Timeseries + NLP

What can we learn about a person from years of writing?

Questions

What can we learn about a person from their writing?

What psychological characterics stand out?

What linguistic characterics stand out?

What psychological patterns are interesting?

What linguistic patterns are interesting?

Datasets

Problem: No datasets available for longitudinal text analysis

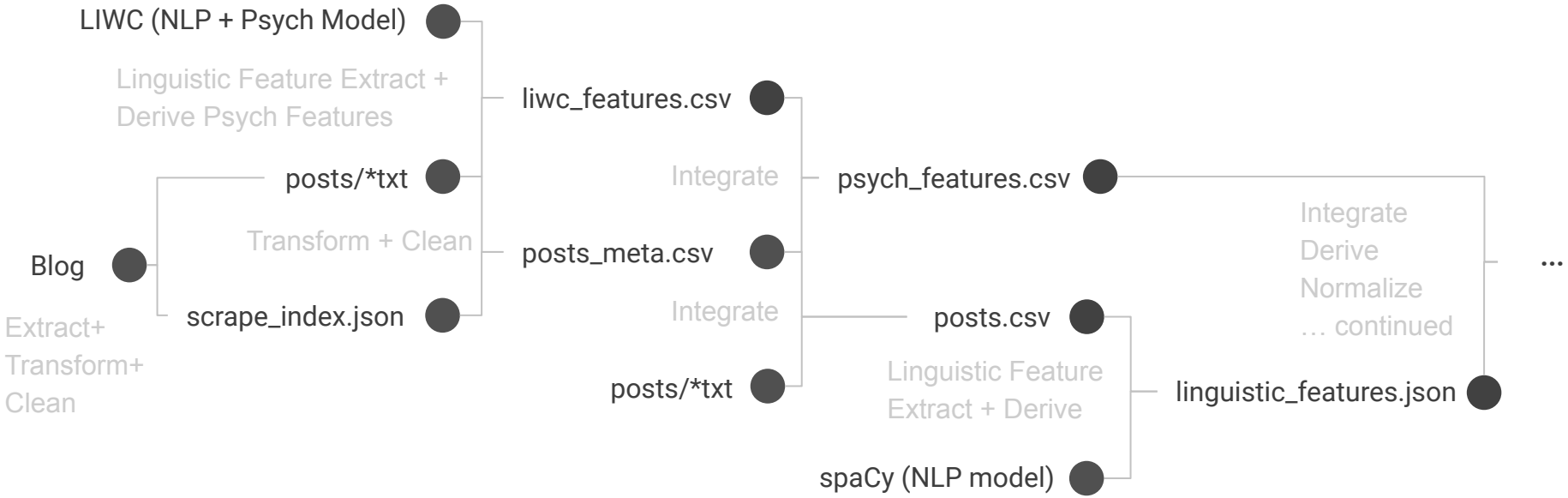
Solution: Collect our own.

- 1 blog
- ~39k posts spanning 17 years
- ~212mb of content in posts/*txt files
- ~10mb of metadata in scrape_index.json
- ~40 hours learning + writing + debugging a custom crawler/scraper
- ~12 hrs crawling time

Data Prep Pipeline

Clean
Transform
Extract
Derive
Integrate
Normalize

… in varying orders



Tools

Coord: Github, Trello, Slack, Zoom

Writing: Google Docs, Slides, VPN

Scraping: Puppeteer (Chromium), CSS, Chrome Dev Tools, Node.js, custom JS

Cleaning / Integration: Funcy, Pandas, Custom Python

Normalizing: Pandas + sklearn

Feature Extraction: spaCy (NLP model) and LIWC2015 (Psych + NLP Model)

Classification: LIWC (Linguistic Inquiry and Word Count)

Dimension reduction: LIWC, pandas, sklearn

Vis: Pandas, Seaborn, Matplotlib

Techniques: Adam

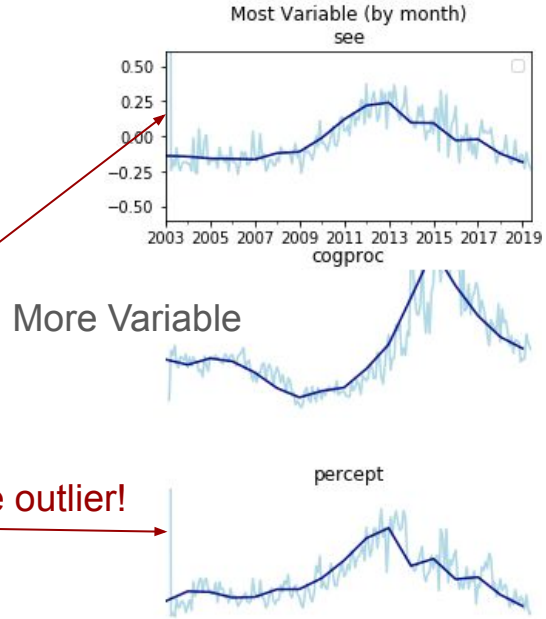
All those in the data pipeline. Plus:

