4502 Group 2 (in-progress draft)

Timeseries + NLP

What can we learn about someone from their writing over time?

Datasets

- Couldn’t find good datasets for longitudinal text analysis.
- Found a blog with 17 years of posts.
- “Scraping” tutorials show overly simple examples with known urls to scrape.

Solution? Custom Scraper.

- ~12 hrs crawling time for one headless browser and 2-3 tabs (# of browser+tabs processor limited)
- 38,723 posts files (212mb). Text only. No pictures, videos, etc.
- Metadata JSON file containing year, month, day, title, author, url (10mb)

How best to store posts and metadata? TBD. Different analyses require different formats.

- Remote: 54mb zip of posts directory and metadata file on github
- ^^ Github complained. And was slooooooow. May need to store separately for commit performance.
- Locally: Metadata file + posts directory. Likely intermediate formats needed to support analyses.

Evaluation

Quantified evaluation is tricky since the desired outcome is exploratory in nature. We are also new to NLP, so there are many unknown unknowns. Overall, we want to explore what we can learn about someone from their writing, from momentary thought patterns, to how such patterns change over time.

Good results include:

- Adapt book concepts (esp chpt 7+8) to the NLP context where possible/practical
- Determine what questions we can ask through related research papers (e.g., personality characteristics, optimism/pessimism, and positive/negative sentiment)
- Learn enough NLP technology to answer as many questions as we can
- Communicate answers intuitively in writing and images