Underthesea

Nội dung

Installation

To install underthesea, simply:

```
$ pip install underthesea
```

1. Sentence Segmentation

Usage

```
>>> # -*- coding: utf-8 -*-
>>> from underthesea import sent_tokenize
>>> text = 'Taylor cho biết lúc đầu cô cảm thấy ngại với cô bạn thân Amanda nhưng rồi mọi thứ trôi qua nhanh chóng. Amanda cũng thoải mái với mối quan hệ này.'

>>> sent_tokenize(text)
[
   "Taylor cho biết lúc đầu cô cảm thấy ngại với cô bạn thân Amanda nhưng rồi mọi thứ trôi qua nhanh chóng.",
   "Amanda cũng thoải mái với mối quan hệ này."
]
```

2. Word Segmentation

Usage

```
>>> # -*- coding: utf-8 -*-
>>> from underthesea import word_tokenize
>>> sentence = 'Chàng trai 9X Quảng Trị khởi nghiệp từ nấm sò'
>>> word_tokenize(sentence)
['Chàng trai', '9X', 'Quảng Trị', 'khởi nghiệp', 'từ', 'nấm', 'sò']
>>> word_tokenize(sentence, format="text")
'Chàng_trai 9X Quảng_Trị khởi_nghiệp từ nấm sò'
```

3. POS Tagging

Usage

```
>>> # -*- coding: utf-8 -*-
>>> from underthesea import pos_tag
>>> pos_tag('Chợ thịt chó nổi tiếng ở Sài Gòn bị truy quét')
[('Chợ', 'N'),
 ('thịt', 'N'),
 ('chó', 'N'),
 ('nổi tiếng', 'A'),
 ('ở', 'E'),
 ('Sài Gòn', 'Np'),
 ('bị', 'V'),
 ('truy quét', 'V')]
```

4. Chunking

Usage

```
>>> # -*- coding: utf-8 -*-
>>> from underthesea import chunk
>>> text = 'Bác sĩ bây giờ có thể thản nhiên báo tin bệnh nhân bị ung thư?'
>>> chunk(text)
[('Bác sĩ', 'N', 'B-NP'),
 ('bây giờ', 'P', 'I-NP'),
 ('có thể', 'R', 'B-VP'),
 ('thản nhiên', 'V', 'I-VP'),
 ('báo tin', 'N', 'B-NP'),
 ('bệnh nhân', 'N', 'I-NP'),
 ('bị', 'V', 'B-VP'),
 ('ung thư', 'N', 'I-VP'),
 ('?', 'CH', 'O')]
```

5. Dependency Parsing

Usage

```
>>> # -*- coding: utf-8 -*-
>>> from underthesea import dependency_parse
>>> text = 'Tối 29/11, Việt Nam thêm 2 ca mắc Covid-19'
>>> dependency_parse(text)
[('Tối', 5, 'obl:tmod'),
 ('29/11', 1, 'flat:date'),
 (',', 1, 'punct'),
 ('Việt Nam', 5, 'nsubj'),
 ('thêm', 0, 'root'),
```

```
('2', 7, 'nummod'),
('ca', 5, 'obj'),
('mắc', 7, 'nmod'),
('Covid-19', 8, 'nummod')]
```

6. Named Entity Recognition

Usage

```
>>> # -*- coding: utf-8 -*-
>>> from underthesea import ner
>>> text = 'Chưa tiết lộ lịch trình tới Việt Nam của Tổng thống Mỹ Donald Trump'
>>> ner(text)
[('Chưa', 'R', 'O', 'O'),
 ('tiết lộ', 'V', 'B-VP', 'O'),
 ('lịch trình', 'V', 'B-VP', 'O'),
 ('tới', 'E', 'B-PP', 'O'),
 ('Việt Nam', 'Np', 'B-NP', 'B-LOC'),
 ('của', 'E', 'B-PP', 'O'),
 ('Tổng thống', 'N', 'B-NP', 'O'),
 ('Mỹ', 'Np', 'B-NP', 'B-LOC'),
 ('Donald', 'Np', 'B-NP', 'B-PER'),
 ('Trump', 'Np', 'B-NP', 'I-PER')]
```

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 đầu tiên ở Premier League bị sa thải sau 4 vòng đấu')
['The thao']
>>> classify('Hội đồng tư vấn kinh doanh Asean vinh danh giải thưởng quốc tế')
['Kinh doanh']

>> classify('Lãi suất từ BIDV rất ưu đãi', domain='bank')
['INTEREST_RATE']
```

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 chất lg,chăn đắp lên dính lông lá khắp người. thất vọng')
negative
>>> sentiment('Sản phẩm hơi nhỏ so với tưởng tượng nhưng chất lượng tốt, đóng gói cẩn thận.')
positive

>>> sentiment('Đky qua đường link ở bài viết này từ thứ 6 mà giờ chưa thấy ai lhe hết', domain='bank')
['CUSTOMER_SUPPORT#negative']
>>> sentiment('Xem lại vẫn thấy xúc động và tự hào về BIDV của mình', domain='bank')
['TRADEMARK#positive']
```

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
```

Up Coming Features

- Machine Translation
- Text to Speech
- Automatic Speech Recognition