4502 Group 2 (in-progress draft)

Timeseries + NLP

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

Description

We intend to find out what we can learn about a person from their writing. So, by scraping this individuals blog posts from the past 17 years we should have enough data to analyze using NLP. We can conclude specific things about this person by analyzing his writing trends such as if he is a optimist or pessimist.

Prior Work

So far we have just found several resources on NLP, time series and web scraping. We have also created a github, slack group and a trello board in order to communicate and stay on track with the project throughout the summer.

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 slooooow. May need to store separately for commit performance.
- Locally: Metadata file + posts directory. Likely intermediate formats needed to support analyses.

Proposed Work

1. Data Quality Research: Good quality data has three attributes: uniqueness, continuity, and consistency. We can test the bias of the data from these three levels.

(a) Verify that the data is duplicated.

(b) There are no missing values between the lowest and highest values of the check attribute, and some values need to be unique. In general, I will pay special attention to these points when testing continuity:

Use of NA value
Use of Null values
Use of null values
Use of special symbols (such as ?, %)
Make use of null values (such as "Unknown", "empty", "unknown")

(c) Check whether the data write format conforms to the unified specification, and whether there will be behavior such as field overload. In the field, there are mainly three types:

Character type (eg CHAR, VARCHAR)
Numerical type (eg INTEGER, DECIMAL)
Date type (eg DATE, TIMESTAMP)

Proposed Work

2. If we find that there is a missing value, we will automatically add the value to those missing values or delete it directly.
3. Smoothing-out noisy data:
Binning:
Smooth ordered data by values around the data (near values). First, sort the data, and then divide the data into equal-frequency boxes. Replace the values in the box with the same indicator values. Therefore, in the binning operation, the larger the width of the box, the more obvious the smoothing effect. The box can be of equal width or can be defined by itself.

Here, through the different indicators, the bin can be divided into:

The box average is smooth: each value in the box is replaced with the average of the data in each box.
The median box is smooth: each value in the box is replaced with the median value of the data in each bin.
The box boundary is smooth: the maximum and minimum values in the defined box are the boundary values, and each value in the box is replaced with the nearest one.

Regression:
Fit a noisy variable to a straight line (linear regression) or a curve (curve regression). The general approach here is to use a linear regression or multiple linear regression to fit a line or multidimensional plane. In turn, the noise data can be replaced by smooth lines or faces.

Outlier Analysis:
Outliers are detected by clustering methods. The values that fall outside the cluster are outliers that can be replaced or directly removed by the cluster's centralized trend indicator.

Tools

Data Scrubbing Tool: Potter's Wheel

Data Auditing Tool: ACL

Data Migration Tool: ETL

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