Authorship & Stylometry: A classic DH topic

& text classification, more generally

J.Nerbonne, reusing slides (18-26) from Dan Jurafsky

Style

- Study of STYLE, characteristic writing habits
  - One aspect of studying literature
    - Styles of Henry James vs. Mark Twain, Thomas Mann vs. Hans Falada
    - Flowery, factual, humorous, monumental, suspenseful, …
- STYLOMETRY measures style to improve objectivity, replicability
  - See interview with John Burrows
- Study of individual style often generalized to study of groups
  - Latin: Classical/Medieval/Neo-Classical; Romantic/Realistic/Surrealist
  - Groups of same age, sex/gender, social class, … (in social media)

Attributing authorship

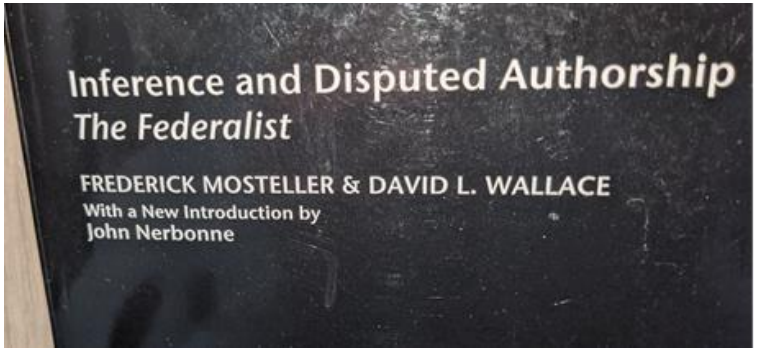
- Attributing authorship based on style is a difficult test
  - Need to characterize individual differences
- Literary scholars, historians of course interested in authors
- Traditional scholars skeptical of inferring authorship using style
  - Prefer evidence on the provenance of manuscripts, witnesses
  - Rudman (2002), cited in JN “The exact analysis of texts”
- Interesting related forensic field, DETECTING FORGERY
  - In forgery, authors (may) intentionally mimic others’ styles
  - Love (2002), also cited in JN “Exact analysis of text”

Attributing authorship

- Attributing authorship based on style is a difficult test
- Literary scholars, historians of course interested in authors
- These are mostly known, but not always
  - Lots of conjecture on authors of the works we call Shakespeare's
  - Junius papers (1769-72), critical of George III
  - Federalist papers (10/1787 – 8/1788) on the American constitution
  - Beatles' songs (Paul McCartney or John Lennon)
  - J.K.Rowling (aka Robert Galbraith) The Cuckoo's Calling
    - Rowling detected as author by stylometrist Juola

Mosteller & Wallace (1964)

- Bayesian statisticians, University of Chicago, Harvard
- Laid the groundwork for a lot of text analysis
- Stylometry, but also SPAM detection



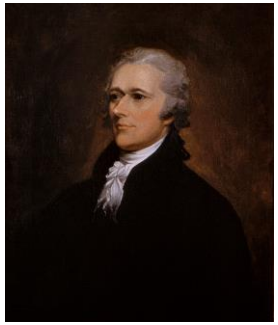
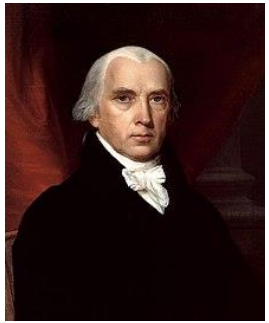
- 3rd ed. with introduction by course instructor

Choice of subject matter

- *Federalist Papers* were written by three authors during less than one year
  - Big advantage: it's easier to distinguish three authors than more (e.g. 30)
- All the papers were published in New York newspapers and all signed by *Publius*
  - Advantage: consistent genre (newspaper) and time (<12 months)
- The essays are important in the history of politics and law
  - The first foundation of constitutional democracy!

