Text as data

Christopher Barrie

Research Training Centre Micro-Methods Workshop April, 2021.

Introduction

Introduction

- Computational techniques for the analysis of text are new

But:

- We also have historical precedents...

Early examples

Librum Prohibitorum, 1564, *Sacred Congregation of the Index*

Early examples

On Deciphering Cryptographic Messages (رسالة في استخراج الكتب المعماة), C.9 AD, Al-Kindi.

Figure 2a. Al-Kindī's tree diagram classification of cipher types as it appears in his manuscript [12].

Early examples

Analysis of Four Newspapers, John Gilmer Speed, 1893, *Forum*.

Results of Speed's Analysis of Four Newspapers Published on April 17, 1881, and April 16, 1893
(Numbers other than percent change are total columns)

Subject	TRIBUNE			WORLD			TIMES			SUN		
	1881	1893	% Change	1881	1893	% Change	1881	1893	% Change	1881	1893	% Change
Editorial	5.00	5.00	0	4.75	4.00	-15.9	6.00	5.00	-16.67	4.00	4.00	0
Religious	2.00	0.00	-100	0.75	0.00	-100	1.00	0.00	-100	0.50	1.00	100
Scientific	1.00	0.75	-25	0.00	2.00	-	1.00	0.00	-100	0.00	2.50	-
Political	3.00	3.75	-25	0.00	10.50	-	1.00	4.00	300	1.00	3.50	250
Literary	15.00	5.00	-66.67	1.00	2.00	100	18.00	12.00	-33.33	5.75	6.00	4.35
Gossip	1.00	23.00	2,200	1.00	63.50	6,250	0.50	16.75	3,250	2.00	13.00	550
Scandals	0.00	1.50	-	0.00	1.50	-	1.00	2.50	150	0.00	2.00	-
Sporting	1.00	6.50	550	2.50	16.00	540	3.00	10.00	233.33	0.50	17.50	3,400
Fiction	0.00	7.00	-	1.50	6.50	333.33	1.00	1.50	50	0.00	11.50	-
Historical	2.50	2.50	0	2.75	4.00	45.45	2.50	1.50	-40	4.25	14.00	229.41
Music, Drama	2.50	4.00	60	1.00	11.00	1,000	4.00	7.00	75	0.00	3.50	-
Crimes	0.00	0.50	-	0.00	6.00	-	0.00	1.00	-	0.00	0.00	-
Art	1.00	1.00	0	3.00	3.00	0	2.00	0.00	-100	0.25	1.25	400%

"...and, to be specific, we will now start to measure, with scissors and compasses, how the contents of newspapers have quantitatively shifted during the last generation."

"From these quantitative accounts, we will proceed to qualitative considerations. We will have to study the stylization of newspapers, how the same problems are discussed inside and outside the newspapers... Then, we may finally approach the point where we have reasons to hope for a slow approximation to our wide-ranging questions."

Max Weber, speech delivered at the first Congress of Sociologists, meeting in Frankfurt, 1910. Translated by Klaus Krippendorff.

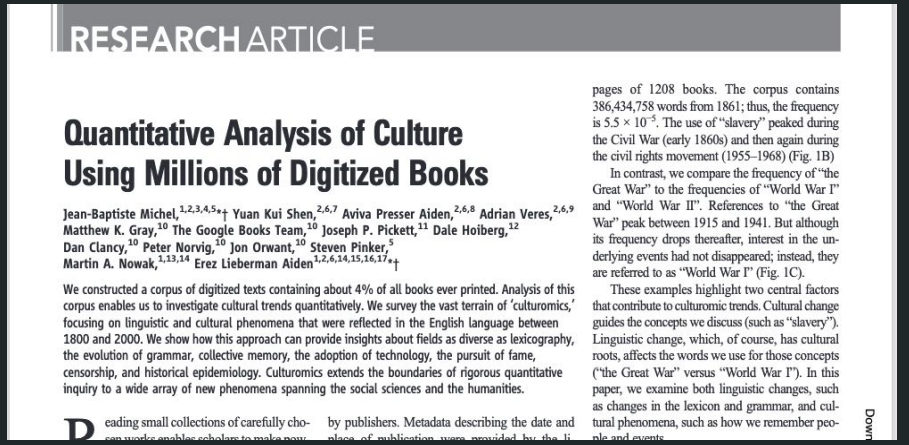
Where is the data?

