**LAB 7: PHÂN TÍCH DỮ LIỆU DẠNG VĂN BẢN VỚI NLTK**

**1. Giới thiệu về thư viện NLTK:**

A screenshot of a computer program

AI-generated content may be incorrect.

A close-up of a computer screen

AI-generated content may be incorrect.



A screenshot of a computer

AI-generated content may be incorrect.

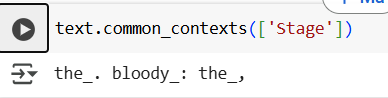
A screenshot of a computer code

AI-generated content may be incorrect.

**2. Tìm 1 từ với NLTK:**

A screenshot of a computer

AI-generated content may be incorrect.



A screenshot of a computer

AI-generated content may be incorrect.

**3. Phân tích tần số của các từ**

A screenshot of a computer code

AI-generated content may be incorrect.

A close-up of a computer screen

AI-generated content may be incorrect.

A screen shot of a computer code

AI-generated content may be incorrect.

A close-up of a computer screen

AI-generated content may be incorrect.

A screenshot of a computer code

AI-generated content may be incorrect.

A computer screen shot of a computer code

AI-generated content may be incorrect.

**4. Lựa chọn các từ trong văn bản**

A screenshot of a computer code

AI-generated content may be incorrect.

A screenshot of a computer screen

AI-generated content may be incorrect.

**5. Bigrams và collocations**

A screenshot of a computer program

AI-generated content may be incorrect.

A screenshot of a computer code

AI-generated content may be incorrect.

**6. Sử dụng văn bản trên mạng**

A computer screen shot of a computer code

AI-generated content may be incorrect.

A computer code with numbers and text

AI-generated content may be incorrect.

A screenshot of a computer code

AI-generated content may be incorrect.

**7. Rút trích văn bản từ trang html**

A screen shot of a computer code

AI-generated content may be incorrect.

A computer code with black text

AI-generated content may be incorrect.

**8. Phân tích cảm xúc người dùng**

A screen shot of a computer

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

A close-up of a computer screen

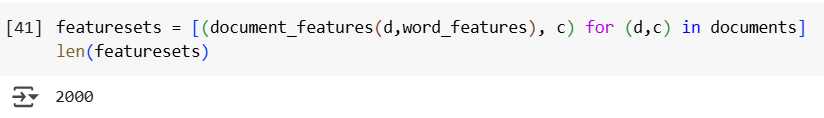
AI-generated content may be incorrect.

A close-up of words

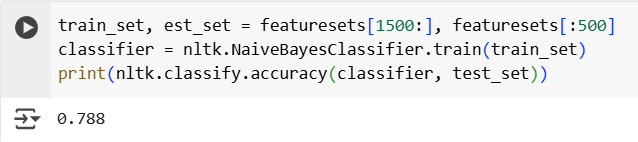
AI-generated content may be incorrect.

A computer code with text

AI-generated content may be incorrect.





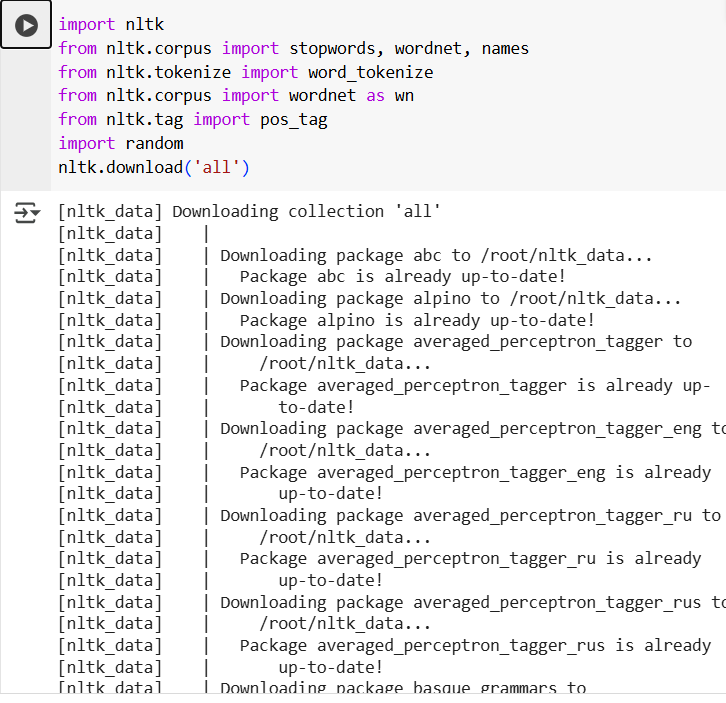


A screenshot of a computer

AI-generated content may be incorrect.