Federalist papers

- Essays on the American constitution then under debate (1787)
- Famous, influential authors
  - Alexander Hamilton, founder of US National Bank
    - Current hero of Broadway musical
  - James Madison, author of *The Bill of Rights,* 4th president
    - *… Madison Square Garden*
  - John Jay, first chief justice


- Hamilton, Madison & Jay
  - Hamilton – strong nat'l government
  - Madison – "Jeffersonian" democrat, protect people from the state
  - Jay – rule of law
- Worth knowing!

Federalist papers still quoted today

- Alexander Hamilton 1790: “The only path to a subversion of the republican system of the country is, by flattering the prejudices of the people, and exciting their jealousies and apprehensions, to throw affairs into confusion, and bring on civil commotion, […]

  When a man unprincipled in private life, desperate in his fortune, bold in his temper…is seen to mount the hobby horse of popularity […] he may ‘ride the storm and direct the whirlwind.’”

Foto von Jon Tyson auf Unsplash

Choice of indicators

- What's makes a style distinctive?
  - Lots of people think of sentence length, complexity, vocabulary domains (sailing, herding), …
  - Sentence length in *Federalist Papers*
  - Complexity hard to measure
  - Subject domains the same!

	mean	sd
Hamilton	34.55	19.2
Madison	34.59	20.3

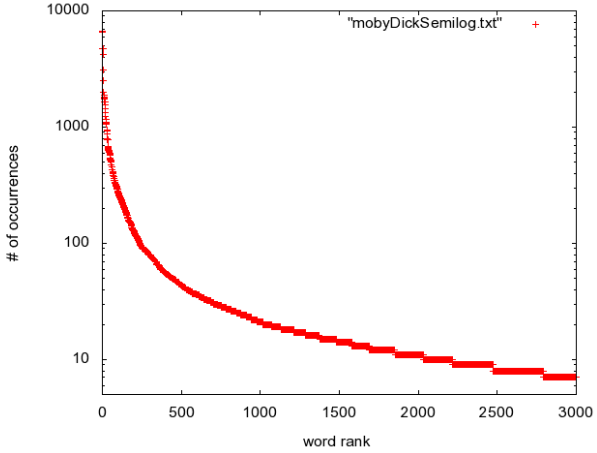
- Mosteller & Wallace look at FUNCTION WORDS
  - Articles, auxiliary verbs, prepositions, conjunctions, pronouns

Why function words?

- Zipfian motivation
  - Function words are frequent, most others are not
  - Sort words by frequency, then the probability of nth-most frequent word

$$p(n) \sim \frac{1}{n}$$

- *the* accounts for 5% tokens (1st freq.)
- How frequent is the 100th (*men*), 1000th (*names*)? (from Project Gutenberg)



- Word freq. in *Moby Dick*
- Note *y*-axis is logarithmic

Why frequent words?

- If the 10.000th MOST FREQUENT WORD (MFW), *calves*, occurs about 10^{-4} as often as *the* (5×10^{-2}), how big does your sample need to be in order to see significant differences?
- Expected frequency of *calves* ~ 5×10^{-6}, so to see 30, you expect to need x, where

$$30 = x \times 5 \times 10^{-6} \; ; \; x = {}^{30}/_{5} \times 10^{6} \; ; \; x = 6 \times 10^{6}$$

- Mathematical problem: I used a rule of thumb from frequentist statistics, i.e. use sample size 30. Is that OK here?

Why frequent words?

- 6 million words ~ 60 novels ~ 4 months of newspaper (*Frankfurter Rundschau*)
- Using MFW you don't need huge samples
- Using MFW you have a better chance of seeing significant differences
- Better chance of seeing the sample words in all the documents

- Any other reasons?

Frequent words are used automatically!

There's a set of exercises on the website. Please turn it in by Mon.

There's a set of exercises on the website. Please turn them in by Mon.