- Now digitized >25m books
- Sourced from university libraries worldwide
- Slowed in recent years

Access datasets here: http://storage.googleapis.com/books/ngrams/books/datasetsv2.html.

Uses of Google books data:

Analysis of cross-temporal national cultural norms: Michel et al. 2011. *Science*.



RESEARCH ARTICLE

Quantitative Analysis of Culture Using Millions of Digitized Books

Jean-Baptiste Michel,[1,2,3,4,5]*† Yuan Kui Shen,[2,6,7] Aviva Presser Aiden,[2,6,8] Adrian Veres,[2,6,9] Matthew K. Gray,[10] The Google Books Team,[10] Joseph P. Pickett,[11] Dale Hoiberg,[12] Dan Clancy,[10] Peter Norvig,[10] Jon Orwant,[10] Steven Pinker,[5] Martin A. Nowak,[1,13,14] Erez Lieberman Aiden[1,2,6,14,15,16,17]*†

We constructed a corpus of digitized texts containing about 4% of all books ever printed. Analysis of this corpus enables us to investigate cultural trends quantitatively. We survey the vast terrain of 'culturomics,' focusing on linguistic and cultural phenomena that were reflected in the English language between 1800 and 2000. We show how this approach can provide insights about fields as diverse as lexicography, the evolution of grammar, collective memory, the adoption of technology, the pursuit of fame, censorship, and historical epidemiology. Culturomics extends the boundaries of rigorous quantitative inquiry to a wide array of new phenomena spanning the social sciences and the humanities.

Reading small collections of carefully chosen works enables scholars to make new

by publishers. Metadata describing the date and place of publication were provided by the li-

pages of 1208 books. The corpus contains 386,434,758 words from 1861; thus, the frequency is 5.5×10^{-5}. The use of "slavery" peaked during the Civil War (early 1860s) and then again during the civil rights movement (1955–1968) (Fig. 1B)

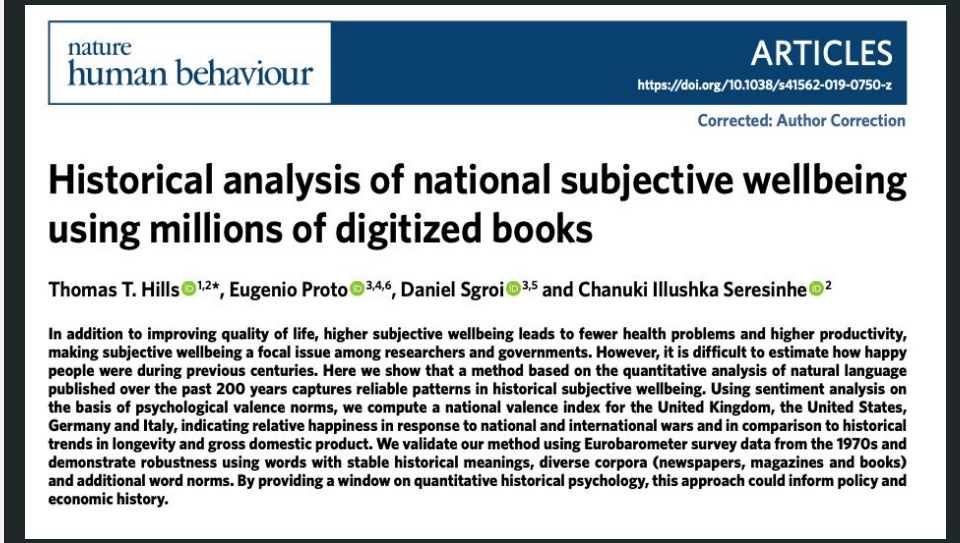
In contrast, we compare the frequency of "the Great War" to the frequencies of "World War I" and "World War II". References to "the Great War" peak between 1915 and 1941. But although its frequency drops thereafter, interest in the underlying events had not disappeared; instead, they are referred to as "World War I" (Fig. 1C).

