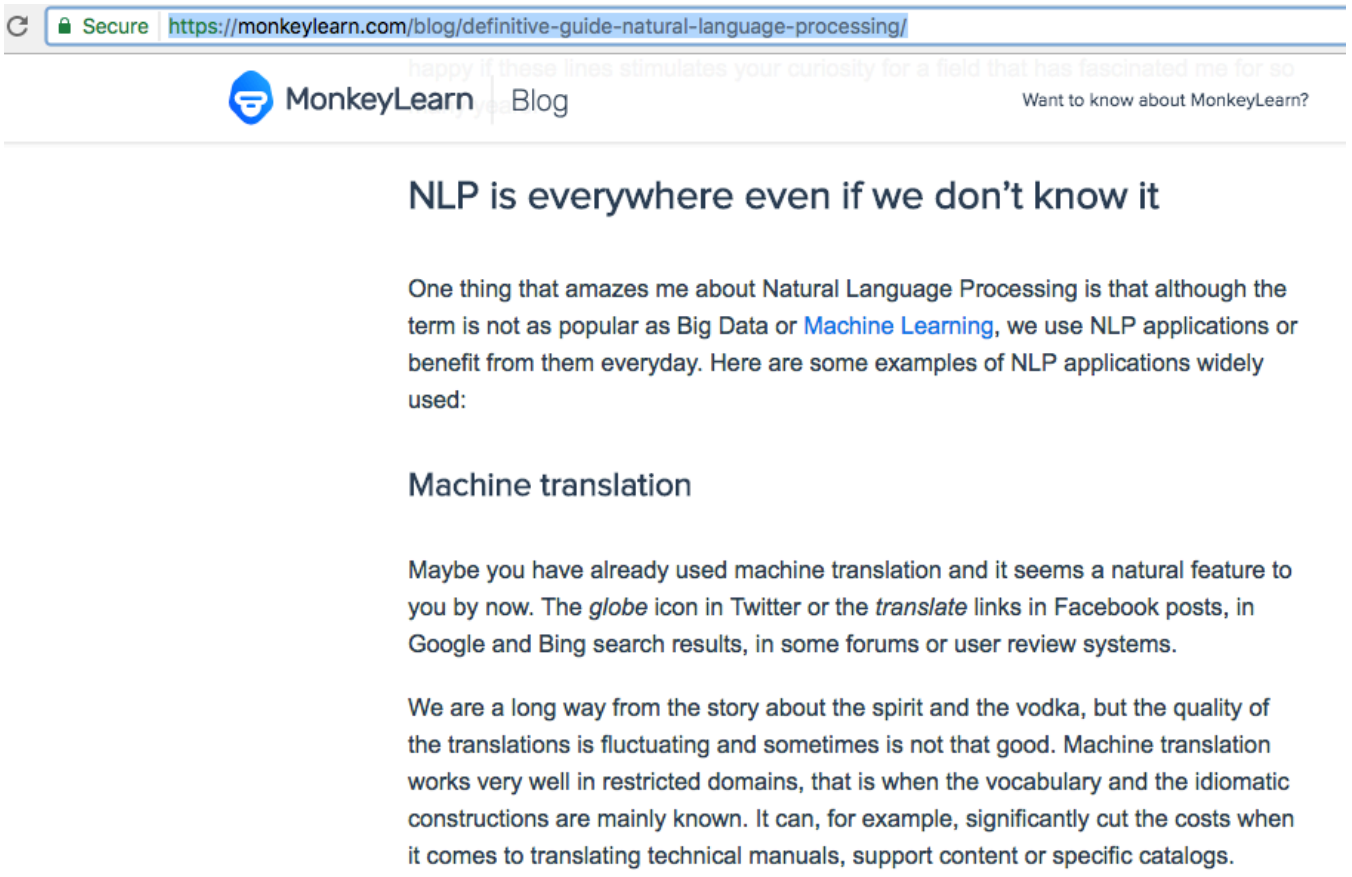
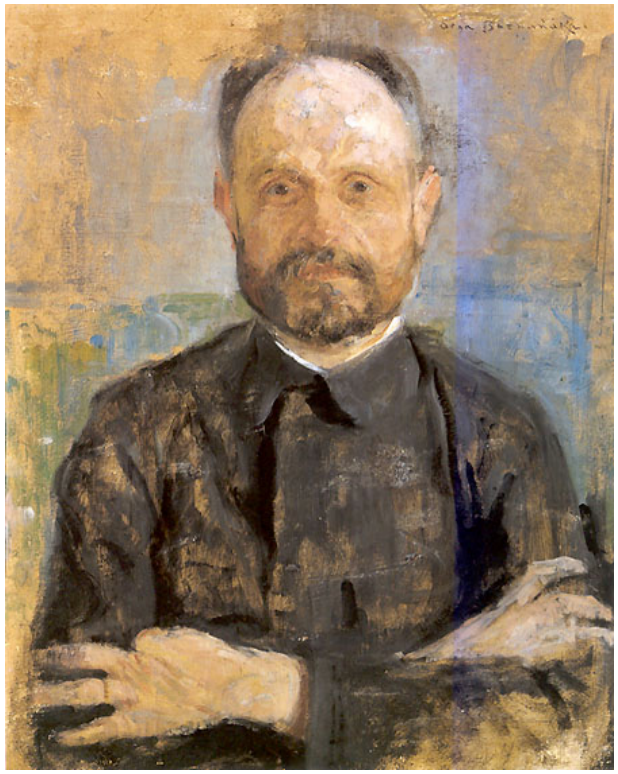
text to data to insight
Data and the Digital World

August 23, 2017

Whitt Kilburn, kilburnw@gvsu.edu, Dept. of Political Science
Matt Schultz, schultzm@gvsu.edu, University Libraries

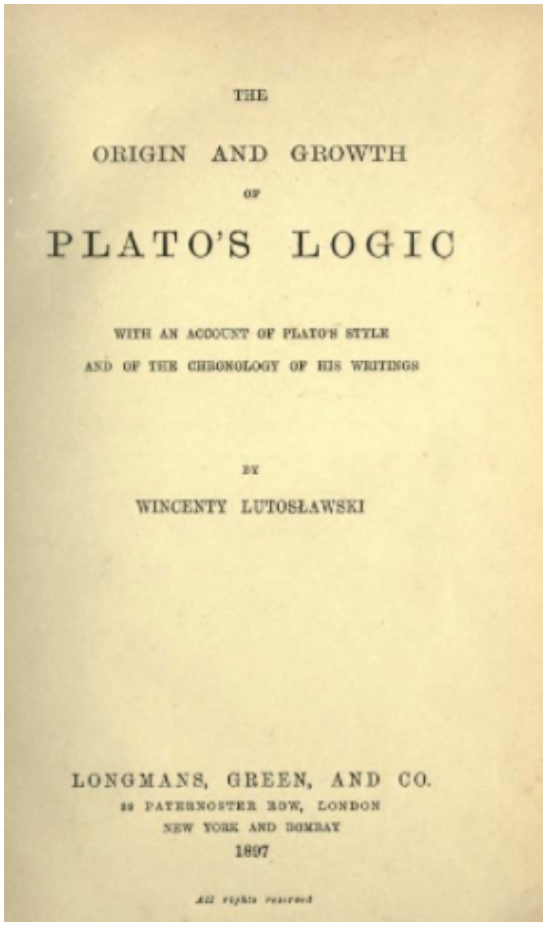
An overview of Natual Language Processing
https://monkeylearn.com/blog/definitive-guide-natural-language-processing/





