Under The Sea Documentation

Release 1.1.8

Vu Anh

Jun 20, 2018

Notes

Vietnamese NLP Toolkit

CHAPTER 1

Underthesea - Vietnamese NLP Toolkit

[English] [Ting Vit]

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: underthesea app
- Facebook Page: https://www.facebook.com/undertheseanlp/

1.1 Installation

To install underthesea, simply:

```
$ pip install underthesea==1.1.8
```

Satisfaction, guaranteed.

1.2 Usage

- *1. Word Segmentation*
- *2. POS Tagging*
- *3. Chunking*
- *4. Named Entity Recognition*
- *5. Text Classification*
- *6. Sentiment Analysis*

1.2.1 1. Word Segmentation

```
>>> # -*- 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'
```

1.2.2 2. POS Tagging

```
>>> # -*- 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')]
```

1.2.3 3. Chunking

```
>>> # -*- 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)
```

(continues on next page)

(continued from previous page)

```
[('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')]
```

1.2.4 4. Named Entity Recognition

```
>>> # -*- 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')]
```

1.2.5 5. Text Classification

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

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 gii thng quc t')
['Kinh doanh']
>>> classify('Đánh giá "rp hát ti gia" Samsung Soundbar Sound+ MS750')
['Vi tinh']
```

1.2.6 6. Sentiment Analysis

```
$ pip install future scipy numpy scikit-learn==0.19.0 joblib
```

Usage

```
>>> # -*- coding: utf-8 -*-
>>> from underthesea import sentiment
>>> sentiment('Gi my ln mà lúc nào cũng là các chuyên viên đang bn ht ', domain='bank
↪')
('CUSTOMER SUPPORT#NEGATIVE',)
>>> sentiment('bidv cho vay hay ko phu thuoc y thich cua thang tham dinh, ko co quy␣
↪dinh ro rang', domain='bank')
('LOAN#NEGATIVE',)
```

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.

CHAPTER 2

Credits

2.1 Development Lead

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

2.2 Contributors

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

2.3 Developer Advocates

- Nhu Bao Vu <nhubaovu@gmail.com>
- Doan Viet Dung <doanvietdung273@gmail.com>

CHAPTER 3

History

3.1 1.1.8 (2018-06-20)

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

3.2 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

3.3 1.1.6 (2017-12-26)

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

3.4 1.1.5 (2017-10-12)

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

3.5 1.1.4 (2017-09-12)

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

3.6 1.1.3 (2017-08-30)

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

3.7 1.1.2 (2017-08-22)

- Add dictionary

3.8 1.1.1 (2017-07-05)

- Support Python 3
- Refactor feature_engineering code

3.9 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

3.10 1.0.0 (2017-03-01)

- First release on PyPI.
- First release on Readthedocs

CHAPTER 4

word_tokenize

underthesea.word_tokenize.**word_tokenize**(*sentence*, *format=None*)

Vietnamese word segmentation

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

Returns **tokens** – tagged sentence

Return type list of text

Examples

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

```
>>> word_tokenize(sentence)
['Bác sĩ', 'bây gi', 'có th', 'thn nhiên', 'báo tin', 'bnh nhân', 'b', 'ung th']
```

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

CHAPTER 5

pos_tag

underthesea.pos_tag.**pos_tag**(*sentence*, *format=None*)

Vietnamese POS tagging

Parameters **sentence** (*{unicode, str}*) – Raw sentence

Returns **tokens** – tagged sentence

Return type list of tuple with word, pos tag

Examples

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

CHAPTER 6

chunking

underthesea.chunking.**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 = "Nghi vn 4 thi th Triu Tiên trôi dt b bin Nht Bn"
>>> chunk(sentence)
[('Nghi 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')]
```

CHAPTER 7

ner

underthesea.ner.**ner**(*sentence*, *format=None*)

Location and classify named entities in text

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

Returns **tokens** – tagged sentence

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

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'),
('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')]
```

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
```

underthesea.classification.**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

Examples

```
>>> # -*- coding: utf-8 -*-
>>> from underthesea import classify
>>> sentence = "HLV ngoi đòi gn t mi tháng dn dt tuyn Vit Nam"
>>> classify(sentence)
['The thao']
```

```
>>> sentence = "Tôi rt thích cách phc v ca nhân viên BIDV"
>>> classify(sentence, domain='bank')
('CUSTOMER SUPPORT',)
```

CHAPTER 9

sentiment

Install dependencies

```
$ pip install future scipy numpy scikit-learn==0.19.0 joblib
```

underthesea.sentiment.**sentiment**(*X*, *domain=None*)

Sentiment Analysis

Parameters

- **X** (*{unicode, str}*) – raw sentence
- **domain** (*{'bank'}*) –

  domain of text

  - bank: bank domain

Returns **tokens** – sentiment of sentence

Return type list

Examples

```
>>> # -*- coding: utf-8 -*-
>>> from underthesea import sentiment
>>> sentence = "Va smartbidv, va bidv online mà li k dùng chung 1 tài khon đăng␣
↪nhp, rc ri!"
>>> sentiment(sentence, domain='bank')
('INTERNET BANKING#NEGATIVE',)
```

CHAPTER 10

Indices and tables

- genindex
- modindex
- search

Index