These examples highlight two central factors that contribute to culturomic trends. Cultural change guides the concepts we discuss (such as "slavery"). Linguistic change, which, of course, has cultural roots, affects the words we use for those concepts ("the Great War" versus "World War I"). In this paper, we examine both linguistic changes, such as changes in the lexicon and grammar, and cultural phenomena, such as how we remember people and events

Down

Uses of Google books data:

Analysis of cross-national subjective wellbeing:
Hills et al. 2019. *Nature Human Behaviour*.



nature human behaviour

ARTICLES

https://doi.org/10.1038/s41562-019-0750-z

Corrected: Author Correction

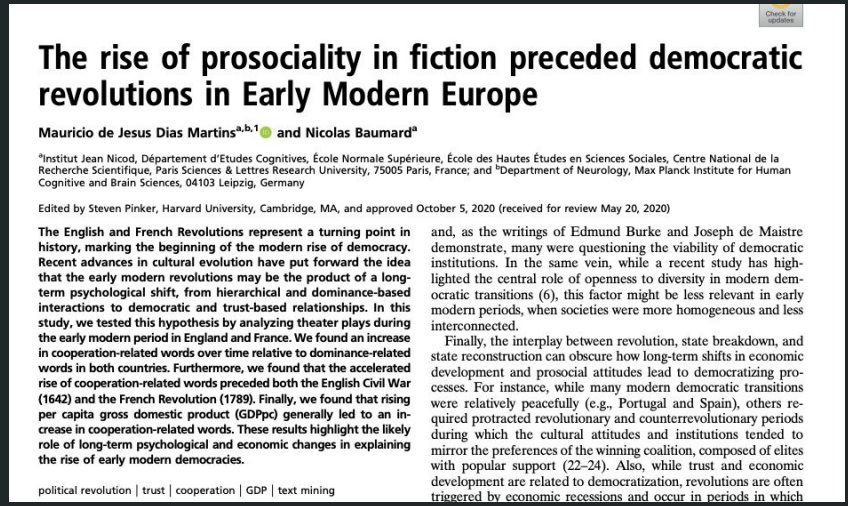
Historical analysis of national subjective wellbeing using millions of digitized books

Thomas T. Hills[1,2*], Eugenio Proto[3,4,6], Daniel Sgroi[3,5] and Chanuki Illushka Seresinhe[2]

In addition to improving quality of life, higher subjective wellbeing leads to fewer health problems and higher productivity, making subjective wellbeing a focal issue among researchers and governments. However, it is difficult to estimate how happy people were during previous centuries. Here we show that a method based on the quantitative analysis of natural language published over the past 200 years captures reliable patterns in historical subjective wellbeing. Using sentiment analysis on the basis of psychological valence norms, we compute a national valence index for the United Kingdom, the United States, Germany and Italy, indicating relative happiness in response to national and international wars and in comparison to historical trends in longevity and gross domestic product. We validate our method using Eurobarometer survey data from the 1970s and demonstrate robustness using words with stable historical meanings, diverse corpora (newspapers, magazines and books) and additional word norms. By providing a window on quantitative historical psychology, this approach could inform policy and economic history.

Uses of literary data:

Analysis of cross-temporal national cultural norms:
Dias Martins, Mauricio de Jesus and Nicolas Baumard. 2020. *PNAS*.



Check for updates

The rise of prosociality in fiction preceded democratic revolutions in Early Modern Europe

Mauricio de Jesus Dias Martins[a,b,1] and Nicolas Baumard[a]

[a]Institut Jean Nicod, Département d'Études Cognitives, École Normale Supérieure, École des Hautes Études en Sciences Sociales, Centre National de la Recherche Scientifique, Paris Sciences & Lettres Research University, 75005 Paris, France; and [b]Department of Neurology, Max Planck Institute for Human Cognitive and Brain Sciences, 04103 Leipzig, Germany