Wincenty Lutoslawski, 1863-1954
coined the term "stylometry"

Source: https://commons.wikimedia.org/wiki/Main_Page



THE

ORIGIN AND GROWTH

OF

PLATO'S LOGIC

WITH AN ACCOUNT OF PLATO'S STYLE
AND OF THE CHRONOLOGY OF HIS WRITINGS

BY

WINCENTY LUTOSŁAWSKI

LONGMANS, GREEN, AND CO.
39 PATERNOSTER ROW, LONDON
NEW YORK AND BOMBAY
1897

Source: https://archive.org/details/origingrowthofpl00lutoiala

'If handwriting can be so exactly determined as to afford certainty as to its identity, so also with style, since style is more personal and characteristic than handwriting'

The Origin and Growth 1897 (p. 60)

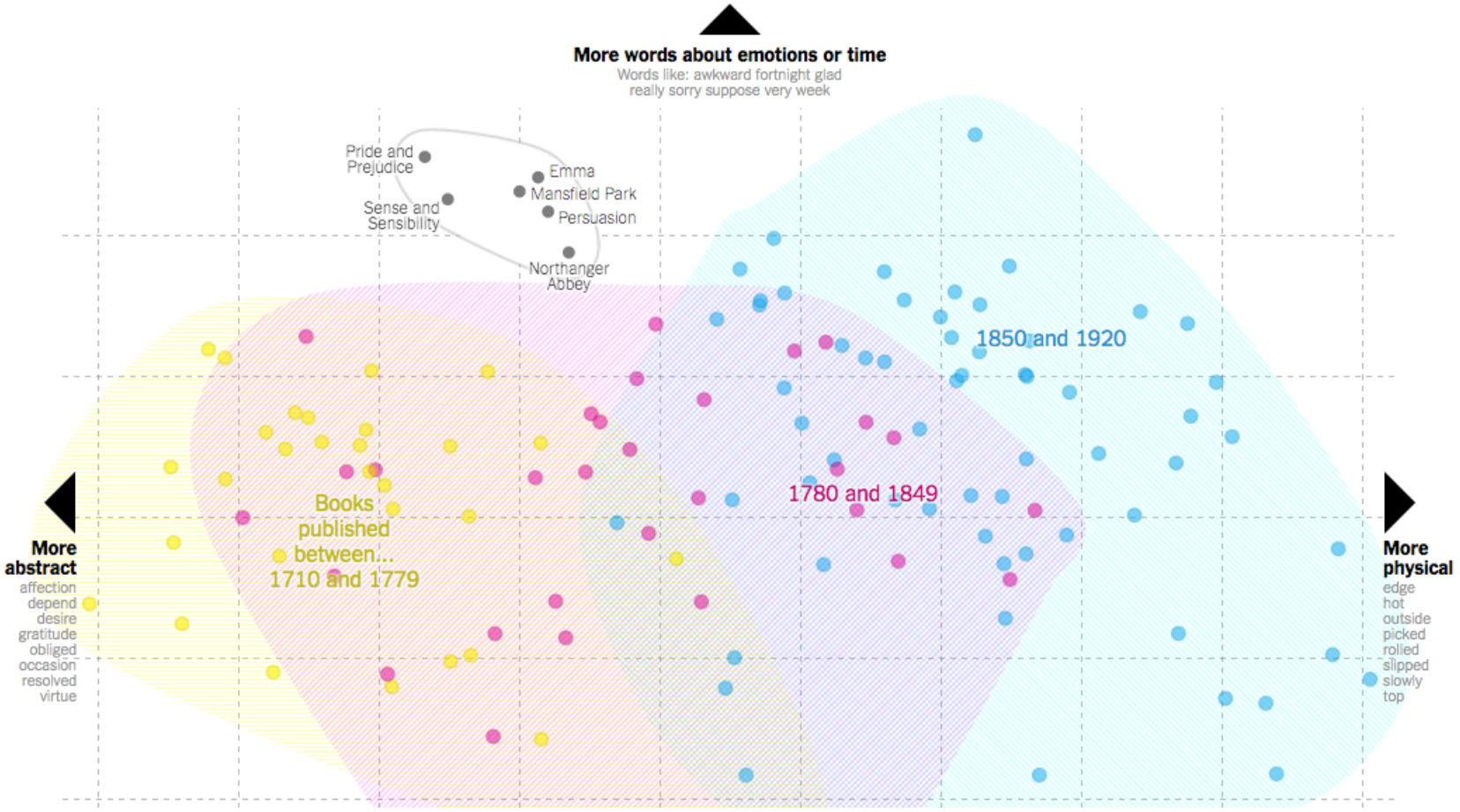
Modern stylometry: relationship between text style and meta characteristics

writing style: patterns of word usage, especially function ('stop') word usage

meta characteristics: text author, gender, chronology, time period, identity, etc.

https://www.nytimes.com/2017/07/06/upshot/the-word-choices-that-explain-why-jane-austen-endures.html

Jane Austen's novels, plotted here along with 125 other British works of narrative fiction published between 1710 and 1920, have a vocabulary that focuses on the abstract more than the physical, and on the quotidian more than the melodramatic.



The location of each book in the chart is based on how often each word in the English language appears in it.

'Function' or 'Stop' words in Jane Austen, *Pride and Prejudice*

It is a truth universally acknowledged, that a single man in possession of a good fortune, must be in want of a wife. However little known the feelings or views of such a man may be on his first entering a neighbourhood, this truth is so well fixed in the minds of the surrounding families, that he is considered the rightful property of some one or other of their daughters.

"My dear Mr. Bennet," said his lady to him one day, "have you heard that Netherfield Park is let at last?"

Mr. Bennet replied that he had not.

"But it is," returned she; "for Mrs. Long has just been here, and she told me all about it."

