TUẦN 2: TIỀN XỬ LÝ TRANG WEB ĐƠN (Phần 1)

1. Mục tiêu:

Sinh viên tìm hiểu và sử dụng các thư viện sẵn có của Python để thực hiện tiền xử lý văn bản web trên ngôn ngữ tiếng Anh:

- Làm sạch (loại bỏ các thẻ HTML, loại bỏ các ký tự đặc biệt)
- Tách câu, tách từ (Sentence Tokenize, word tokenize)
- Loại bỏ hư từ (stopwords)
- Chuẩn hóa từ (Stemming and Lemmatization)

2. Nội dung:

- Đầu vào: Tập dữ liệu các nội dung trang web. Nội dung mỗi trang web được lưu dưới dạng 1 file .txt.
- Xử lý: thực hiện các bước cần thiết để có được danh sách các từ (đọc nội dung từng file→làm sạch →tách câu → tách từ→loại bỏ hư từ → chuẩn hóa từ → ghi từ vựng thu được ra file)
- Yêu cầu đối với chương trình:
    - Cho phép người dùng chọn đường dẫn thư mục đầu vào, chương trình tự động xử lý mọi tập tin *.txt nằm trong thư mục đầu vào và thư mục con của thư mục đầu vào.
    - Cho phép người dùng lựa chọn thư mục đầu ra, mọi các tập tin đầu ra được lưu trữ trong thư mục này (không cần theo cấu trúc giống như thư mục đầu vào)
    - Danh sách từ của văn bản sau khi xử lý được ghi vào tập tin tương ứng có phần mở rộng là <original name>_word.txt.

  Vd: tập tin đầu vào là abc1.txt thì tập tin đầu ra là abc1_word.txt
- Yêu cầu nộp bài:

  Sinh viên nộp file nén MSSV.rar/zip trong đó bao gồm:

  + File source code: MSSV.py

  + File Readme: Ghi chú các hướng dẫn để chạy chương trình.
- Sinh viên cần chuẩn bị một thư mục các file test và sẵn sàng chạy chương trình khi giáo viên yêu cầu.

3. Hướng dẫn thực hiện

1. **Đọc file từ thư mục**:

Sinh viên tìm hiểu các module: os, pathlib hoặc một số module khác của Python để thực hiện bước này.

- Duyệt và lấy danh sách file thuộc một thư mục bằng os.walk. Trong đó path là biến lưu đường dẫn đến thư mục muốn thao tác.

```
import os
list_path=[]
for root, dirs, files in os.walk(path):
    for file in files:
        list_path.append(root+"/"+file)
```

- Từ danh sách đường dẫn đến các file thu được, lần lượt duyệt qua từng file, đọc thông tin và thực hiện tiền xử lý.

```
i=0
for i in range (len(list_path)):
    read_file=open(list_path[i],"r")
    a=read_file.read()
    #xử lý...
```

2. Tiền xử lý dữ liệu

Ở các bước tiền xử lý tiếp theo, sinh viên có thể tìm hiểu một số thư viện hỗ trợ của python như: re, BeautifulSoup và nltk:

```
import re # import thư viện re
from bs4 import BeautifulSoup # import thư viện BeautifulSoup
import nltk # import thư viên nltk
from string import punctuation #import tập dấu câu từ thư viện string
from nltk.corpus import stopwords # import tập hư từ
from nltk.tokenize import word_tokenize, sent_tokenize # import các hàm xử lý tách từ, tách câu
my_stopwords = set(stopwords.words('english') + list(punctuation))
```

a. Làm sạch dữ liệu:

- Lấy dữ liệu thừ file:

```
#Lấy dữ liệu từ file
def get_text(file):
    read_file=open(file,"r")
    text=read_file.readlines()
    text = ' '.join(text)
    return text
```

- Xóa các thẻ html trong file:

```
#Loại bỏ các thẻ của html trong file
def clean_htlm(text):
    soup=BeautifulSoup(text,'html.parser')
    return soup.get_text()
```

- Xóa các ký tự đặc biệt và chuẩn hóa các khoảng trắng thừa

```
#Loại bỏ các ký tự đặc biệt
def remove_special_character(text): # text là một string
    #Thay thế các ký tự đặc biệt bằng ''
    string = re.sub('[^\w\s]','',text)
    #Xử lý các khoảng trắng thừa ở giữa chuỗi
    string = re.sub('\s+', ' ', string)
    #Xử lý các khoảng trắng thừa ở đầu và cuối chuỗi
    string = string.strip()
    return string
```

b. Tách câu, tách từ, loại bỏ hư từ:

```
my_stopwords = set(stopwords.words('english') + list(punctuation))
i=0
for i in range (len(list_path)):
    text = get_text(list_path[i])
    text_cleaned = clean_html(text)
    #Tách câu
    sents = sent_tokenize(text_cleaned)
    #Loại bỏ ký tự đặc biệt trong câu
    sents_cleaned = [remove_special_character(s) for s in sents]
    #Nối các câu lại thành text
    text_sents_join = ''.join(sents_cleaned)
    #Tách từ
    words = word_tokenize(text_sents_join)
    #Đưa về dạng chữ thường
    words = [word.lower() for word in words]
    #Loại bỏ hư từ
    words = [word for word in words if word not in my_stopwords]
```

c. Chuẩn hóa từ:

```
from nltk.stem import PorterStemmer
    ps = PorterStemmer()
    words = [ps.stem(word) for word in words]
```

3. Ghi file ra thư mục

Sinh viên tự tìm hiểu