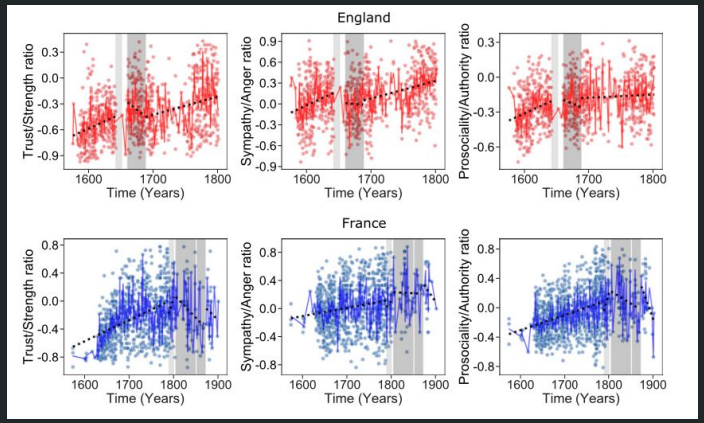
Edited by Steven Pinker, Harvard University, Cambridge, MA, and approved October 5, 2020 (received for review May 20, 2020)

The English and French Revolutions represent a turning point in history, marking the beginning of the modern rise of democracy. Recent advances in cultural evolution have put forward the idea that the early modern revolutions may be the product of a long-term psychological shift, from hierarchical and dominance-based interactions to democratic and trust-based relationships. In this study, we tested this hypothesis by analyzing theater plays during the early modern period in England and France. We found an increase in cooperation-related words over time relative to dominance-related words in both countries. Furthermore, we found that the accelerated rise of cooperation-related words preceded both the English Civil War (1642) and the French Revolution (1789). Finally, we found that rising per capita gross domestic product (GDPpc) generally led to an increase in cooperation-related words. These results highlight the likely role of long-term psychological and economic changes in explaining the rise of early modern democracies.

political revolution | trust | cooperation | GDP | text mining

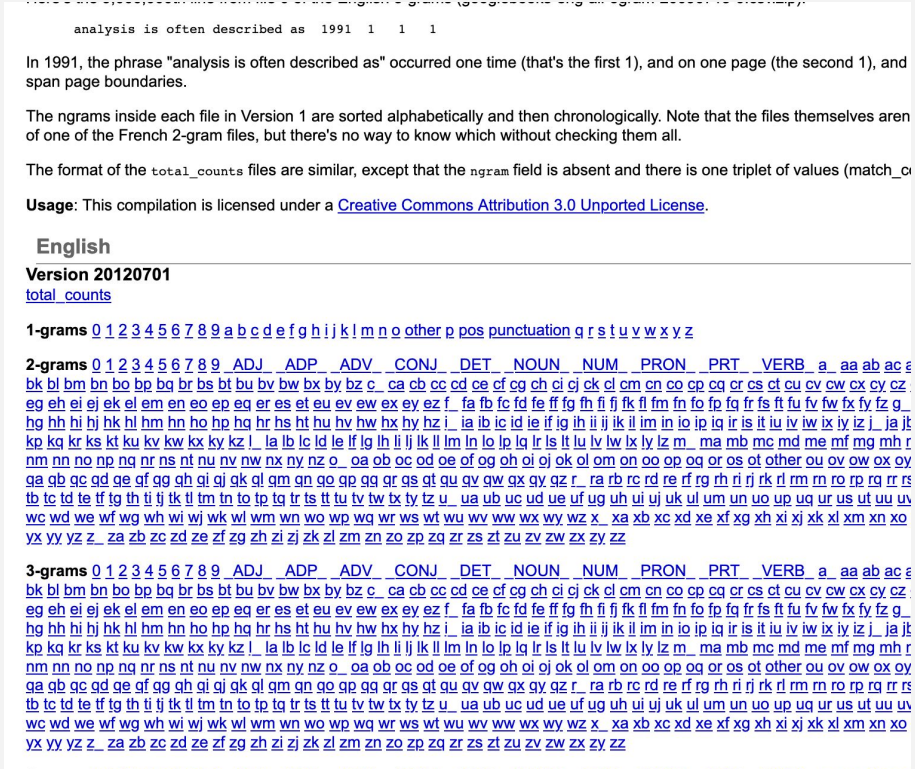
and, as the writings of Edmund Burke and Joseph de Maistre demonstrate, many were questioning the viability of democratic institutions. In the same vein, while a recent study has highlighted the central role of openness to diversity in modern democratic transitions (6), this factor might be less relevant in early modern periods, when societies were more homogeneous and less interconnected.

