Finesse Challenge

1. **Your reasoning behind why certain approaches would work best here, how you would modify them (or not) and how you would go about implementing them**

I used Selenium to scrape TikTok videos from the “explore” page with a query of “fashion”. The query is configurable, and if used for mass data gathering, multiple queries should be used in parallel. I used the default “Top” tab and scrolled the page until no more trending videos were available (~140 posts).

I opted to use Selenium as I’ve used it in the past and could get a solution running quickly. Also, a Chromedriver is included within Selenium which requires one less variable component within the Docker container. Selenium has library support, like selenium-stealth to rotate the user agent, which I used to avoid bot detection and Captchas blocking my queries.

This approach works well since I can scrape trending posts and their view count (views are not visible on a video page). Alternatively, I use the explore feed’s “Videos” tab, or compile accounts from the “Accounts” tab and scrape those accounts’ videos. The optimal solution is likely a combination of scraping TikTok’s various tabs/feeds.

Web scraping is often a brittle solution, meaning if TikTok ever changes their UI, the scraper will break. It also is only set up to work with the Explore feed. An alternative place to scrape is to have the scraper infinitely scroll the default TikTok page. However, this presents two issues:

- We have to add (several) queries to the author’s profile, find the matching video and add the viewcount to this scrape. (views are visible as an overlay element on videos displayed on a profile)
- We have to sign into an account that’s been “primed” with fashion preferences. This means that the account has been used in a way that suggests to TikTok that it is interested in fashion.
    - Subproblems: is it primed correctly? Primed to the type of fashion we want? Account demographic matches what we are aiming at?

TikTok’s own API’s only allow researching of videos for non-profit organizations. There are other freemium API’s that allow querying TikTok’s explore feed, sometimes in magnitudes of 30 videos at a time. From my research, it seems that these solutions use similar web-scraping techniques, considering there is no official TikTok solution for commercial use. These solutions are expensive:

- TikAPI - $190 monthly for 10,000 requests/day
- EnsembleData - $1400/month for 50,000 requests/day

TikTok for business (lead generation specifically) is an avenue worth pursuing, but was not applicable for this project.

2. **The scalability of your system (data throughput, deployment, challenges/costs at scale etc.)**

The script scrapes nearly 140 posts in about 7 minutes (~1200 posts/hour). It is easily deployed using the Docker container. Parallelization can be used to run multiple scrapers at once. At scale, we can manually parallelize instances of the scraper, or use a solution with Selenium Grid to speed up setup.

A huge slowdown in my scraping process was querying comments. I currently have the scraper limited to scraping min(20, numComments) from the top to gather the most relevant surface-level comments. This process could be improved by removing the need to call `findElements()` for all visible comments on every scoll.

There are nearly 34 million videos posted on TikTok per day. Scraping 340,000 (1%) of these posts would require 283 hours with this scraper running linearly. With 10 instances, we can scale to scrape these posts in 28 hours. The server costs for storage would be minimal (plaintext), but there could be rising costs for memory to scroll and view videos.

3. **Describe how you assess the quality of the data and suggest ways the quality metrics or the data itself could be improved.**

As described in (1), the quality of the videos can be improved by querying from different places than just the explore tab. However, the trending videos tab (where I currently am scraping from) is likely to contain good videos for fashion trends. We can sort our data by likes, views, ratio of likes-views, and sentiment analysis on the bots.

We can potentially discard lower-quality posts that don't meet a certain number of views, likes, shares, or a ratio of these metrics.

We can

4. Any possible extensions/improvements you would implement if you had no time constraints.

    - Querying using different words than fashion (call scraper again with other words)
    - Run scraper in parallel with different queries
    - Scrape infinitely from the "For you page" on one or multiple "primed" TikTok accounts
    - Scrape more/less comments from each post (maybe a % of total comments rather than X comments)
    - Sort by like/view count, ratio of likes/views/shares, and scrape posts that are at least of X threshold
    - Scrape popular accounts' videos (accounts found through explore with queries like "fashion")
    - Add more parameters for customizing script
        - how many comments to scrape per post
        - which tabs of the explore feed to scrape
        - thresholds for minimum likes/views/shares/comments
    - Code quality
        - Modularize code further (separate logic for explore page vs. post)
        - Better documentation for what each CSS-class variable represents with screenshots
            - Some classes may not have been turned to labeled constants

Additional Written Questions

If you could improve one thing about FINESSE right now - as you experience it – what would it be and why?

I wish that I had the option to change my vote for the clothing item I clicked on. Also, once I click to vote, the numbers show, but after refreshing, I can vote again. This makes it unclear to me whether my vote went through and this second vote can count for duplicate votes.

If I understand correctly, the votes are used to determine how much of each item should be produced. So, if I am equally interested in multiple options, I'd like to vote for all three.

Instead, I wish that I could cast a vote on all three like a checkbox selection vote. Once I submit it,, I could change it if I decide I was also interested in another piece. What I voted on should reappear. Maybe this is not feasible due to browser memory not being guaranteed to persist without a user account (I do not have an account).

Your plane crashes and you land on an island. What do you do? How do you plan your survival?

My first step would be to ensure that everybody is safe and accounted for. I'd also try to set up one of the mobile devices to be used for SOS signalling if possible in the area that we land in.

Then, I would survey the area to determine sources of wood, water, shelter, and food (fish, fruit, animals). Assuming nothing from the cargo is available, I would prioritize finding a way to contain water, fire, food, and then shelter. Humans can go a long time without eating, but a very short time without water. Fire is essential to boil the water to make it drinkable and cook any food. Shelter is also important, but isn't a bare necessity to survive.

What matters to you most and why?

The most important thing to me is sustainable, long-term well-being, for myself and others. My priority when I am solving a problem, helping out a friend, or planning for my own life, is to find a solution that ensures the most amount of benefit for the longest period of time. This principle is why I choose to go to the gym, or why I try to teach others the root cause of their problems if I am helping them out. If I am building something, then I plan it in such a way that it would be easy to maintain, expand on, and serve the best functionality for its intended lifetime.

Short term actions often seem helpful or easy to turn to, but long-term solutions have always brought me the most happiness. I feel myself making the most impact on others when I am able to teach them something they use in the future, or if I am able to change their way of thinking that would benefit them in the future.