- Dimensionality Reduction: Manual removal
- Dimensionality Reduction: Subsets by LIWC class, year, month
- Visualization
- Clustering: Automated - Sklearn OPTICS
- Correlation Analysis
- Correlation Change Analysis

Results: Adam

Correlation Analysis (Meh. Plotting dimensions is more useful.)
Correlation Change Analysis (Ditto)

More cleaning needed. Especially contextual outlier handling for scenarios like few-word-posts in few-post-months.



All March Posts

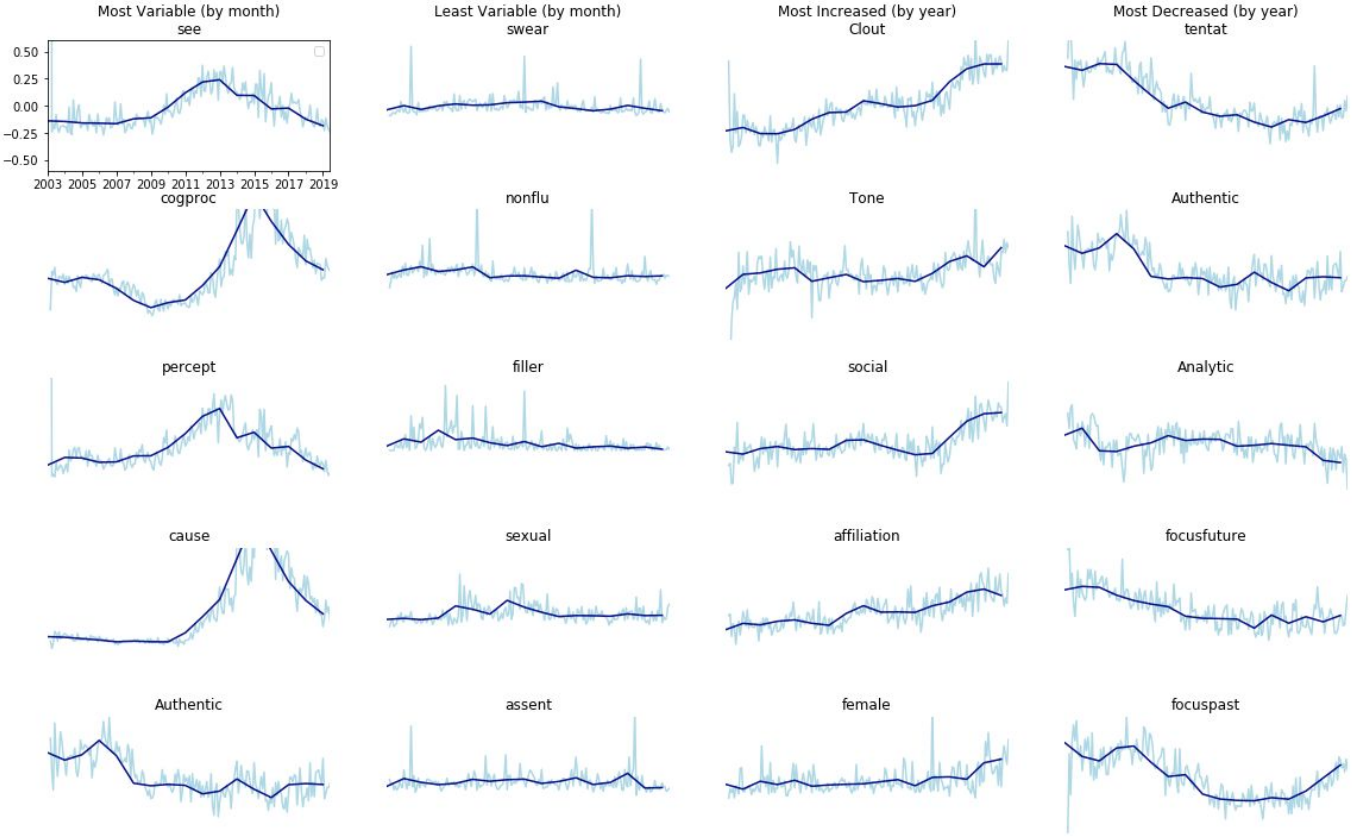
date	title	see (Sensory Category)	WC (Word Count)
2003-03-01 01:00:00	Google Search	60	5
2003-03-19 01:00:00	Not-So-Hidden Agenda	0.3	338

Results: Adam

Infinite ways to slice and visualize.