Finally, the interplay between revolution, state breakdown, and state reconstruction can obscure how long-term shifts in economic development and prosocial attitudes lead to democratizing processes. For instance, while many modern democratic transitions were relatively peacefully (e.g., Portugal and Spain), others required protracted revolutionary and counterrevolutionary periods during which the cultural attitudes and institutions tended to mirror the preferences of the winning coalition, composed of elites with popular support (22–24). Also, while trust and economic development are related to democratization, revolutions are often triggered by economic recessions and occur in periods in which



- *unigram:* ‘new’ ‘york’ ‘city’
- *bigram(s)*: ‘new_york’ ‘york_city’
- *trigram(s)*: ‘new_york_city’

Access datasets here:
http://storage.googleapis.com/books/ngrams/books/datasetsv2.html



```
analysis is often described as  1991  1   1   1
```

In 1991, the phrase "analysis is often described as" occurred one time (that's the first 1), and on one page (the second 1), and span page boundaries.

The ngrams inside each file in Version 1 are sorted alphabetically and then chronologically. Note that the files themselves aren of one of the French 2-gram files, but there's no way to know which without checking them all.

The format of the `total_counts` files are similar, except that the `ngram` field is absent and there is one triplet of values (match_c

Usage: This compilation is licensed under a Creative Commons Attribution 3.0 Unported License.

English

Version 20120701
total_counts

1-grams 0 1 2 3 4 5 6 7 8 9 a b c d e f g h i j k l m n o other p pos punctuation q r s t u v w x y z

2-grams 0 1 2 3 4 5 6 7 8 9 _ADJ_ _ADP_ _ADV_ _CONJ_ _DET_ _NOUN_ _NUM_ _PRON_ _PRT_ _VERB_ a_ aa ab ac a
bk bl bm bn bo bp bq br bs bt bu bv bw bx by bz c_ ca cb cc cd ce cf cg ch ci cj ck cl cm cn co cp cq cr cs ct cu cv cw cx cy cz
eg eh ei ej ek el em en eo ep eq er es et eu ev ew ex ey ez f_ fa fb fc fd fe ff fg fh fi fj fk fl fm fn fo fp fq fr fs ft fu fv fw fx fy fz g_
hg hh hi hj hk hl hm hn ho hp hq hr hs ht hu hv hw hx hy hz i_ ia ib ic id ie if ig ih ii ij ik il im in io ip iq ir is it iu iv iw ix iy iz j_ ja jt
kp kq kr ks kt ku kv kw kx ky kz l_ la lb lc ld le lf lg lh li lj lk ll lm ln lo lp lq lr ls lt lu lv lw lx ly lz m_ ma mb mc md me mf mg mh r
nm nn no np nq nr ns nt nu nv nw nx ny nz o_ oa ob oc od oe of og oh oi oj ok ol om on oo op oq or os ot other ou ov ow ox oy
qa qb qc qd qe qf qg qh qi qj qk ql qm qn qo qp qq qr qs qt qu qv qw qx qy qz r_ ra rb rc rd re rf rg rh ri rj rk rl rm rn ro rp rq rr rs
tb tc td te tf tg th ti tj tk tl tm tn to tp tq tr ts tt tu tv tw tx ty tz u_ ua ub uc ud ue uf ug uh ui uj uk ul um un uo up uq ur us ut uu uv
wc wd we wf wg wh wi wj wk wl wm wn wo wp wq wr ws wt wu wv ww wx wy wz x_ xa xb xc xd xe xf xg xh xi xj xk xl xm xn xo
yx yy yz z_ za zb zc zd ze zf zg zh zi zj zk zl zm zn zo zp zq zr zs zt zu zv zw zx zy zz

