webSummerizer

Hong Huy Hoang

November 2024

1 Introduction

My second project using LLM, using Google's Gemini generative AI

2 Requirements

- language Python 3.11.10
- beautifulsoup4 4.12.2
- ipython 8.15.0
- requests 2.31.0
- google-generative-ai 0.3.0

3 Setup

I recommend using miniconda, although every other python virutal env managers will work fine

The command bellow will install a conda environtment. You can change that of the environment in environtment.yml before running it

```
conda env create -f environment.yml
conda activate websummerizer
```

4 How this works?

4.1 Load Environment and API from Google AI Studio

```
#load env
import requests
import google.generativeai as genai
```

```
import os
from bs4 import BeautifulSoup
from ipykernel.kernelapp import IPKernelApp
from IPython.display import Markdown, display
```

```
#load API
os.environ["GOOGLE_API_KEY"] = "your_api_here"
genai.configure(api_key="your_api_here")
generation_config = {"temperature" : 0.9, "top_p" : 1, "top_k" : 1}
model = genai.GenerativeModel("gemini-pro", generation_config=generation_config)
```

4.2 Define a class represent a webpage

```
#define a class represent a webpage
#beautifulSoup is used for web scanning and extract information from html

class Website:
    url : str
    title : str
    text : str

    def __init__(self, url):
        self.url = url
        response = requests.get(url)
        soup = BeautifulSoup(response.text, 'html.parser')
        self.title = soup.title.string if soup.title else "No title found"

        #remove unnecessary things
        for irrelevant in soup.body(["script", "style", "img", "input"]):
            irrelevant.decompose()

        #strip text
        self.text = soup.body.get_text(separator = "\n", strip = True)

```

4.3 Define system prompt and user prompt for AI to know more about context

System prompt

```
system_prompt = "You are an assistant that analyzes the contents of a website \
    and provides a short summary, ignoring text that might be navigation related. \
```

```
    Respond in markdown."

```

User prompt

```
    def user_prompt_for(website):
    user_prompt = f"You are looking at a website titled {website.title}"
    user_prompt += "The contents of this website is as follows; \
        please provide a short summary of this website in markdown.\
        If it includes news or annoucements, then summerize these too \n\n"
    user_prompt += website.text
    return user_prompt
```

4.4 Time to bring all together, very simple!

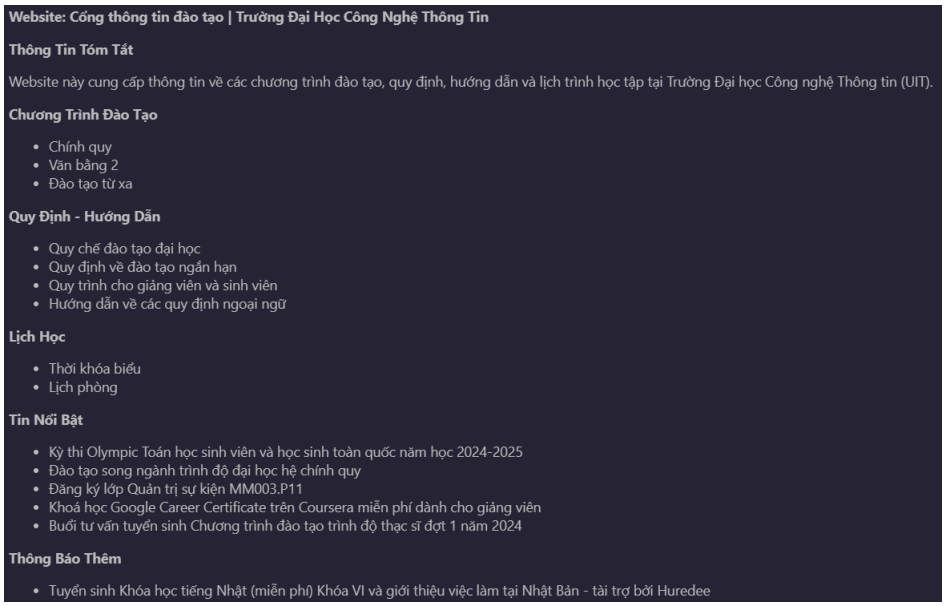
```
    def summarize(url):
    website = Website(url)
    full_prompt = f"System: {system_prompt} User: {user_prompt_for(website)}"
    response = model.generate_content(full_prompt)
    return response.text
```

4.5 Testing Gemini AI model

Example url: https://daa.uit.edu.vn/

```
    summarize('https://daa.uit.edu.vn/')
    display_summary('https://daa.uit.edu.vn/')
```

Output would be something like this



Website: Cổng thông tin đào tạo | Trường Đại Học Công Nghệ Thông Tin

Thông Tin Tóm Tắt

Website này cung cấp thông tin về các chương trình đào tạo, quy định, hướng dẫn và lịch trình học tập tại Trường Đại học Công nghệ Thông tin (UIT).

Chương Trình Đào Tạo

- Chính quy
- Văn bằng 2
- Đào tạo từ xa

Quy Định - Hướng Dẫn

- Quy chế đào tạo đại học
- Quy định về đào tạo ngắn hạn
- Quy trình cho giảng viên và sinh viên
- Hướng dẫn về các quy định ngoại ngữ

Lịch Học

- Thời khóa biểu
- Lịch phòng

Tin Nổi Bật

- Kỳ thi Olympic Toán học sinh viên và học sinh toàn quốc năm học 2024-2025
- Đào tạo song ngành trình độ đại học hệ chính quy
- Đăng ký lớp Quản trị sự kiện MM003.P11
- Khoá học Google Career Certificate trên Coursera miễn phí dành cho giảng viên
- Buổi tư vấn tuyển sinh Chương trình đào tạo trình độ thạc sĩ đợt 1 năm 2024

Thông Báo Thêm

- Tuyển sinh Khóa học tiếng Nhật (miễn phí) Khóa VI và giới thiệu việc làm tại Nhật Bản - tài trợ bởi Huredee

Figure 1: Example of output