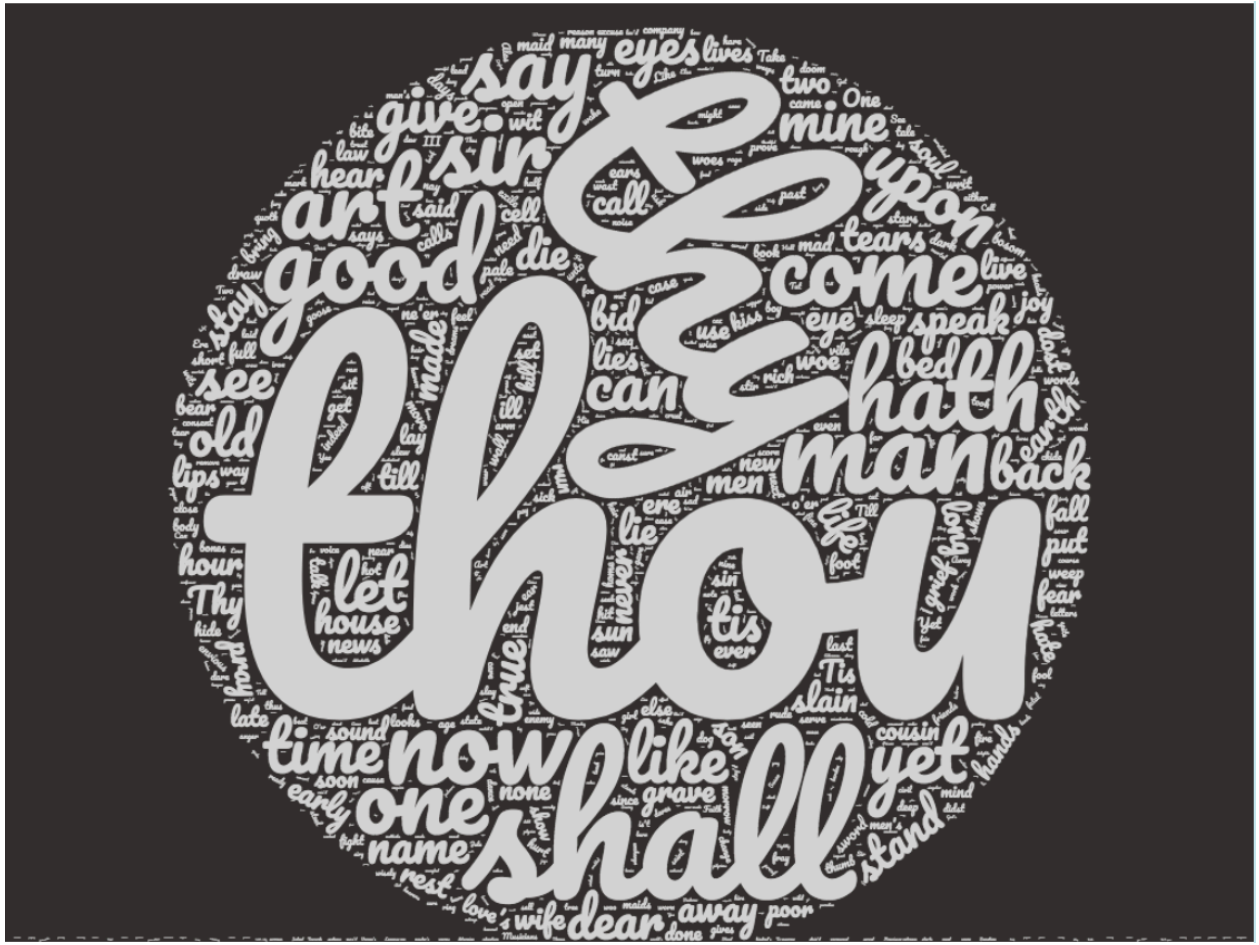
Mr. Bennet made no answer.

"Do you not want to know who has taken it?" cried his wife impatiently.

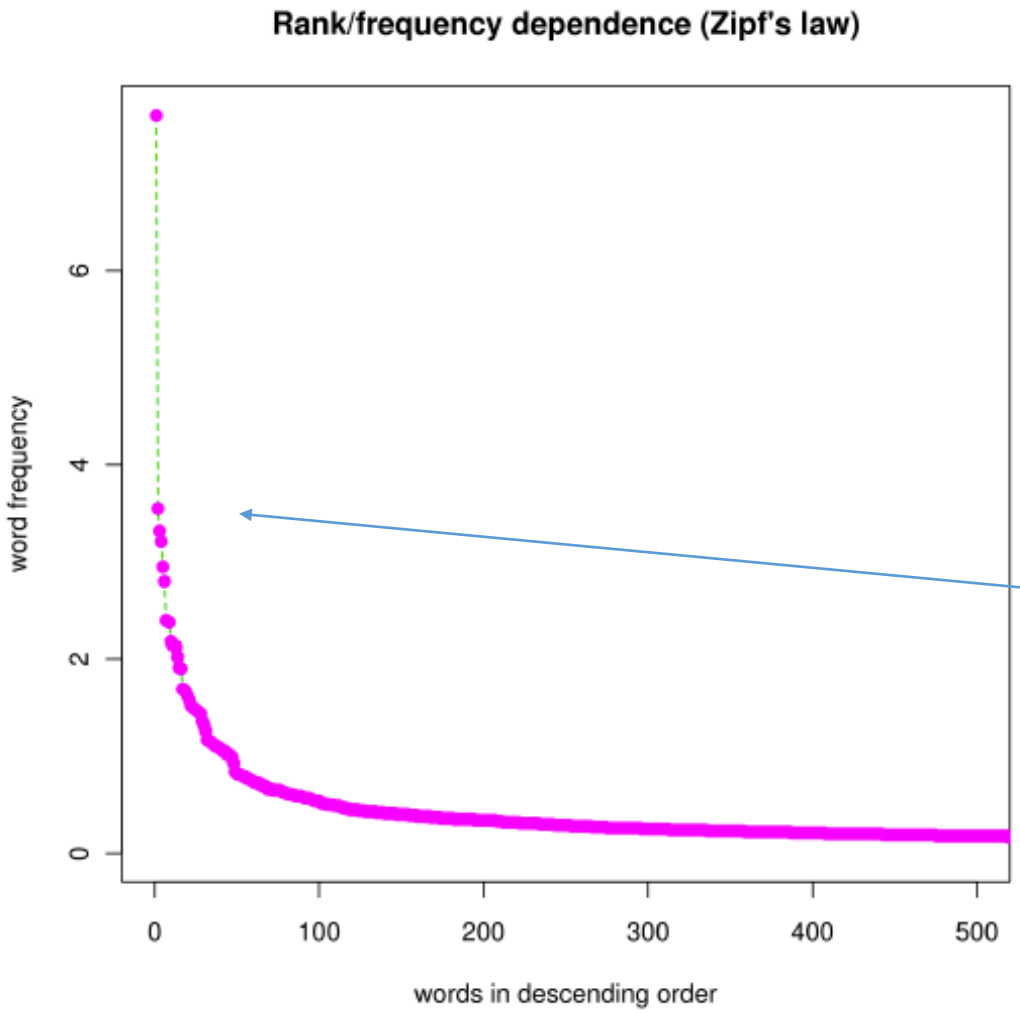
"*You* want to tell me, and I have no objection to hearing it."

This was invitation enough.

Romeo and Juliet: A word cloud (of mostly stop or function words)



