Meeting Notes: Data Scraping Efforts

Date: January 20, 2025

Overview of Activities: We explored various websites and methods to scrape data from social media sites and blogs. Below is a detailed account of the tools, libraries, and challenges encountered:

---

1. Twitter Data Scraping

Tools Used:

- **TwCommentExport v1.1.4 (Chrome Extension):** Used to scrape Twitter comments efficiently.
- **Snscrape:** Python library for bulk scraping of Twitter posts.
- **Selenium:** Used to navigate Twitter's website and extract data directly.

Challenges:

- Twitter's restrictions on scraping large volumes of data.
- Anti-scraping measures that limit the extraction of comprehensive data.

Reference:

- Multilogin Blog on Twitter Scraping

2. Medium Data Scraping

Getting Started:

- Scraping data from Medium using Python and BeautifulSoup.
- Guide referenced: Scraping Medium with Python

3. Costco Customer Comments Research

Platforms:

- **Quora:**
    - Researched customer opinions on Costco products.
    - Common themes: product satisfaction, membership services, and shopping experience.
    - Link: Quora Search for Costco
- **Instagram:**
    - Tool: **Apify Instagram Scraper**

- Objective: Collect customer opinions and experiences shared in Instagram comments related to Costco products.
- Link: Apify Instagram Scraper

4. Reddit Data Scraping

Objective: Scrape posts and comments from the Costco_alcohol subreddit.

Approach:

- Used the praw library to interact with Reddit's API.
- Retrieved posts, comments, and metadata (e.g., scores, URLs, timestamps).
- Data saved in CSV format for further analysis.

Challenges: None significant; the process was successful for the top 1000 posts.

References:

- Reddit Scraping Tutorial
- RStudio Reddit Comments Scraping

5. Influenster Website Scraping

Objective: Scrape product reviews for Costco.

Approach:

- Used the cloudscraper library to bypass Cloudflare security.
- Parsed HTML with BeautifulSoup to extract reviews.

Challenges:

- Cloudflare's security measures prevented successful scraping.

Outcome: Successful.

6. Instagram Data Scraping

Objective: Scrape data about Costco's engagement in the US region, including posts, deals, and user comments.

Tools Tried:

- **Selenium, BeautifulSoup, Instaloader:** Limited to basic metrics (followers, total comments, hashtags).
- **Third-party APIs (e.g., Apify, Phantombuster):** Effective but costly and time-intensive.

Effective Solution:

- **Instagrapi Library:**
    - Successfully retrieved real-time data, including comments.
    - Authenticated with Instagram using credentials.
    - Fetched post details and comments.

Challenges:

- Privacy restrictions and Instagram's anti-scraping measures.
- Costly subscription fees for third-party services.

References:

- Instagrapi on GitHub
- Phantombuster Documentation

7. Facebook Data Scraping

Objective: Scrape Facebook comments for Costco-related posts.

Tools Tried:

- **Apify Scrapers:** Effective for single-post scraping.
- **Python (facebook-scraper library):**
    - Used cookies to set up sessions

Challenges:

- Errors in session setup.
- Found github source where data is fetched in separate CSV files, requiring post-processing.

References:

- Scraping Facebook Comments with Python
- Facebook Scraper Library

8. Llama 2 Model for Automated Scraping

Objective: Use Llama 2's language model for intelligent parsing of HTML content.

Approach:

- Loaded Llama 2 model via Hugging Face's Transformers library.
- Extracted structured data (e.g., product names, brands, prices, reviews).

Challenges:

- Insufficient GPU memory on Google Colab caused frequent OutOfMemoryError issues.
- Implementation pending testing on a high-end machine.

References:

- Hugging Face Discussion on Llama 2

Key Takeaways

- **Privacy Restrictions:** Scraping social media platforms is challenging due to privacy laws and anti-scraping measures.
- **Tool Selection:** While many tools exist, their effectiveness depends on the platform and specific requirements.
- **Effective Solutions:** Instagrapi proved to be a reliable tool for Instagram scraping.
- **Future Steps:**
    - Investigate alternative methods for bypassing Cloudflare.
    - Optimize Llama 2 implementation using advanced hardware.
    - Research robust APIs or third-party tools for seamless scraping.