)
`classify()` (*in module underthesea*), [21](#)

N

`ner()` (*in module underthesea*), [20](#)

P

`pos_tag()` (*in module underthesea*), [20](#)

S

`sentiment()` (*in module underthesea*), [21](#)

U

`underthesea` (*module*), [19](#)

V

`viet2ipa()` (*in module underthesea.pipeline.ipa*), [22](#)

W

`word_tokenize()` (*in module underthesea*), [19](#)