Zipf's 1st Law: Rank-Frequency dependence



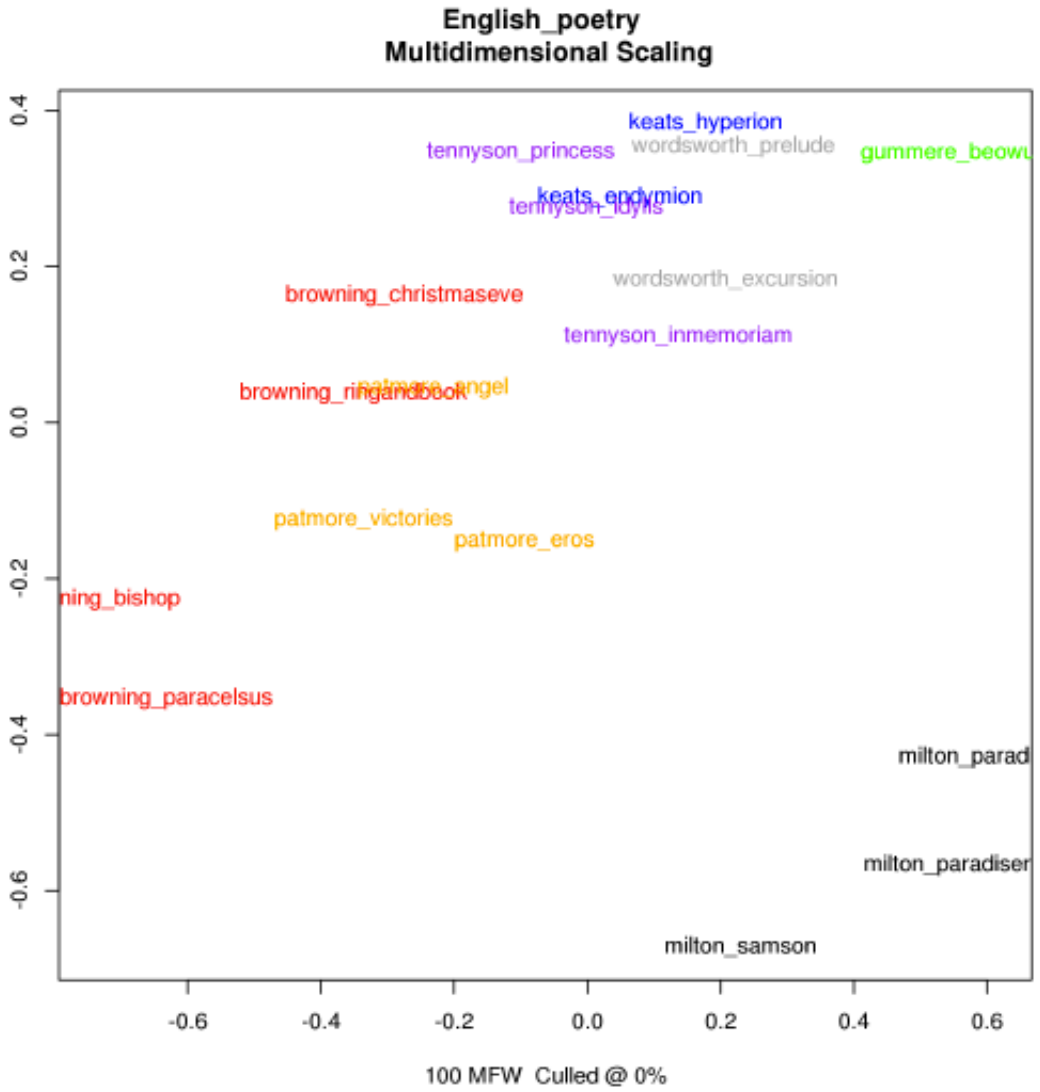
Patterns of usage of the most frequent words ---- the function words --- reveal the author's "fingerprint"

- Prepositions: of, at, in, without, between.
- Pronouns: he, they, anybody, it, one.
- Determiners: the, a, that, my, more, much, either, neither.
- Conjunctions: and, that, when, while, although, or.
- Auxiliary: verbs be (is, am, are), have, got, do.
- Particles: no, not, nor, as.

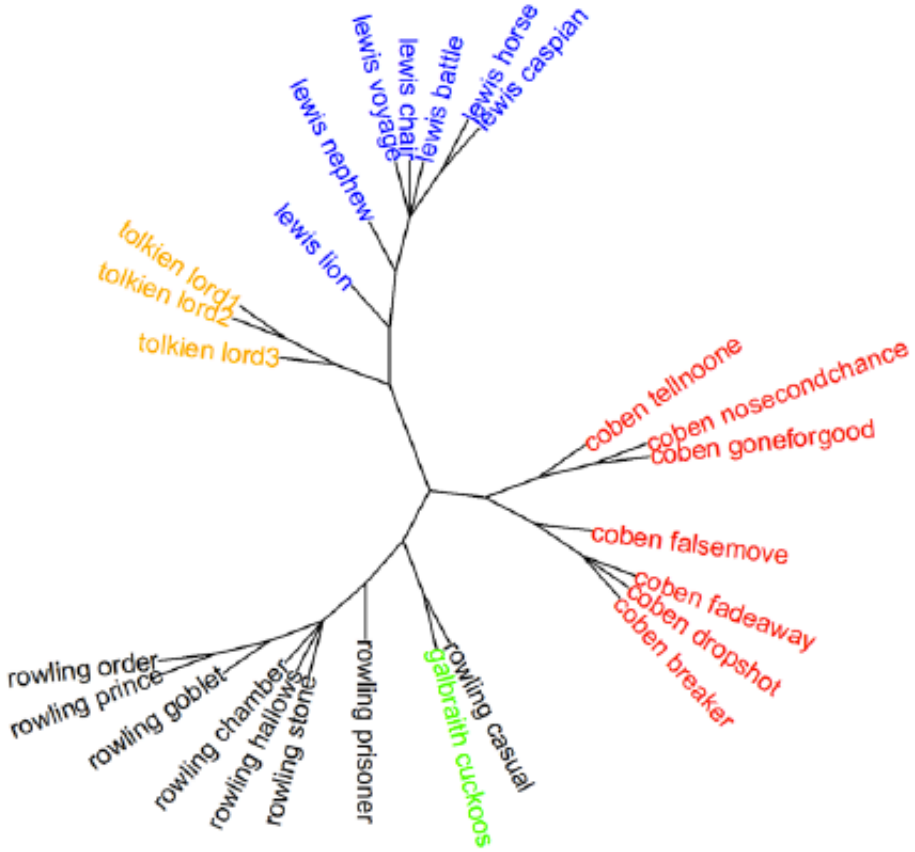
such as modern English 'the', 'in', 'of', 'or', 'I', 'is'

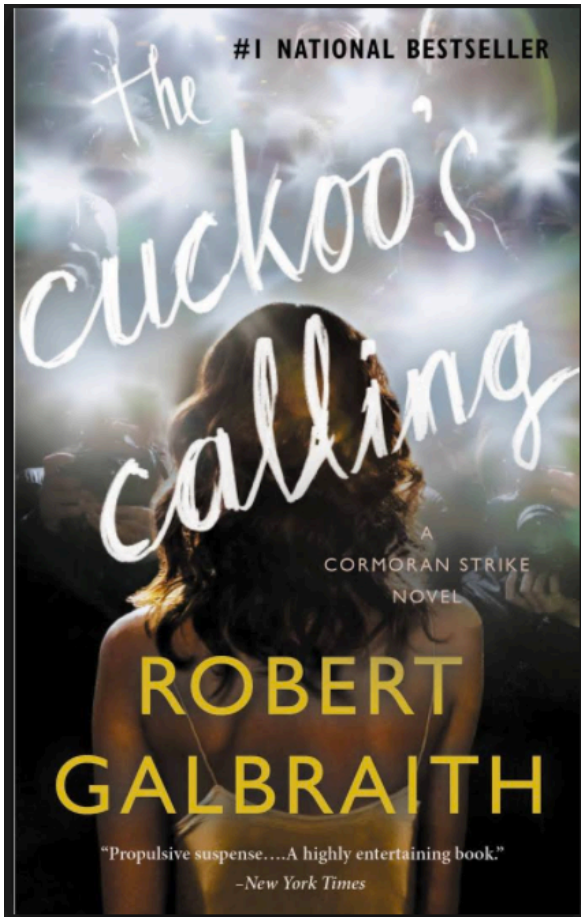
Stylometry: usually beginning with a table of frequencies

