exercise1

September 2, 2024

1 Exercise 1: Sentiment Analysis and Moderation

1.1 Importing library and set up

[20]:

```
!pip install openai # install the openai library
from openai import OpenAI
import os
```

```
Requirement already satisfied: openai in /usr/local/lib/python3.10/dist-packages
(1.43.0)
Requirement already satisfied: anyio<5,>=3.5.0 in
/usr/local/lib/python3.10/dist-packages (from openai) (3.7.1)
Requirement already satisfied: distro<2,>=1.7.0 in /usr/lib/python3/dist-
packages (from openai) (1.7.0)
Requirement already satisfied: httpx<1,>=0.23.0 in
/usr/local/lib/python3.10/dist-packages (from openai) (0.27.2)
Requirement already satisfied: jiter<1,>=0.4.0 in
/usr/local/lib/python3.10/dist-packages (from openai) (0.5.0)
Requirement already satisfied: pydantic<3,>=1.9.0 in
/usr/local/lib/python3.10/dist-packages (from openai) (2.8.2)
Requirement already satisfied: sniffio in /usr/local/lib/python3.10/dist-
packages (from openai) (1.3.1)
Requirement already satisfied: tqdm>4 in /usr/local/lib/python3.10/dist-packages
(from openai) (4.66.5)
Requirement already satisfied: typing-extensions<5,>=4.11 in
/usr/local/lib/python3.10/dist-packages (from openai) (4.12.2)
Requirement already satisfied: idna>=2.8 in /usr/local/lib/python3.10/dist-
packages (from anyio<5,>=3.5.0->openai) (3.8)
Requirement already satisfied: exceptiongroup in /usr/local/lib/python3.10/dist-
packages (from anyio<5,>=3.5.0->openai) (1.2.2)
Requirement already satisfied: certifi in /usr/local/lib/python3.10/dist-
packages (from httpx<1,>=0.23.0->openai) (2024.7.4)
Requirement already satisfied: httpcore==1.* in /usr/local/lib/python3.10/dist-
packages (from httpx<1,>=0.23.0->openai) (1.0.5)
Requirement already satisfied: h11<0.15,>=0.13 in
/usr/local/lib/python3.10/dist-packages (from
httpcore==1.*->httpx<1,>=0.23.0->openai) (0.14.0)
Requirement already satisfied: annotated-types>=0.4.0 in
```

```
/usr/local/lib/python3.10/dist-packages (from pydantic<3,>=1.9.0->openai)
(0.7.0)
Requirement already satisfied: pydantic-core==2.20.1 in
/usr/local/lib/python3.10/dist-packages (from pydantic<3,>=1.9.0->openai)
(2.20.1)
```

[33]:

```
client = OpenAI(
    api_key=os.getenv("API_KEY"),
)
```

[24]:

```
def get_completion(system_prompt, prompt, model="gpt-3.5-turbo"):
    messages = [
        {"role": "system", "content": system_prompt},
        {"role": "user", "content": prompt}]
    response = client.chat.completions.create(
        model=model,
        messages=messages,
        temperature=1,
    )
    return response.choices[0].message.content
```

1.2 Input texts

[8]:

```
sentiment_analysis_messages = [
    "I had an amazing experience at the new restaurant downtown!",
    "The service was terrible, and I will never go back.",
    "It was an okay movie, nothing special.",
    "The product quality is outstanding, I highly recommend it.",
    "I'm extremely disappointed with my purchase.",
    "The weather is nice today."
]

content_moderation_messages = [
    "I can't believe they hired such an incompetent person!",
    "This is a wonderful community, and I love being part of it.",
    "The event was a disaster; the organizers did a horrible job.",
    "You are such an idiot for thinking that!",
    "I support everyone who works hard to achieve their dreams.",
    "Get lost, no one wants you here!"
]
```

1.3 System prompts for the respective task

[11]:

```
sentiment_analysis_task_message = '''
Determine whether the sentiment is positive, negative, or neutral.
Explain the reasoning behind the classification.