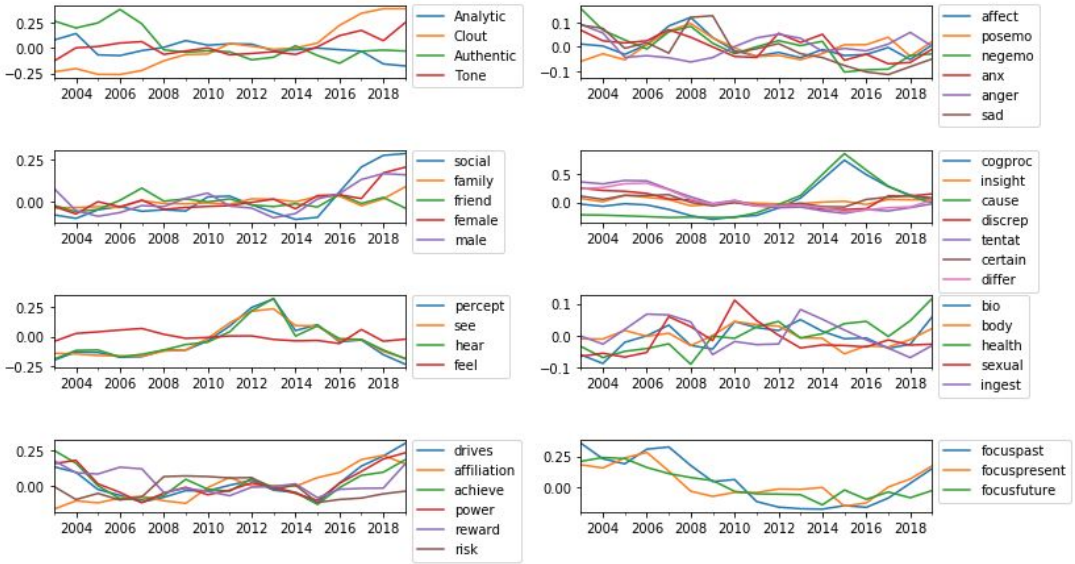
Simple FTW. Manual clustering + vis by year + month.

Measurement ranges matter. Month increase/decrease == Year variability.

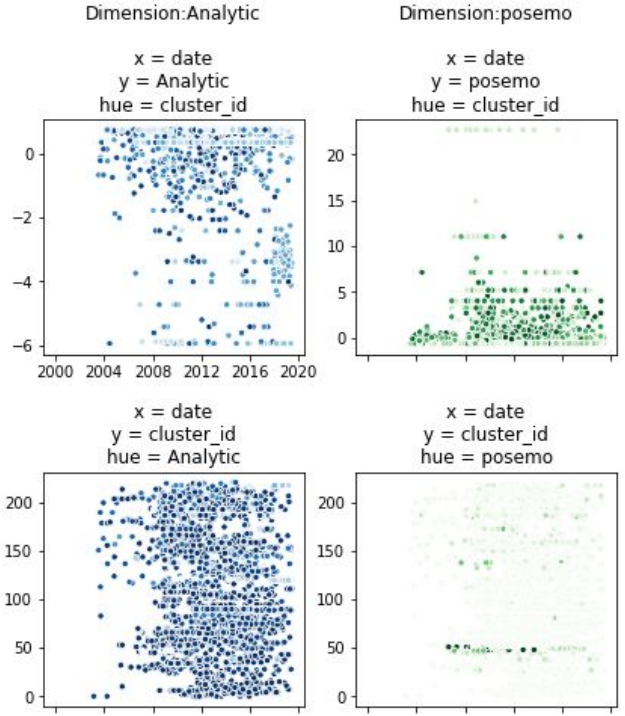


Results: Adam

Manual dimensionality reduction
(by LIWC group, grouped by year)



