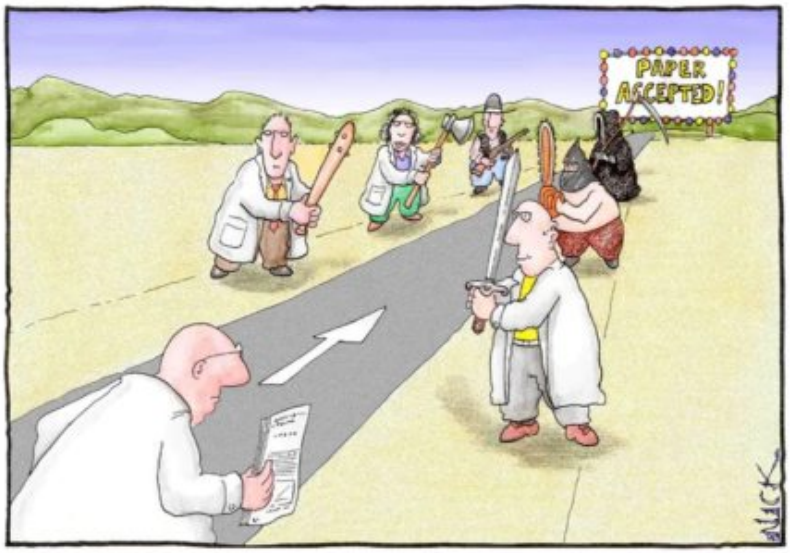
Basic Python - Work with Text Data

Hoàng-Nguyên Vũ

1. **Mô tả: Làm quen với thư viện newspaper3k và nlpk**



- **Thư viện newspaper3k** là một thư viện Python mã nguồn mở giúp bạn trích xuất dữ liệu từ các bài báo trực tuyến. Thư viện này hỗ trợ nhiều trang web tin tức khác nhau, bao gồm: VnExpress, Tuổi Trẻ, Thanh Niên, Zing News, VTC News, ... Các tính năng nổi bật của thư viện bao gồm:
  + **Trích xuất dữ liệu:** Newspaper3k có thể trích xuất nhiều loại dữ liệu từ các trang web báo chí, bao gồm tiêu đề, bài viết, tóm tắt, tác giả, ngày tháng, hình ảnh, video, v.v.
  + **Hỗ trợ nhiều trang web:** Newspaper3k hỗ trợ hơn 100 trang web báo chí khác nhau, bao gồm cả các trang web tiếng Việt như VnExpress, Tuổi Trẻ, Thanh Niên, v.v.
  + **Dễ sử dụng:** Newspaper3k cung cấp một API đơn giản để trích xuất dữ liệu từ các trang web báo chí.
  + **Mã nguồn mở:** Newspaper3k là một thư viện mã nguồn mở, vì vậy bạn có thể sử dụng và sửa đổi nó miễn phí.
- **Thư viện nltk (Natural Language Toolkit)** là một thư viện mã nguồn mở được phát triển bởi Python. Nó cung cấp một bộ công cụ mạnh mẽ để xử lý ngôn ngữ tự nhiên (NLP) trong Python. Các tính năng chính của thư viện bao gồm:
  + **Phân tích cú pháp:** NLTK có thể phân tích cú pháp của các câu tiếng Anh để xác định cấu trúc và thành phần của chúng.
  + **Phân loại từ:** NLTK có thể xác định loại từ (danh từ, động từ, tính từ, v.v.) của các từ trong một câu.

+ **Gán nhãn ngữ nghĩa:** NLTK có thể gán nhãn ngữ nghĩa (tên riêng, địa điểm, tổ chức, v.v.) cho các từ trong một câu.
+ **Tóm tắt văn bản:** NLTK có thể tóm tắt các văn bản dài thành các văn bản ngắn hơn.
+ **Dịch máy:** NLTK có thể dịch văn bản từ ngôn ngữ này sang ngôn ngữ khác.

- **Cách cài đặt và sử dụng một số tính năng:**
  - Để cài đặt thư hiện thư viện newspaper3k, ta sẽ cài thông qua câu lệnh:

```
!pip install newspaper3k
!pip install nltk
```

  - Cách sử dụng các tính năng chính của thư viện:
    + **Thư viện newspaper3k:**

```
from newspaper import Article

# Tạo một đối tượng Article từ URL của bài báo
article = Article('https://vnexpress.net/thoi-tiet-mien-bac-mien-trung-mien-nam-ngay-14-3-4518045.html')

# Tải bài báo
article.download()
article.parse()

# In bài báo
print(article.text)

# Lấy toàn bộ ảnh trong bài báo
print(article.images)
```

    + **Kết quả:** Tòa án quân sự Mỹ thông báo Ryan Mays, thủy thủ bị tố đốt tàu đổ bộ USS Bonhomme Richard, được trắng án....

      {Link đường dẫn ảnh của bài báo...}

    + **Thư viện nltk:**

```
import nltk
from nltk.tokenize import word_tokenize

nltk.download('punkt')

data = "Tôi thích học AI và Toán"
# Bước 1: Tokenization data
tokenization = word_tokenize(data)
# Bước 2: Gọi thư viện Pos tagging
result = nltk.pos_tag(tokenization)
print(result)
```

    + **Kết quả:** [('Tôi', 'NNP'), ('thích', 'NN'), ('học', 'NN'), ('AI', 'NNP'), ('và', 'NN'), ('Toán', 'NNP')]

2. **Bài tập:**

- **Câu 1:** Thực hiện đọc và tóm tắt bài báo tại đường dẫn sau: Bài báo VnExpress
- **Câu 2:** Thực hiện pos tagging bài báo trên với thư viện NLTK.

Kết quả:

+ Câu 1: Ngày 13/3, Cognition Labs, startup về công nghệ trí tuệ nhân tạo tại Mỹ, công bố kỹ sư phát triển phần mềm AI đầu tiên trên thế giới. Với Devin, các kỹ sư có thể tập trung vào những vấn đề thú vị hơn, các đội kỹ thuật có thể nỗ lực cho những mục tiêu tham vọng hơn", Cognition cho biết...

+ Câu 2: [('Kỹ', 'NNP'), ('sư', 'NN'), ('phần', 'NN'), ('mềm', 'NN'), ...]