TRENDS IN JOURNALISM STYLOMETRY

Project Proposal

Prepared by: Rebekkah Ismakov

January 11, 2017

BRIEF SUMMARY

Motivation

The way we consume news and media reveals a lot of the society we live in, as well as of the changing style of communication and writing throughout location and time. Journalism, unlike other forms of writing, is relied on being objective and presenting pure facts without infusion of biases and opinions. In reality however, we are intuitively aware that this not the case. Journalism naturally encompasses the authors assumptions and biases within the text, visible not only through the words chosen, but also through the very structure of the author's style of writing.

This allows us to look at many features, as well as many potential predictors to attempt to extrapolate what journalism and news media expose about both the writers and the reader consumers. This is useful both as an analytical tool as well as an extrapolation of predictors that will allow businesses to better understand their changing and versatile audience.

In addition to examining stylometry between media sources and across time, it would be of interest to see if writing style could predict the gender of the author. This will likely have useful applications in many cases where gender is not disclosed; if a company wants to target a specific sex, this would provide a method of deciphering gender from purely text. Additionally, a topic of great concern within industry and tech, among other domains, is the wide gender gap present. It is claimed that this leads to implementations of products, designs, and policies being fully decided by one gender only, thus potentially hindering progress by ignoring different modes of thinking. If a clear distinction can be seen from purely analyzing political and world text, this would provide strong evidence to this claim that the genders do in fact have varying ways of observing the world, to the extent where it can even be picked out from words that should ideally be objective in nature.

To investigate the stylometry of various news sites and the properties of their articles, we will need to parse out various indicators of style within the text. This type of feature analysis will be implemented by looking at text properties such as mean length of words, frequency of given words, and variance of sentence length, among other parameters (see Methods below for more details). Data will be obtained through web scraping, as well as through available APIs.

QUESTIONS AND GOALS

Analysis Questions

- Are there any significant differences in stylometry between different news media sites?
- Has the style of writing changed throughout time?
- Are there differences in stylometry of the most popular articles depending on the medium through which it is shared (most viewed, most shared, most emailed)?
- Is there a gender imbalance of journalists? Does this discrepancy extend to different news sites?

Modeling Goals

- Predicting the the media source of an article given just its text [Naive Bayes].
- Create an app that allows a user to copy in text and outputs which new site writing style it is most similar to.
- Predicting author gender given an article [Naive Bayes, Logistic regression, Random Forest].

Project Outline

The project will be broken up into the following steps:

1. Scrape all the news sites to be used in analysis (in process).
2. Examine differences in features between news sites and between author genders.
3. Use Naive Bayes to predict new site from article.
4. Use Naive Bayes, Logistic Regression, Random Forest to try to predict gender of author from article.
5. Use Time Series to look at writing style throughout time (use NYT data going back to 1980s).
6. Perform statistical analysis on NYT most popular articles to extrapolate differences between different sharing mediums.
7. Create an app that will tell you which writing style your text is most similar to.
8. Create a predictor that guesses your gender.
9. Showcase results and models in aesthetically pleasing way.

DATA AND METHODS

News Site Datasets

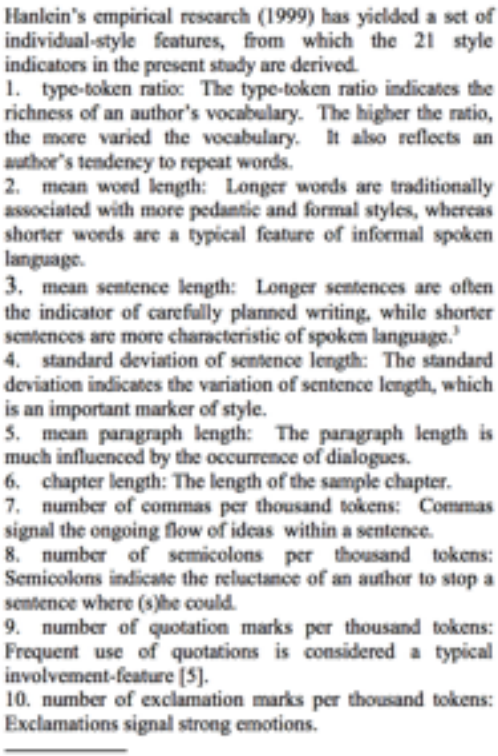
- Buzzfeed: 11,308 datapoints scraped.
- TIME magazine: 12,693 datapoints scraped.
- Slate: 4059 datapoints scraped.
- Reuters: 303 url links to scrape.
- Atlantic: 72,485 url links to scrape.
- APIs: NYT, BBC, Associated Press
- https://en.wikipedia.org/wiki/Lis_of_news_media_APIs

Extrapolation of Style Features

Certain properties of text are able to reveal information on the style of writing. For example, the mean lengths of words and mean lengths of sentences may say something of the formality of writing, as well as the frequency of given words and punctuation marks.

The features used for extrapolating style from an article include:

- type-token ratio: The type-token ration indicates the richness of an author's vocabulary.
- mean word length: Longer words are traditionally associated with more pedantic and formal styles.
- mean sentence length: Longer sentences are often the indicator of carefully planned writing, while shorter sentences are more characteristic of spoken language.
- variability of sentence length: The standard deviation indicates the variation of sentence length, an important marker of style.
- frequency of commas: Commas signify the ongoing flow of ideas within a sentence.
- frequency of semi-commas: Indication of more formal writing.



Hanlein's empirical research (1999) has yielded a set of individual-style features, from which the 21 style indicators in the present study are derived.

1. type-token ratio: The type-token ratio indicates the richness of an author's vocabulary. The higher the ratio, the more varied the vocabulary. It also reflects an author's tendency to repeat words.
2. mean word length: Longer words are traditionally associated with more pedantic and formal styles, whereas shorter words are a typical feature of informal spoken language.
3. mean sentence length: Longer sentences are often the indicator of carefully planned writing, while shorter sentences are more characteristic of spoken language.[3]
4. standard deviation of sentence length: The standard deviation indicates the variation of sentence length, which is an important marker of style.
5. mean paragraph length: The paragraph length is much influenced by the occurrence of dialogues.
6. chapter length: The length of the sample chapter.
7. number of commas per thousand tokens: Commas signal the ongoing flow of ideas within a sentence.
8. number of semicolons per thousand tokens: Semicolons indicate the reluctance of an author to stop a sentence where (s)he could.
9. number of quotation marks per thousand tokens: Frequent use of quotations is considered a typical involvement-feature [5].
10. number of exclamation marks per thousand tokens: Exclamations signal strong emotions.

Fig 1. List of features used as indicators of style. Taken from *Using Machine Learning Techniques for Stylometry* (Ramyaa and Khaled, 2004).

Fig 2. Richness of vocabulary indicated by type token ratio plotted against time. News site media plotted by color. Slate appears to have a lower type token ratio compared to TIME and Buzzfeed. Buzzfeed vocabulary richness appears to be decreasing. Preliminary results.