	Milton *Samson*	Milton *Paradise*	Keats *Hyperion*	Patmore *Eros*	Browning *Bishop*	...
"the"	4.57	4.24	4.25	4.19	4.47	...
"to"	3.11	3.29	3.43	3.14	3.71	...
"and"	3.19	3	3.08	2.85	2.81	...
"of"	2.6	3	2.63	2.43	2.86	...
"I"	2.17	2.2	2.13	2.42	2.22	...
"a"	2.24	1.92	1.92	2.21	1.92	...
⋮	⋮	⋮	⋮	⋮	⋮	⋱


English_poetry
Multidimensional Scaling
keats_hyperion
tennyson_princess
wordsworth_prelude
gummere_beowu
keats_endymion
tennyson_idylls
wordsworth_excursion
browning_christmaseve
tennyson_inmemoriam
browning_ringandbook
patmore_angel
patmore_victories
patmore_eros
ning_bishop
browning_paracelsus
milton_parad
milton_paradiser
milton_samson
0.4
0.2
0.0
-0.2
-0.4
-0.6
-0.6
-0.4
-0.2
0.0
0.2
0.4
0.6
100 MFW Culled @ 0%

The mystery of Mr. Galbraith and *The Cuckoo's Calling*





Reassessing authorship of the Book of Mormon using delta and nearest shrunken centroid classification*

Matthew L. Jockers; Daniela M. Witten; Craig S. Criddle

Lit Linguist Computing (2008) 23 (4): 465-491.
DOI: https://doi.org/10.1093/llc/fqn040
Published: 06 December 2008

Abstract

Mormon prophet Joseph Smith (1805–44) claimed that more than two-dozen ancient individuals (Nephi, Mormon, Alma, etc.) living from around 2200 BC to 421 AD authored the *Book of Mormon* (1830), and that he translated their inscriptions into English. Later researchers who analyzed selections from the *Book of Mormon* concluded that differences between selections supported Smith's claim of multiple authorship and ancient origins. We offer a new approach that employs two classification techniques: 'delta' commonly used to determine probable authorship and 'nearest shrunken centroid' (NSC), a more generally applicable classifier. We use both methods to

DSH Digital Scholarship in the Humanities

eadh European Association for Digital Humanities

ADHO Alliance of Digital Humanities Organizations

Issues | Advance articles | Submit | Purchase | Alerts | About | All Digital Scholarshi | Advanced Search

No cover image available

Volume 13, Issue 3
September 1998

This article was originally published in Literary and Linguistic Computing

< Previous Next >

The Evolution of Stylometry in Humanities Scholarship

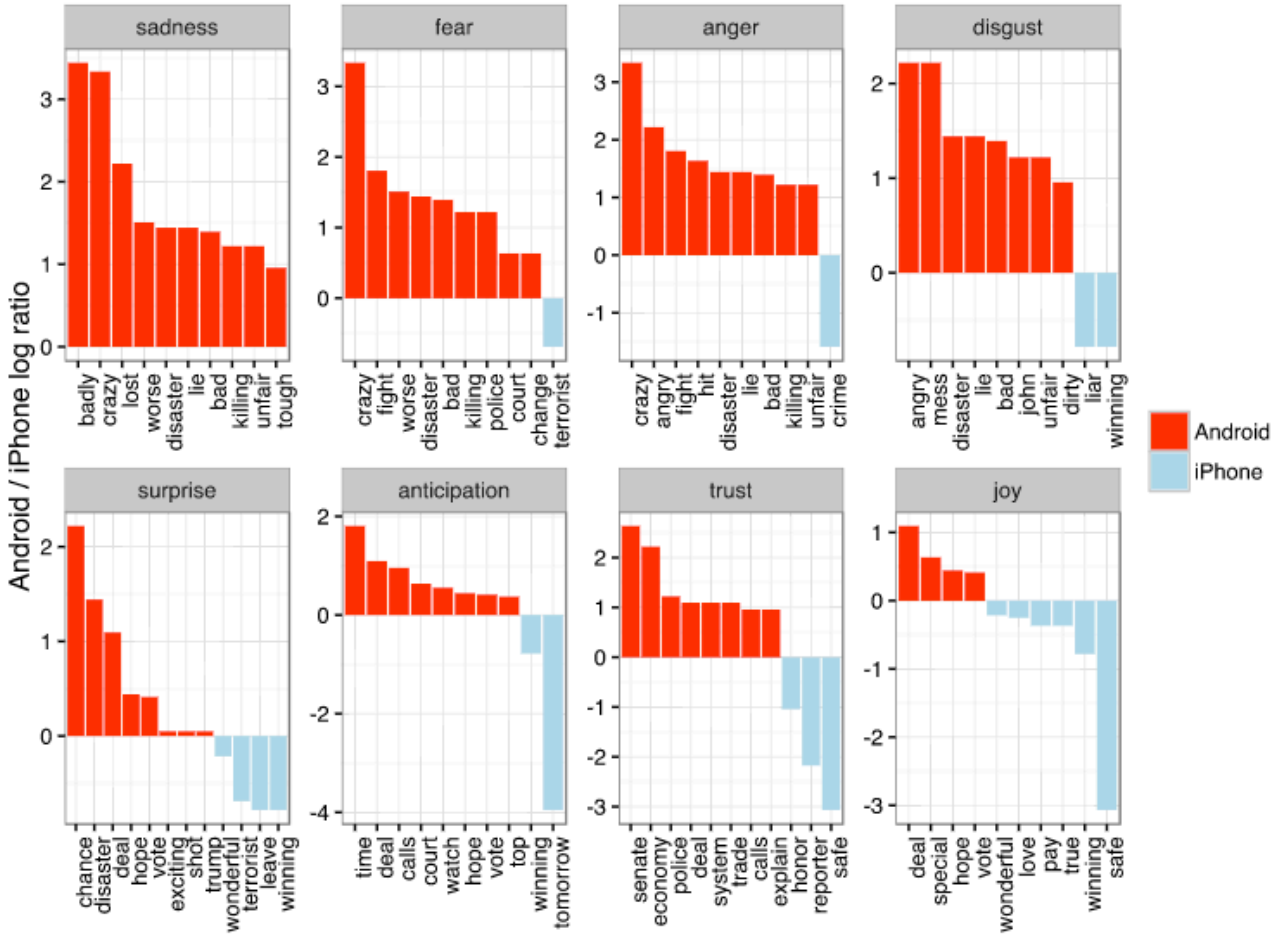
DAVID I. HOLMES

Literary and Linguistic Computing, Volume 13, Issue 3, 1 September 1998, Pages 111–117,
https://doi.org/10.1093/llc/13.3.111
Published: 01 September 1998