Clustering: Automated - Sklearn OPTICS
- 210 clusters
- 90% of data points non-clustered
- Unclear clustering results
- Unclear vis from many clusters + dims



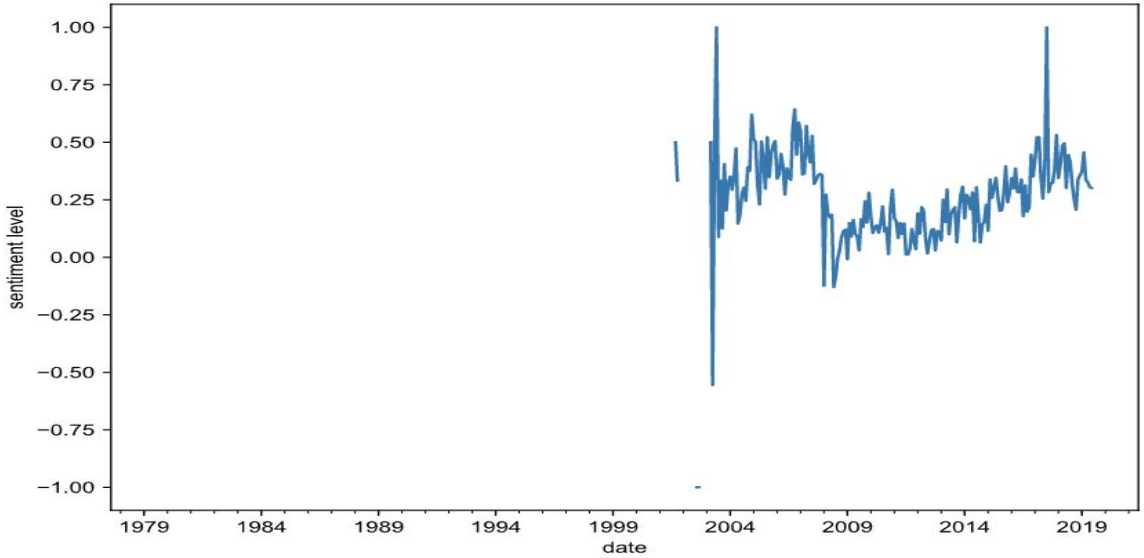
Techniques: Henry

Read the data through python pandas, archive the article data of the author for about 40 years, and calculate the emotional value for each article, using the library VADER(VADER (Valence Aware Dictionary and sEntiment Reasoner) is a lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media, and works well on texts from other domains.), after getting the emotion of each article. By re-sampling the entire data in the monthly dimension, it was found that the author's article was positive for most of the time and the author's frequency of publication was very high, so he was inferred to be a positive and diligent person.

Results: Henry

The data shows the author's article was positive for most of the time. So we can think of him as an positive person. At the same time, statistics are made on the author's writing style. Almost all of the author's articles use a lot of scientific data to support their views. Using data to prove the correctness of one's point of view has almost become the author's main rhetorical and writing characteristics. In addition, the author has added pictures or videos to almost all of his articles to help readers understand the views they want to make. Combining articles and pictures to express one's point of view is also a great writing feature of the author.

```
ut[30]: Text(0,0.5,'sentiment level')
```



Techniques: Kyle

For my analysis I decided to look at if the author of these blog posts is overall a positive or negative person based on what he writes about. LIWC provides us with two statistics pertaining to this: negative and positive emotion words. So, I read in the data using a pandas dataframe and took the average of many different categories that LIWC provided. However, the easiest to clearly see was the positive and negative emotion words.

Results: Kyle

After taking the averages of many columns in the dataframe, I ended up analyzing the positive and negative emotion words. His negative emotion average was 1.48 and his positive emotion average was 2.67, almost double the negative emotion average. This leads me to believe the author is a more positive person overall.

Applications

We can use our results to make assumptions about the author of these blog posts. Using VADAR and analyzing the positive and negative words we can conclude that he is mostly a positive person. There are countless other applications using Timeseries and NLP, such as analyzing the author’s psychological and linguistic characteristics.

Thanks for Watching!

https://github.com/a-laughlin/data-mining-group

Adam Laughlin
Computer Science
University of Colorado Boulder
adam.laughlin@colorado.edu

Jiaheng Zhao
Computer Science
University of Colorado Boulder
jizh3194@colorado.com

Kyle Bremont
Computer Science
University of Colorado Boulder
kyle.bremont@colorado.edu