**9. Bài tập áp dụng**

- Cài các thư viện cần thiết



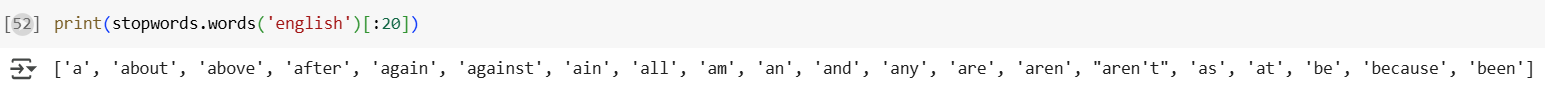
1. Viết chương trình Python với thư viện NLTK để liệt kê các tên của copus.



2. Viết chương trình Python với thư viện NLTK để liệt kê danh sách các stopword bằng các ngôn ngữ khác nhau.



3. Viết chương trình Python với thư viện NLTK để kiểm tra danh sách các stopword bằng các ngôn ngữ khác nhau.



4. Viết chương trình Python với thư viện NLTK để loại bỏ các stopword từ một văn bản đã cho.

A computer screen shot of text

AI-generated content may be incorrect.

5. Viết chương trình Python với thư viện NLTK bỏ qua các stopword từ danh sách các stopword.

A screenshot of a computer

AI-generated content may be incorrect.

6. Viết một chương trình Python với thư viện NLTK để tìm định nghĩa và ví dụ của một từ đã cho bằng WordNet từ Wikipedia,

A computer screen shot of a computer code

AI-generated content may be incorrect.

7. Viết chương trình Python với thư viện NLTK để tìm tập hợp các từ đồng nghĩa và trái nghĩa của một từ nào đó.

A computer screen shot of a computer

AI-generated content may be incorrect.

8. Viết chương trình Python với thư viện NLTK để có cái nhìn tổng quan về bộ tag, chi tiết của một tag cụ thể trong bộ tag và chi tiết về một số bộ tag liên quan, sử dụng biểu thức chính quy.

A screenshot of a computer program

AI-generated content may be incorrect.

9. Viết chương trình Python với thư viện NLTK để so sánh sự giống nhau của hai danh từ đã cho.

A computer code with red and blue text

AI-generated content may be incorrect.

10. Viết chương trình Python với thư viện NLTK để so sánh sự giống nhau của hai động từ đã cho.

A screenshot of a computer

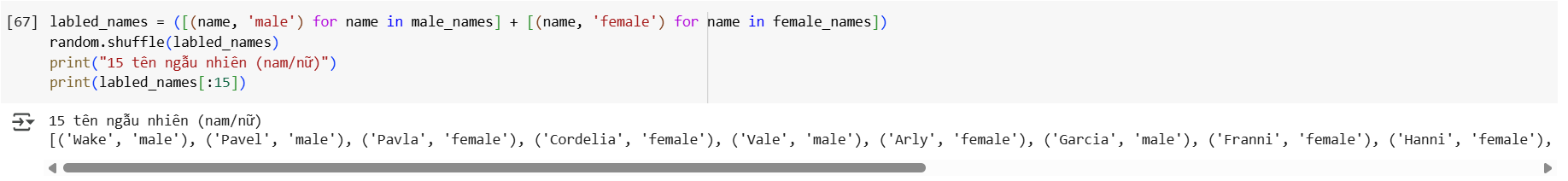
AI-generated content may be incorrect.

11. Viết chương trình Python với thư viện NLTK để tìm số lượng tên nam và nữ trong các tên kho ngữ liệu. In tên 10 nam và nữ đầu tiên. Lưu ý: Kho văn bản tên chứa tổng cộng khoảng 2943 nam (male.txt) và 5001 nữ (Female.txt) tên. Kho được biên soạn bởi Kantrowitz, Ross.

A screenshot of a computer code

AI-generated content may be incorrect.

12. Viết chương trình Python với thư viện NLTK để in 15 kết hợp ngẫu nhiên đầu tiên được gắn nhãn nam và được gắn nhãn tên nữ từ kho tên.

****

13. Viết chương trình Python với thư viện NLTK để trích xuất ký tự cuối cùng của tất cả các tên được gắn nhãn và tạo mảng mới với chữ cái cuối cùng của mỗi tên và nhãn được liên kết