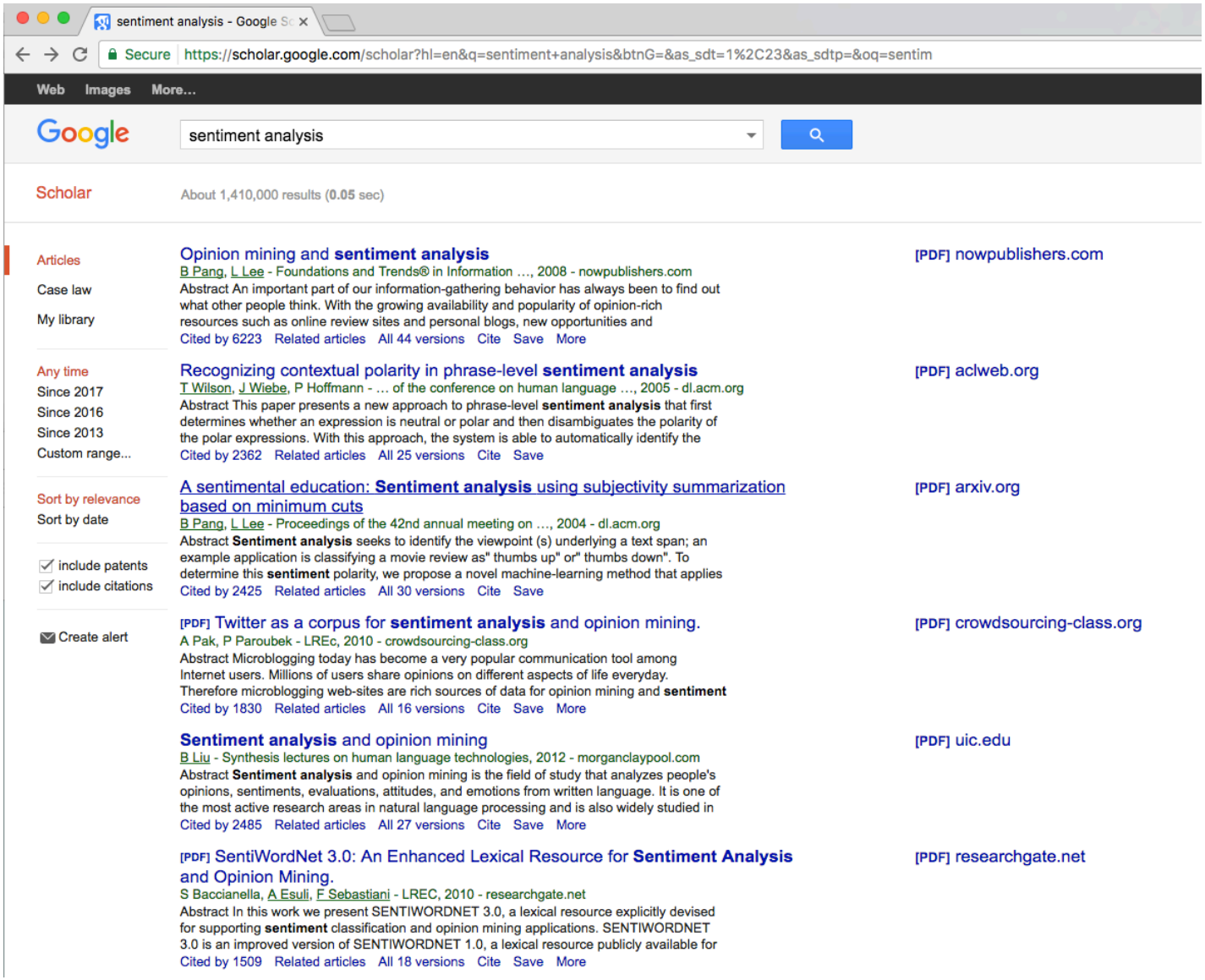
Abstract

This paper traces the historical development of the use of statistical methods in the analysis of literary style. Commencing with stylometry's early origins, the paper looks at both successful and unsuccessful applications, and at the internal struggles as statisticians search for a proven methodology. The growing power of the computer and the ready availability of machine-readable texts are transforming modern stylometry, which has now attracted the attention of the media. Stylometry's interaction with more traditional literary scholarship is also discussed.

Sentiment Analysis: Trump Tweets. *Negative tweets are from one phone type, Android*



Source: http://varianceexplained.org/r/trump-tweets/


sentiment analysis - Google Sc
Secure https://scholar.google.com/scholar?hl=en&q=sentiment+analysis&btnG=&as_sdt=1%2C23&as_sdtp=&oq=sentim
Web Images More...
Google
sentiment analysis
Scholar
About 1,410,000 results (0.05 sec)
Articles
Case law
My library
Any time
Since 2017
Since 2016
Since 2013
Custom range...
Sort by relevance
Sort by date
include patents
include citations
Create alert
Opinion mining and sentiment analysis
B Pang, L Lee - Foundations and Trends® in Information ..., 2008 - nowpublishers.com
Abstract An important part of our information-gathering behavior has always been to find out what other people think. With the growing availability and popularity of opinion-rich resources such as online review sites and personal blogs, new opportunities and
Cited by 6223 Related articles All 44 versions Cite Save More
[PDF] nowpublishers.com
Recognizing contextual polarity in phrase-level sentiment analysis
T Wilson, J Wiebe, P Hoffmann - ... of the conference on human language ..., 2005 - dl.acm.org
Abstract This paper presents a new approach to phrase-level sentiment analysis that first determines whether an expression is neutral or polar and then disambiguates the polarity of the polar expressions. With this approach, the system is able to automatically identify the
Cited by 2362 Related articles All 25 versions Cite Save
[PDF] aclweb.org
A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts
B Pang, L Lee - Proceedings of the 42nd annual meeting on ..., 2004 - dl.acm.org
Abstract Sentiment analysis seeks to identify the viewpoint (s) underlying a text span; an example application is classifying a movie review as" thumbs up" or" thumbs down". To determine this sentiment polarity, we propose a novel machine-learning method that applies
Cited by 2425 Related articles All 30 versions Cite Save
[PDF] arxiv.org
[PDF] Twitter as a corpus for sentiment analysis and opinion mining.
A Pak, P Paroubek - LREc, 2010 - crowdsourcing-class.org
Abstract Microblogging today has become a very popular communication tool among Internet users. Millions of users share opinions on different aspects of life everyday. Therefore microblogging web-sites are rich sources of data for opinion mining and sentiment
Cited by 1830 Related articles All 16 versions Cite Save More
[PDF] crowdsourcing-class.org
Sentiment analysis and opinion mining
B Liu - Synthesis lectures on human language technologies, 2012 - morganclaypool.com
Abstract Sentiment analysis and opinion mining is the field of study that analyzes people's opinions, sentiments, evaluations, attitudes, and emotions from written language. It is one of the most active research areas in natural language processing and is also widely studied in
Cited by 2485 Related articles All 27 versions Cite Save More
[PDF] uic.edu
[PDF] SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining.
S Baccianella, A Esuli, F Sebastiani - LREC, 2010 - researchgate.net
Abstract In this work we present SENTIWORDNET 3.0, a lexical resource explicitly devised for supporting sentiment classification and opinion mining applications. SENTIWORDNET 3.0 is an improved version of SENTIWORDNET 1.0, a lexical resource publicly available for
Cited by 1509 Related articles All 18 versions Cite Save More
[PDF] researchgate.net

JOURNAL ARTICLE

DETECTING MULTIPLE AUTHORSHIP OF UNITED STATES SUPREME COURT LEGAL DECISIONS USING FUNCTION WORDS

Jeffrey S. Rosenthal and Albert H. Yoon

The Annals of Applied Statistics
Vol. 5, No. 1 (March 2011), pp. 283-308

Published by: Institute of Mathematical Statistics
Stable URL: http://www.jstor.org/stable/23024829
Page Count: 26

Topics: Legal judgments, Function words, Authorship attribution, Computer software, Judicial rulings, Dissent, Literary style, Judgment, Null hypothesis, United States Supreme Court

Where text mining usually begins:

Tokenization --- the process of breaking up a text into units of analysis, whether characters, words, sentences, or paragraphs….

