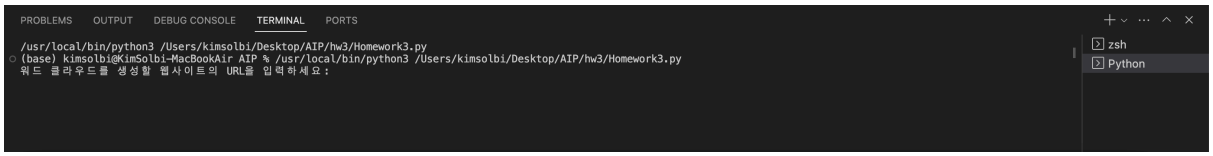
Homework 3 프로그램 소개 문서

<웹 크롤링을 기반으로 한 워드 클라우드 생성>

먼저 프로그램을 실행하면 "워드 클라우드를 생성할 웹사이트의 URL을 입력하세요:" 라는 메시지가 출력됩니다.



웹 페이지의 URL을 입력하면, 프로그램이 해당 웹 페이지의 본문을 분석하여 워드 클라우드를 생성해줍니다.

결과 화면은 sherlock.png 이미지를 마스크로 가져와서 해당 사진에 워드 클라우드가 생성됩니다.

마스크로 사용한 이미지는 다음과 같습니다.

<코드 분석>

```
from selenium import webdriver
from selenium.webdriver.chrome.service import Service
from selenium.webdriver.chrome.options import Options
from webdriver_manager.chrome import ChromeDriverManager
from selenium.webdriver.common.by import By
import wordcloud
import matplotlib.pyplot as plt
import numpy as np
from PIL import Image

# 크롬 옵션 설정
chrome_options = Options()
chrome_options.add_argument("--headless")

# 웹 드라이버 설치 및 실행
service = Service(ChromeDriverManager().install())
driver = webdriver.Chrome(service=service, options=chrome_options)

# URL 입력 받기
url = input("워드 클라우드를 생성할 웹사이트의 URL 을 입력하세요: ")
driver.get(url)

# 페이지에서 본문 텍스트 가져오기
conan_data = driver.find_element(By.XPATH,
'//*[@id="JqFElvLbD"]/div[2]/div/div/div/div/div/div[1]/div[7]/div/div/
div/div/div/div/div[2]/div/div/div/div/div/div/div[11]/div/div/div/div/
div[1]/div/div[23]/div').text

# 필요없는 단어들 목록 만들기
s_words =
wordcloud.STOPWORDS.union({'네이버','카페','지부','온리전','공식','사이트는'
,'셈','티켓값이','교보문고','감청의 권'})

# 워드 클라우드 생성
output_filename = "conan_data_sherlock.png"
sherlock_data =
Image.open("/Users/kimsolbi/Desktop/AIP/hw3/sherlock.png")
sherlock_mask = np.array(sherlock_data)

wordcloud_instance = wordcloud.WordCloud(
```

```
    background_color="white",
    max_words=2000,
    mask=sherlock_mask,
    stopwords=s_words,
    min_font_size=10,
    max_font_size=100,
    font_path='/System/Library/Fonts/Supplemental/AppleGothic.ttf',
    width=1000,
    height=700
).generate(conan_data)

wordcloud_instance.to_file(output_filename)

plt.figure(figsize=(40,30))
plt.imshow(wordcloud_instance, interpolation="bilinear")
plt.axis("off")
plt.show()

driver.quit()
```

이때 사용한 url은
https://namu.wiki/w/%EB%AA%85%ED%83%90%EC%A0%95%20%EC%BD%94%EB%82%9C 입니다.
해당 사이트에서 <7.3. 한국에서> 부분을 xpath로 설정하여 크롤링 하였습니다.