Basic Python - Work with Text Data

Hoàng-Nguyên Vũ

1. **Mô tả: Làm quen với thư viện Underthesea**

   - **Underthesea** là một bộ công cụ mã nguồn mở hỗ trợ xử lý ngôn ngữ tự nhiên Tiếng Việt (NLP). Nó được phát triển bởi cộng đồng nghiên cứu NLP tại Việt Nam và được công bố lần đầu tiên vào năm 2017.

Bảng 1: Tính năng, ưu điểm của thư viện Underthesea

Tính năng	
Phân chia câu	Cắt một đoạn văn bản thành các câu riêng biệt.
Phân loại từ	Xác định loại từ (danh từ, động từ, tính từ, v.v.) của mỗi từ trong câu.
Gắn thẻ POS	Gán thẻ cho mỗi từ với thông tin ngữ pháp (danh từ, động từ, tính từ, v.v.).
Nhận dạng thực thể tên riêng	Xác định các thực thể tên riêng (người, địa điểm, tổ chức, v.v.) trong văn bản.
Phân loại văn bản	Phân loại văn bản vào các chủ đề hoặc thể loại khác nhau.
Tóm tắt văn bản	Tạo tóm tắt ngắn gọn cho một đoạn văn bản dài.
Trích xuất quan điểm	Xác định các quan điểm và ý kiến trong văn bản.
Dịch máy	Dịch văn bản từ tiếng Việt sang tiếng Anh và ngược lại.
Ưu điểm	
Mã nguồn mở	Có thể sử dụng và sửa đổi miễn phí.
Dễ sử dụng	Cung cấp API đơn giản và dễ hiểu.
Hiệu quả	Đã được chứng minh hiệu quả trên nhiều tập dữ liệu tiếng Việt.
Cộng đồng	Được hỗ trợ bởi cộng đồng nghiên cứu NLP Việt Nam năng động.

   - **Cách cài đặt và sử dụng một số tính năng:**
     - Để cài đặt thư hiện Underthesea, ta sẽ cài thông qua câu lệnh:

```
!pip install underthesea
```

     - Các tính năng nổi bật trong thư viện Underthesea:
       + **Gán nhãn từ loại (POS tagging):**

```
from underthesea import pos_tag
pos_tag('Học sinh đang học toán')
```

       + **Kết quả:** [('Học sinh', 'N'), ('đang', 'R'), ('học', 'V'), ('toán', 'N')]

+ **Phân loại văn bản (Text classification):**

```
from underthesea import classify
classify('giá cổ phiếu đang có nhiều biến động trong thời
    gian qua')
```

+ **Kết quả:** ['kinh_doanh']

+ **Phân tích cảm xúc (Sentiment Analysis):**

```
from underthesea import sentiment
sentiment('Sản phẩm mình đặt về không như quảng cáo')
```

+ **Kết quả:** 'negative'

+ **Phân đoạn câu văn (Sentence Segmentation):**

```
from underthesea import sent_tokenize
text = 'Những đứa trẻ nghèo thế hệ tôi biết đọc biết viết,
    thành người bằng những cuốn giáo khoa đi mượn như thế.
    Cũng có những đứa nhà nghèo quá, không mượn đâu được bộ
     sách cho tử tế, môn được môn không, càng học càng đúp,
     cuối cùng bỏ dở giữa chừng.'
sent_tokenize(text)
```

+ **Kết quả:** ['Những đứa trẻ nghèo thế hệ tôi biết đọc biết viết, thành người bằng những cuốn giáo khoa đi mượn như thế.', 'Cũng có những đứa nhà nghèo quá, không mượn đâu được bộ sách cho tử tế, môn được môn không, càng học càng đúp, cuối cùng bỏ dở giữa chừng.']

+ **Phân đoạn từ ngữ (Word Segmentation):**

```
from underthesea import word_tokenize
sentence = 'Công trình của PGS Vân đã thay thế bạch kim
    trong pin nhiên liệu, giúp giảm giá thành mà pin có độ
    bền cao hơn.'
word_tokenize(sentence)
```

+ **Kết quả:** ['Công trình', 'của', 'PGS Vân', 'đã', 'thay thế', 'bạch kim', 'trong', 'pin', 'nhiên liệu', ',', 'giúp', 'giảm', 'giá thành', 'mà', 'pin', 'có', 'độ', 'bền', 'cao', 'hơn', '.']

2. **Bài tập:**

- Hãy thực hiện các task: POS Tagging, Text Classsification, Sentiment Analysis, Sentence Segmentation, Word Segmentation đoạn văn bản sau:

  Công cụ Suno AI nhanh chóng nhận được sự chú ý từ người dùng khi có thể tạo bài hát chỉ với vài câu lệnh. Phiên bản mới nhất V3 Alpha mới được giới thiệu cuối tháng 2, có bản miễn phí với 10 bài hát mỗi ngày.

  Kết quả:

+ **POS Tagging:** [('Công cụ', 'N'), ('Suno', 'Np'),('AI', 'P'),...]
+ **Text Classsification:** ['vi_tinh']
+ **Sentiment Analysis:** 'positive'
+ **Sentence Segmentation:** ['Công cụ Suno AI nhanh chóng nhận được sự chú ý từ người dùng khi có thể tạo bài hát chỉ với vài câu lệnh.', 'Phiên bản mới nhất V3 Alpha mới được giới thiệu cuối tháng 2, có bản miễn phí với 10 bài hát mỗi ngày.']
+ **Word Segmentation:** ['Công cụ', 'Suno', 'AI', 'nhanh chóng', ...]