- Lots of people don't notice the difference, including people that read the sentences aloud!
  - We tend to remember *what* was said, not *how* (Bransford & Franks 1972).
  - Most notice unusual words and unusual uses of words
- Similarly, writers don't focus on how MFW are used
- So we can regard them as unconscious traces of style

Document Similarity – Burrow's Delta Δ_B

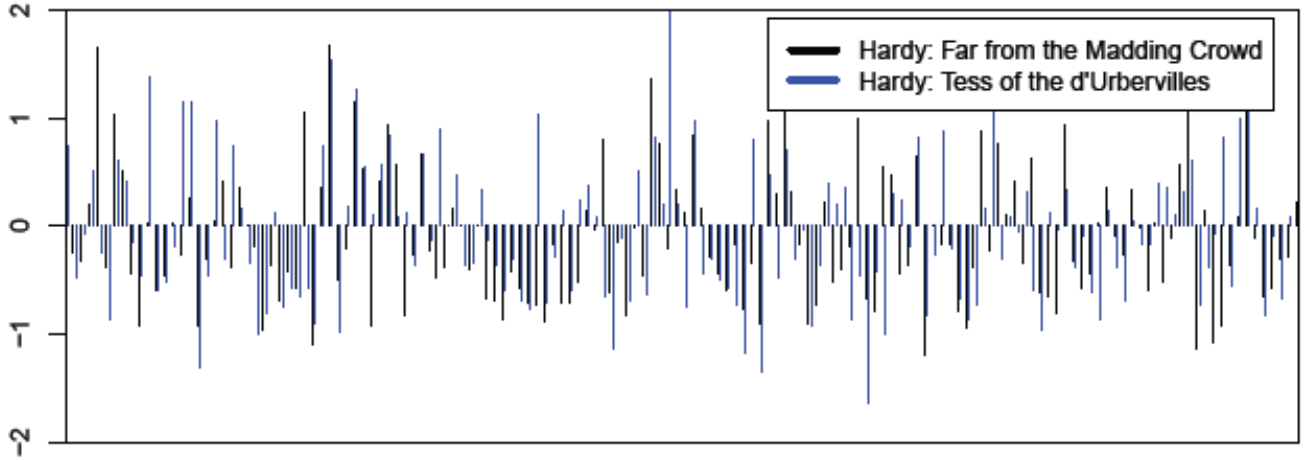
- Based on MFWs (see above)
- Raw frequencies bias measures toward very FW, since nth MFW has frequency $1/n \times f_{MFW_1}$, where MFW_1 is most freq. word
- So normalize relative frequency of MFW w_i in doc D to z-value

$$z_{w_i}(D) = \frac{f_{w_i}(D) - \mu_{w_i}}{\sigma_{w_i}}$$

- … where $f_{w_i}(D)$ is the normalized rel. frequency of w_i in D, μ_{w_i} mean rel. freq. of w_i across documents, σ_{w_i} standard deviation

Example (Jannidis et al.) transforming to z_w

	the	*and*	*to*	*of*	*a*	*I*	*in*	*was*	*that*	*he*	*her*	
z(Madding Crowd) = (	.53,	-.23,	-.32,	.20,	1.66,	-.37,	1.04,	.52,	-.44,	-.92,	.03,	...)
z(Tess of the d'U.) = (	.75,	-.48,	-.08,	.51,	-.24,	-.87,	.60,	.41,	-.14,	-.47,	1.39,	...)
z(Oliver Twist) = (	1.05,	.15,	-.71,	-.56,	.37,	-1.01,	-.06,	-.74,	-.28,	.48,	-.94,	...)

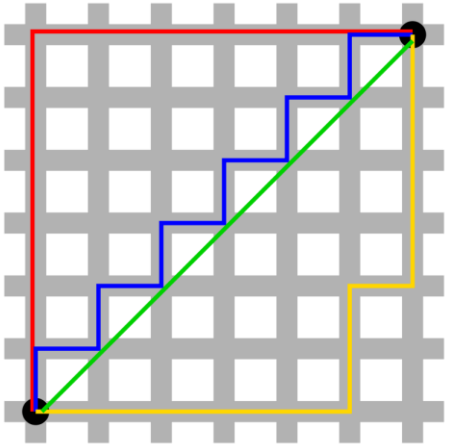


Aggregating over MFWs à la Burrows

$$\Delta_B(D_1, D_2) = \sum_1^{|MFW|} |z_i(D_1) - z_i(D_2)|,$$

where $z_i(D)$ is the z-value of the i-th word in D

- aka Manhattan, city-block distance
- aka L^1norm $\|z(D_1) - z(D_2)\|_1$



Wikipedia Manhattan distance

Lots of other potential traces of authors

- Number of MFW: 100, 500, 1.000?
- Bigrams w. one frequent word: *different than/different from,*
- Letter n-grams, esp. 4-grams
- Word lengths
- …
- Chance to try these in *Stylo*!