Where stylometry (usually) begins: table of frequencies

	Milton *Samson*	Milton *Paradise*	Keats *Hyperion*	Patmore *Eros*	Browning *Bishop*	...
"the"	4.57	4.24	4.25	4.19	4.47	...
"to"	3.11	3.29	3.43	3.14	3.71	...
"and"	3.19	3	3.08	2.85	2.81	...
"of"	2.6	3	2.63	2.43	2.86	...
"I"	2.17	2.2	2.13	2.42	2.22	...
"a"	2.24	1.92	1.92	2.21	1.92	...
⋮	⋮	⋮	⋮	⋮	⋮	⋱

Where stylometry (usually) begins: table of frequencies

	Milton *Samson*	Milton *Paradise*	Keats *Hyperion*	Patmore *Eros*	Browning *Bishop*	...
"the"	4.57	4.24	4.25	4.19	4.47	...
"to"	3.11	3.29	3.43	3.14	3.71	...
"and"	3.19	3	3.08	2.85	2.81	...
"of"	2.6	3	2.63	2.43	2.86	...
"I"	2.17	2.2	2.13	2.42	2.22	...
"a"	2.24	1.92	1.92	2.21	1.92	...
⋮	⋮	⋮	⋮	⋮	⋮	⋱

$|a_1 - b_1|$

Where stylometry (usually) begins: table of frequencies

	Milton *Samson*	Milton *Paradise*	Keats *Hyperion*	Patmore *Eros*	Browning *Bishop*	...
"the"	4.57	4.24	4.25	4.19	4.47	...
"to"	3.11	3.29	3.43	3.14	3.71	...
"and"	3.19	3	3.08	2.85	2.81	...
"of"	2.6	3	2.63	2.43	2.86	...
"I"	2.17	2.2	2.13	2.42	2.22	...
"a"	2.24	1.92	1.92	2.21	1.92	...
⋮	⋮	⋮	⋮	⋮	⋮	⋱

$$|a_1 - b_1| + |a_2 - b_2|$$

Where stylometry (usually) begins: table of frequencies

	Milton *Samson*	Milton *Paradise*	Keats *Hyperion*	Patmore *Eros*	Browning *Bishop*	...
"the"	4.57	4.24	4.25	4.19	4.47	...
"to"	3.11	3.29	3.43	3.14	3.71	...
"and"	3.19	3	3.08	2.85	2.81	...
"of"	2.6	3	2.63	2.43	2.86	...
"I"	2.17	2.2	2.13	2.42	2.22	...
"a"	2.24	1.92	1.92	2.21	1.92	...
⋮	⋮	⋮	⋮	⋮	⋮	⋱

$$|a_1 - b_1| + |a_2 - b_2| + |a_3 - b_3|$$

Where stylometry (usually) begins: table of frequencies

	Milton *Samson*	Milton *Paradise*	Keats *Hyperion*	Patmore *Eros*	Browning *Bishop*	...
"the"	4.57	4.24	4.25	4.19	4.47	...
"to"	3.11	3.29	3.43	3.14	3.71	...
"and"	3.19	3	3.08	2.85	2.81	...
"of"	2.6	3	2.63	2.43	2.86	...
"I"	2.17	2.2	2.13	2.42	2.22	...
"a"	2.24	1.92	1.92	2.21	1.92	...
⋮	⋮	⋮	⋮	⋮	⋮	⋱

$$|a_1 - b_1| + |a_2 - b_2| + |a_3 - b_3| + \ldots$$

Where stylometry (usually) begins: table of frequencies

	Milton *Samson*	Milton *Paradise*	Keats *Hyperion*	Patmore *Eros*	Browning *Bishop*	...
"the"	4.57	4.24	4.25	4.19	4.47	...
"to"	3.11	3.29	3.43	3.14	3.71	...
"and"	3.19	3	3.08	2.85	2.81	...
"of"	2.6	3	2.63	2.43	2.86	...
"I"	2.17	2.2	2.13	2.42	2.22	...
"a"	2.24	1.92	1.92	2.21	1.92	...
⋮	⋮	⋮	⋮	⋮	⋮	⋱

$$|a_1 - b_1| + |a_2 - b_2| + |a_3 - b_3| + \ldots + |a_n - b_n|$$

Where stylometry (usually) begins: table of frequencies

	Milton *Samson*	Milton *Paradise*	Keats *Hyperion*	Patmore *Eros*	Browning *Bishop*	...
"the"	4.57	4.24	4.25	4.19	4.47	...
"to"	3.11	3.29	3.43	3.14	3.71	...
"and"	3.19	3	3.08	2.85	2.81	...
"of"	2.6	3	2.63	2.43	2.86	...
"I"	2.17	2.2	2.13	2.42	2.22	...
"a"	2.24	1.92	1.92	2.21	1.92	...
⋮	⋮	⋮	⋮	⋮	⋮	⋱

$$|a_1 - b_1| + |a_2 - b_2| + |a_3 - b_3| + \ldots + |a_n - b_n| = \sum_{i=1}^{n} |a_i - b_i|$$

Where stylometry (usually) begins: table of frequencies

	Milton *Samson*	Milton *Paradise*	Keats *Hyperion*	Patmore *Eros*	Browning *Bishop*	...
"the"	4.57	4.24	4.25	4.19	4.47	...
"to"	3.11	3.29	3.43	3.14	3.71	...
"and"	3.19	3	3.08	2.85	2.81	...
"of"	2.6	3	2.63	2.43	2.86	...
"I"	2.17	2.2	2.13	2.42	2.22	...
"a"	2.24	1.92	1.92	2.21	1.92	...
⋮	⋮	⋮	⋮	⋮	⋮	⋱

$$\sum_{i=1}^{n} |a_i - b_i|$$

$$\delta_{(b,c)}$$

Where stylometry (usually) begins: table of frequencies

	Milton *Samson*	Milton *Paradise*	Keats *Hyperion*	Patmore *Eros*	Browning *Bishop*	...
"the"	4.57	4.24	4.25	4.19	4.47	...
"to"	3.11	3.29	3.43	3.14	3.71	...
"and"	3.19	3	3.08	2.85	2.81	...
"of"	2.6	3	2.63	2.43	2.86	...
"I"	2.17	2.2	2.13	2.42	2.22	...
"a"	2.24	1.92	1.92	2.21	1.92	...
⋮	⋮	⋮	⋮	⋮	⋮	⋱

