**Final project process book**

**Xiao Zhou**

**Date of entry:** 10 May 2025

**What I’ve worked on:**

I tried to find a good topic and idea. I found that when reporting news, there is some very subtle differences in language usage. The words used, might have some systematic differences and bias toward different countries and/or regions. When talking about climate justice, I feel like the report and news also should be considered. So I want to use this final project to examine the potential differences in reporting climate news. Maybe the frequencies of reporting, maybe the slightly difference in sentiment.

I started to check what traditional news media can be used for this project. Initially I plan to use multiple news media to make a comparison, to see if different media will have different “attitudes” towards different countries and regions. However, after searching, I found that some of the news website have anti-scraping mechanisms, which is a bit hard for me to cope. And Some of the news they require subscription and membership to view some of the articles. Finally, I decided to use BBC for my final project. It is open access and there are many examples online about scraping news from it.

**What problems I encountered:**

When thinking of the ideal and start to read the news, I found a problem: that is: in the news, sometimes, in one sentence, there are multiple countries or regions involved. I feel like it is hard to split them when deciding which one the sentiments is meant towards. It will cause some error if I do this. And there might be comments from experts from another country, which will affect the analysis precision.

**What I learned:**

BBC news is open access and easy to scrape.

**Which resources did I use:**

BBC news: <https://www.bbc.com/news>

Some BBC news scraper examples: (haven’t tried but at least the examples show that this is doable.)

<https://medium.com/@mohsin.shaikh324/project-3-scrape-2000-bbc-news-using-python-e8ef8d2c8e21>

<https://github.com/LuChang-CS/news-crawler>

<https://github.com/Ruchi2507/BBC_NEWS_CRAWLER>

---------------------------------------------------------------------

**Date of entry:** 11 May 2025

**What I’ve worked on:**

I am working on how to do sentiment analysis and determine the country they are really meant for.

I thought about ChatGPT, it is very handy at context analysis. I heard about the API tokens but didn’t know how that works. So, I start to search about this. I also asked ChatGPT how this works.

**What problems I encountered:**

The API tokens seems require fees. I don’t know how much it would cost for this project, maybe I have to ensure the project is not that big to make it affordable…

**What I learned:**

BBC news is open access and easy to scrape.

**Which resources did I use:**

ChatGPT website for pricing: <https://openai.com/api/pricing/?utm_source=chatgpt.com>

<span class="container\_\_headline-text" data-editable="headline" data-zjs="click" data-zjs-href="https://www.cnn.com/2024/05/26/weather/extreme-heat-risk-tracker-dg" data-zjs-component\_position="1" data-zjs-canonical\_url="https://www.cnn.com/search" data-zjs-cms\_id="cms.cnn.com/\_pages/search@published" data-zjs-traits-search\_term="climate solution" data-zjs-traits-search\_results\_number="1421">US heat risk forecast: See where Americans are at the most risk for extreme heat this year</span>

--------------------------------------------------------------

**Date of entry:** 13 May 2025

**What I’ve worked on:**

I started coding and writing scrapers. It did not work very well. I searched the websites and chatgpt, it said that bbc news websites might be javascript, so I should use the selenium library.

**What problems I encountered:**

I tried many times, so I changed for almost six to eight hours before it can actually work.

But I cannot turn pages.

**What I learned:**

The selenium library. It can open a new page while running the python program.

**Which resources did I use:**

ChatGPT website.

--------------------------------------------------------------

**Date of entry:** 16 May 2025

**What I’ve worked on:**

I still cannot turn page on BBC, but I find another way to make it work. I turn the pages manually, the scraper can scrape news from each page.

**What problems I encountered:**

There is no big problem. I still need to write another scraper for another news website maybe.

**What I learned:**

Mannual turning pages…

**Which resources did I use:**

I used ChatGPT website for debugging.

--------------------------------------------------------------

**Date of entry:** 18 May 2025

**What I’ve worked on:**

I write another scraper, from the csv file in which I scraped from BBC news list and includes BBC news title and link, I get the whole articles.

**What problems I encountered:**

There’s no much problem today.

**What I learned:**

I get the news articles in a new file.

**Which resources did I use:**

I used ChatGPT website for debugging.

--------------------------------------------------------------

**Date of entry:** 19 May 2025

**What I’ve worked on:**

I write new scrapers, modified from the two BBC scrapers. CNN news websites has clear page structure, so I can now change page automatically.

**What problems I encountered:**

There’s no much problem today.

**What I learned:**

I find the webpage structure in the link. This is where the page structure is hide.

**Which resources did I use:**

I used ChatGPT website for debugging.

--------------------------------------------------------------

**Date of entry:** 20 May 2025

**What I’ve worked on:**

I modified the codes in cnn scraper to clean some of the unrelated information.

**What problems I encountered:**

There’s no much problem today.

**What I learned:**

I find the webpage structure in the link. This is where the page structure is hide.

**Which resources did I use:**

I used ChatGPT website for debugging.

--------------------------------------------------------------

**Date of entry:** 22 May 2025

**What I’ve worked on:**

I use the vaderSentiment.vaderSentiment library for sentiment analysis. And I also use the Spacy library to identify country or location names.

**What problems I encountered:**

I feel like the workload is enough, so maybe I should quit the part of the plan to use chatgpt to use analysis.

**What I learned:**

I learnt the vaderSentiment.vaderSentiment library. It can be used for sentiment analysis.

**Which resources did I use:**

I used ChatGPT website for debugging. And also chatgpt for advice about what library to use to do sentiment analysis.

--------------------------------------------------------------

**Date of entry:** 23 May 2025

**What I’ve worked on:**

I modified some of the codes in cnn\_content.py, which scrape news contents. I modified it to rule out unrelated words like"© 2025 Cable News Network", "CNN Sans ™ & © 2016", "Related article"

**What problems I encountered:**

I feel like the workload is enough, so maybe I should quit the part of the plan to use chatgpt to use analysis.

**What I learned:**

Rule out unrelated information

**Which resources did I use:**

I used ChatGPT website for debugging. And also chatgpt for advice about what library to use to do sentiment analysis.

--------------------------------------------------------------

**Date of entry:** 24 May 2025

**What I’ve worked on:**

Visualization of the final page and made the powerpoint for final presentation

**What problems I encountered:**

I feel like the workload is enough, I was planning to do some more processing on the sentiment csv on identifying locations, but later give up.

**What I learned:**

Not really…

**Which resources did I use:**

I used ChatGPT website for debugging. And also, chatgpt for advice about what library to use to do sentiment analysis.