

---

# **Under The Sea Documentation**

***Release 1.1.9***

**Vu Anh**

**Apr 26, 2020**



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



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
- Documentation: <https://underthesea.readthedocs.io>

- Live demo: [undertheseanlp.com](http://undertheseanlp.com)
- Facebook Page: <https://www.facebook.com/undertheseanlp/>
- 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

### 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_
↳nh 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 tng Sài Gòn b truy quét')
[('Ch', 'N'),
 ('tht', 'N'),
 ('chó', 'N'),
 ('ni tng', '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. 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)
```

(continues on next page)

(continued from previous page)

```
[('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

Install dependencies and download default model

```
$ pip install git+https://github.com/facebookresearch/fastText.git@v0.2.0
$ pip install unidecode
$ underthesea download tc_general
$ underthesea download 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 đăi', domain='bank')
['INTEREST_RATE']
```

## 1.2.7 7. Sentiment Analysis

Install dependencies

```
$ pip install git+https://github.com/facebookresearch/fastText.git@v0.2.0
$ pip install unidecode
$ underthesea download sa_general
$ underthesea download sa_bank
```

Usage

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

>>> sentiment('hàng kém cht lg, chẵn 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 lhe 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']
```

## 1.3 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 1.1.16 (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)
- Tổng hợp cách đặt tên và quản lý model (GH-225)

#### 3.2 1.1.12 (2019-03-13)

- Add sentence segmentation feature

#### 3.3 1.1.9 (2019-01-01)

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

### 3.4 1.1.8 (2018-06-20)

- Fix word\_tokenize error when text contains tab (t) character
- Fix regex\_tokenize with url

### 3.5 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

### 3.6 1.1.6 (2017-12-26)

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

### 3.7 1.1.5 (2017-10-12)

- New feature: named entity recognition
- Refactor and update model for word\_sent, pos\_tag, chunking

### 3.8 1.1.4 (2017-09-12)

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

### 3.9 1.1.3 (2017-08-30)

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

### 3.10 1.1.2 (2017-08-22)

- Add dictionary



### **3.11 1.1.1 (2017-07-05)**

- Support Python 3
- Refactor feature\_engineering code

### **3.12 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

### **3.13 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`