More Mosteller & Wallace groundwork

- Analysis technique – Bayesian
- Bayesian inversion – later lecture
- But for now, note that authorship attribution is like many other CL problems in TEXT CLASSIFICATION
  - SPAM DETECTION (Serious or not)
  - INFORMATION RETRIEVAL (Relevant or not)

Is this spam?

Subject: **Important notice!**
From: Stanford University <newsforum@stanford.edu>
Date: October 28, 2011 12:34:16 PM PDT
To: undisclosed-recipients:;

Greats News!

You can now access the latest news by using the link below to login to Stanford University News Forum.

http://www.123contactform.com/contact-form-StanfordNew1-236335.html

Click on the above link to login for more information about this new exciting forum. You can also copy the above link to your browser bar and login for more information
about the new services.

Male or female author?

1. By 1925 present-day Vietnam was divided into three parts under French colonial rule. The southern region embracing Saigon and the Mekong delta was the colony of Cochin-China; the central area with its imperial capital at Hue was the protectorate of Annam…
2. Clara never failed to be astonished by the extraordinary felicity of her own name. She found it hard to trust herself to the mercy of fate, which had managed over the years to convert her greatest shame into one of her greatest assets…

S. Argamon, M. Koppel, J. Fine, A. R. Shimoni, 2003. “Gender, Genre, and Writing Style in Formal Written Texts,” Text, volume 23, number 3, pp. 321–346

Positive or negative movie review?

- unbelievably disappointing

- Full of zany characters and richly applied satire, and some great plot twists

- this is the greatest screwball comedy ever filmed

- It was pathetic. The worst part about it was the boxing scenes.

What is the subject of this article?

MEDLINE Article


Brain and Cognition

Syntactic frame and verb bias in aphasia: Plausibility judgments of undergoer-subject sentences

?

MeSH Subject Category Hierarchy

- Antagonists and Inhibitors
- Blood Supply
- Chemistry
- Drug Therapy
- Embryology
- Epidemiology
- …

Text Classification

- Assigning subject categories, topics, or genres
- Spam detection
- Authorship identification
- Age/gender identification
- Language Identification
- Sentiment analysis
- …

Text Classification: definition

- *Input*:
  - a document d
  - a fixed set of classes $C = \{c_1, c_2, \ldots, c_J\}$
- *Output*: a predicted class $c \in C$

Classification Methods: Hand-coded rules

- Rules based on combinations of words or other features
  - spam: black-list-address OR (“dollars” AND “have been selected”)
- Accuracy can be high
  - If rules carefully refined by expert
- But building and maintaining these rules is expensive

Classification Methods: Supervised Machine Learning

- Machine learning is SUPERVISED when it's trained on data with correct answers
- *Input:*
  - a document *d,* a fixed set of classes $C = \{c_1, c_2, \ldots, c_J\}$
  - A training set of *m* hand-labeled documents $(d_1,c_1),\ldots,(d_m,c_m)$
- *Output: function mapping docs to classes* $\gamma(d) = c_i$
- *Lots of ML techniques, including clustering, logistic regression, support vector machines, and Naïve Bayes*

Next times

- Stylo, including exercise on *The Federalist*
- Bayes, Naïve Bayes
- Reports on *Federalist* Exercise

- *Fachschaft* reminds you of all-purpose tutorials (open space w. tutors)
- Please provide feedback, using QR or URL:
  - https://docs.google.com/forms/d/1d1UmAR5oNeNV7WcQAncmYQy0KdCIGRJiP7hZ53H3V4/edit