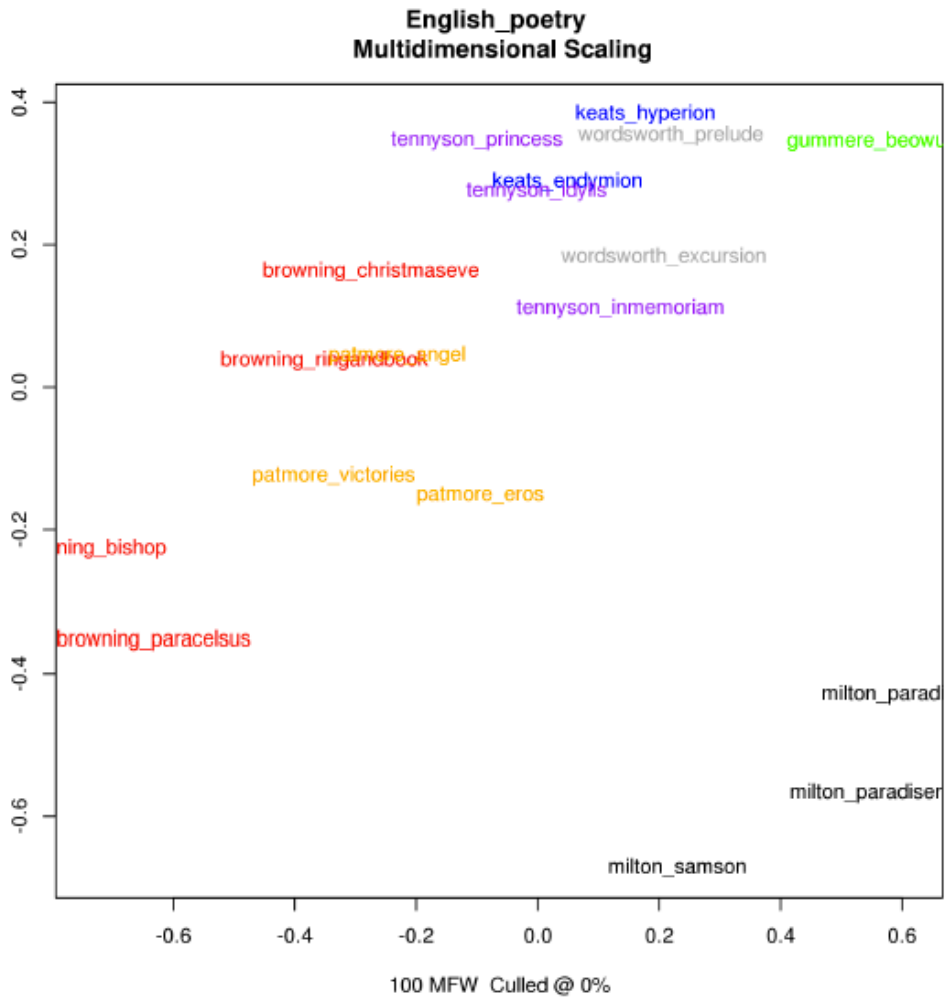
$$\sum_{i=1}^{n} |a_i - b_i|$$

$$\delta_{(b,c)} \quad \delta_{(c,d)} \quad \delta_{(d,e)} \quad \delta_{(a,c)} \quad \delta_{(b,d)} \quad \delta_{(c,e)} \quad \delta_{(a,d)} \quad \cdots$$

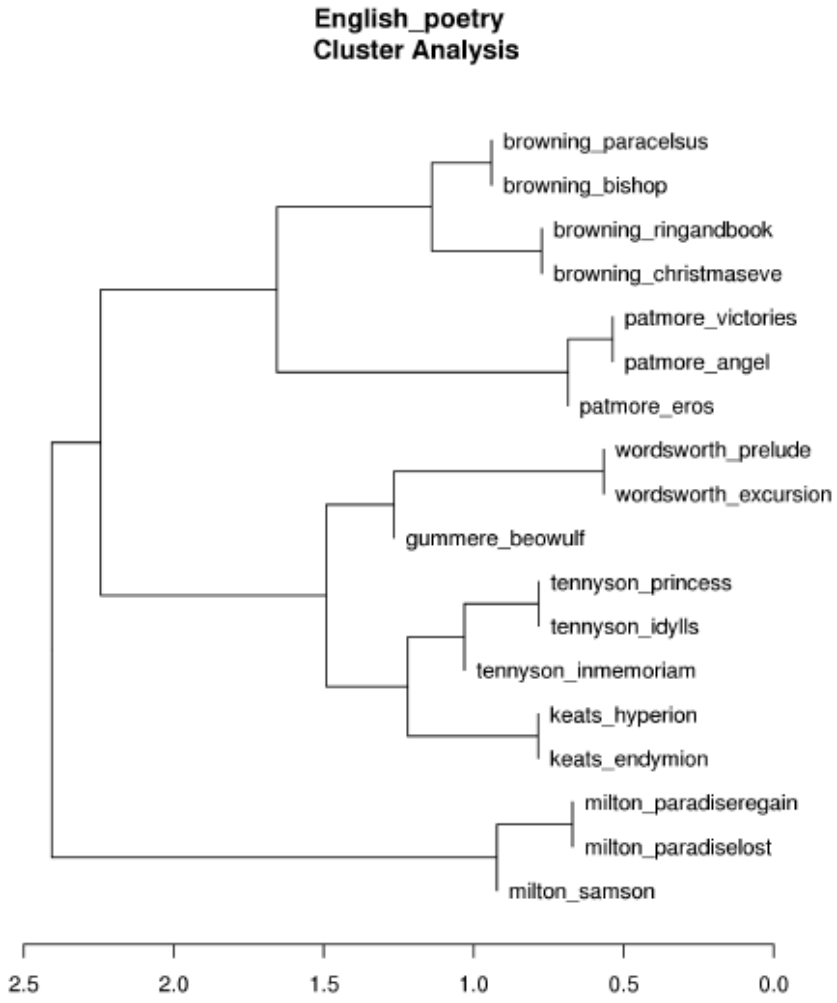
Table of calculated distances

	Milton *Samson*	Milton *Paradise*	Keats *Hyperion*	Patmore *Eros*	Browning *Bishop*	...
Milton *Samson*	0	0.9839	1.12	1.0493	1.0864	...
Milton *Paradise*	0.9839	0	1.0891	1.089	1.1047	...
Keats *Hyperion*	1.12	1.0891	0	1.128	1.11	...
Patmore *Eros*	1.0493	1.089	1.128	0	1.1128	...
Browning *Bishop*	1.0864	1.1047	1.11	1.1128	0	...
⋮	⋮	⋮	⋮	⋮	⋮	⋱

Explanatory approach: Multidimensional Scaling



Cluster analysis: 100 MFWs



Quick Notes:

Cluster method in stylo() is *hierarchical clustering.*

Distances metrics: Burrow's Delta (classic Delta) is *Manhattan distance* on standardized (z distributed) scores.

Other metrics: *Euclidean.*

Consult the stylo() manual for further info….And Google!

Brief overview of sentiment analysis: extraction of an author's emotional intent from a text

Popular Emotional Lexicon: Mohammad's NRC
Dictionary of crowd-sourced terms associated with 8 emotional states, from theory of psychologist Robert Plutchik:

1) anger
2) fear
3) sadness
4) disgust
5) surprise
6) anticipation
7) trust
8) joy

In the bag of words approach, a text's sentiment is scored by the presence of words associated with an emotional state.

http://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm