**Technical Documentation**

**Project Overview:**

Built a web scraper using Node.js with Puppeteer to extract influencer profiles from Collabstr. The scraper collects structured data—name, role type, and profile URL—while handling pagination, filtering invalid entries, and saving results as CSV files. It supports two role types (UGC Creator and Video Creator) with separate CSVs stored in a dedicated data/ folder.

**Tools & Approach:**

• Puppeteer: Controls a browser to navigate pages and extract DOM elements.

• Detects pagination and iterates through pages until the requested number of profiles is collected.

• Profiles are filtered using a blacklist of brand-like names.

• Dynamically calculates how many pages to visit based on the requested profile count.

• json2csv is used to generate CSV files.

• Optional login via .env credentials enables access to additional pages when profile counts exceed 50.

**Challenges & Solutions:**

• Dynamic content & headless mode: Collabstr delays content rendering or detects automation. Solved with explicit waits, longer waitForSelector timeouts, and user-agent spoofing.

• Login handling: Automatically triggers login only when the number of profiles exceeds 50 to efficiently access additional pages.

• Data validation: Checking for broken links and incomplete profiles was challenging. Implemented strict filtering to remove placeholders, brand-like names, and inaccessible links.

**Data Validation Steps:**

• Excluded profiles with brand-like names or placeholder entries.

• Verified profile URLs for completeness and accessibility.

• Limited the dataset to the user-defined maximum number of profiles.

• CSV outputs contain only validated fields (name, role, profileLink) to prevent malformed records.

**Estimated Time Spent:**

• Puppeteer setup: ~3 hours

• Pagination, filtering, login module, and link validation: ~5 hours

• CSV export and testing: ~2 hours

Total: ~10 hours

**Scalability:**

• Efficient for 1,000+ profiles by dynamically handling pagination and storing results in a Map to prevent duplicates.

• Login mechanism ensures additional pages are accessible when scraping large datasets.

• CSV outputs are lightweight and portable, suitable for integration with databases or analytics pipelines.

• Future improvements: parallel page scraping and batching for faster performance.