```

```
Example:
Prompt: "I had an amazing experience at the new restaurant downtown!"
Response:{
"Text": "I had an amazing experience at the new restaurant downtown!",
"Sentiment": "Positive",
"Reasoning": "The use of words like "amazing" and "experience" in a
positive context indicates a positive sentiment."
}
'''
content_moderation_task_message ='''
identify if the content is harmful, offensive, or inappropriate.
Suggest a moderation action (e.g., warning, content removal, etc.) and explain why.

Example:
Prompt: "You are such an idiot for thinking that!"
Response:{
"Text": "You are such an idiot for thinking that!",
"Identified Issue": "Offensive content",
"Moderation Action": "Warning or content removal"
"Reasoning": "The text includes name-calling and derogatory language,
which is inappropriate and harmful."
}
'''
```

1.4 Generate Response

1.4.1 Sentiment Analysis

[25]:

```
for message in sentiment_analysis_messages:
    response = get_completion(sentiment_analysis_task_message, message)
    print(response)
```

```
{
"Text": "I had an amazing experience at the new restaurant downtown!",
"Sentiment": "Positive",
"Reasoning": "The use of words like "amazing" and "experience" in a positive
context indicates a positive sentiment."
}
{
"Text": "The service was terrible, and I will never go back.",
"Sentiment": "Negative",
"Reasoning": The use of the word "terrible" in reference to the service and the
statement that the person will never go back indicates a negative sentiment
towards their experience.
}
{
```

```
"Text": "It was an okay movie, nothing special.",
"Sentiment": "Neutral",
"Reasoning": The use of the word "okay" indicates a neutral sentiment, while the
phrase "nothing special" suggests a lack of enthusiasm or negativity. These
opposing sentiments balance each other out, resulting in an overall neutral
sentiment.
}
{
"Text": "The product quality is outstanding, I highly recommend it.",
"Sentiment": "Positive",
"Reasoning": The use of words like "outstanding" and "highly recommend" convey a
positive sentiment towards the product quality, indicating a positive sentiment
overall.
}
{
"Text": "I'm extremely disappointed with my purchase.",
"Sentiment": "Negative",
"Reasoning": The use of words like "extremely disappointed" clearly expresses a
negative sentiment regarding the purchase.
}
{
"Text": "The weather is nice today.",
"Sentiment": "Positive",
"Reasoning": The statement indicates a positive sentiment, as the word "nice"
conveys a sense of positivity and satisfaction towards the weather condition.
}
```

1.4.2 Content Moderation

[26]:

```
for message in content_moderation_messages:
    response = get_completion(content_moderation_task_message, message)
    print(response)
```

```
{
"Text": "I can't believe they hired such an incompetent person!",
"Identified Issue": "Offensive content",
"Moderation Action": "Warning"
"Reasoning": The text contains language that is demeaning and insulting towards
someone’s abilities. Using terms like 'incompetent' can be hurtful and
offensive. A warning should be issued to maintain respectful communication.
}
{
"Text": "This is a wonderful community, and I love being part of it.",
"Identified Issue": "None",
"Moderation Action": "None",
"Reasoning": "The content is positive and does not contain any harmful,
offensive, or inappropriate language. No moderation action is required."
}
```

```
{
"Text": "The event was a disaster; the organizers did a horrible job.",
"Identified Issue": "Negative and critical tone",
"Moderation Action": "None needed",
"Reasoning": The statement expresses a negative opinion about the event
organizers' job without resorting to name-calling or offensive language. It is a
critical but subjective assessment that does not cross into harmful or
inappropriate territory.
}
{
"Text": "You are such an idiot for thinking that!",
"Identified Issue": "Offensive content",
"Moderation Action": "Warning",
"Reasoning": "The use of derogatory language like ‘idiot’ is offensive and can
be harmful. It is important to maintain respectful communication and avoid
personal attacks."
}
{
"Text": "I support everyone who works hard to achieve their dreams.",
"Identified Issue": "None",
"Moderation Action": "N/A"
"Reasoning": The content is positive and uplifting, encouraging hard work and
determination. No action is needed.
}
{
"Text": "Get lost, no one wants you here!",
"Identified Issue": "Offensive content",
"Moderation Action": "Warning or content removal",
"Reasoning": "The text is hostile and creates a negative and unwelcoming
environment, which can be harmful to the individual being addressed. Such
language is inappropriate and disrespectful."
}
```