3-grams 0 1 2 3 4 5 6 7 8 9 _ADJ_ _ADP_ _ADV_ _CONJ_ _DET_ _NOUN_ _NUM_ _PRON_ _PRT_ _VERB_ a_ aa ab ac a
bk bl bm bn bo bp bq br bs bt bu bv bw bx by bz c_ ca cb cc cd ce cf cg ch ci cj ck cl cm cn co cp cq cr cs ct cu cv cw cx cy cz
eg eh ei ej ek el em en eo ep eq er es et eu ev ew ex ey ez f_ fa fb fc fd fe ff fg fh fi fj fk fl fm fn fo fp fq fr fs ft fu fv fw fx fy fz g_
hg hh hi hj hk hl hm hn ho hp hq hr hs ht hu hv hw hx hy hz i_ ia ib ic id ie if ig ih ii ij ik il im in io ip iq ir is it iu iv iw ix iy iz j_ ja jt
kp kq kr ks kt ku kv kw kx ky kz l_ la lb lc ld le lf lg lh li lj lk ll lm ln lo lp lq lr ls lt lu lv lw lx ly lz m_ ma mb mc md me mf mg mh r
nm nn no np nq nr ns nt nu nv nw nx ny nz o_ oa ob oc od oe of og oh oi oj ok ol om on oo op oq or os ot other ou ov ow ox oy
qa qb qc qd qe qf qg qh qi qj qk ql qm qn qo qp qq qr qs qt qu qv qw qx qy qz r_ ra rb rc rd re rf rg rh ri rj rk rl rm rn ro rp rq rr rs
tb tc td te tf tg th ti tj tk tl tm tn to tp tq tr ts tt tu tv tw tx ty tz u_ ua ub uc ud ue uf ug uh ui uj uk ul um un uo up uq ur us ut uu uv
wc wd we wf wg wh wi wj wk wl wm wn wo wp wq wr ws wt wu wv ww wx wy wz x_ xa xb xc xd xe xf xg xh xi xj xk xl xm xn xo
yx yy yz z_ za zb zc zd ze zf zg zh zi zj zk zl zm zn zo zp zq zr zs zt zu zv zw zx zy zz

National Library of Scotland
Leabharlann Nàiseanta na h-Alba

https://data.nls.uk/data/digitised-collections/

- Digitized text data of e.g.:
  - Encyclopaedia Britannica, 1768-1860
  - Scottish school exam papers, 1888-1963

Other sources

- Project Gutenberg
  - Has dedicated R package to download text: https://cran.r-project.org/web/packages/gutenbergr/vignettes/intro.html
- Digitized periodicals
  - https://www.pnas.org/content/114/4/E457
- Other English corpora
  - https://www.english-corpora.org/
- Internet Archive
  - https://archive.org/details/datasets
- Wikipedia
  - https://dumps.wikimedia.org/
  - https://en.wikipedia.org/wiki/Wikipedia:Database_download

- Site of important political communication
- Site of important public discussion
- Easily accessible API
- Newly expanded Academic API,
  - See: https://github.com/cjbarrie/academictwitteR

Uses of Twitter data

Twitter as barometer of public opinion, Flores, 2017, *American Journal of Sociology*

Do Anti-Immigrant Laws Shape Public Sentiment? A Study of Arizona's SB 1070 Using Twitter Data[1]

René D. Flores
University of Washington

Scholars have debated whether laws can influence public opinion, but evidence of these "feedback" effects is scant. This article examines the effect of Arizona's 2010 high-profile anti-immigrant law, SB 1070, on both public attitudes and behaviors toward immigrants. Using sentiment analysis and a difference-in-difference approach to analyze more than 250,000 tweets, the author finds that SB 1070 had a negative impact on the average sentiment of tweets regarding immigrants, Mexicans, and Hispanics, but not on those about Asians or blacks. However, these changes in public discourse were not caused by shifting attitudes toward immigrants but by the mobilization of anti-immigrant users and by motivating new users to begin tweeting. While some scholars propose that punitive laws can shape people's attitudes toward targeted groups, this study shows that policies are more likely to influence behaviors. Rather than placating the electorate, anti-immigrant laws may stir the pot further, mobilizing individuals already critical of immigrants.

Uses of Twitter data

Twitter as experimental platform, Munger 2017, *Political Behavior.*

Tweetment Effects on the Tweeted: Experimentally Reducing Racist Harassment

Kevin Munger[1]

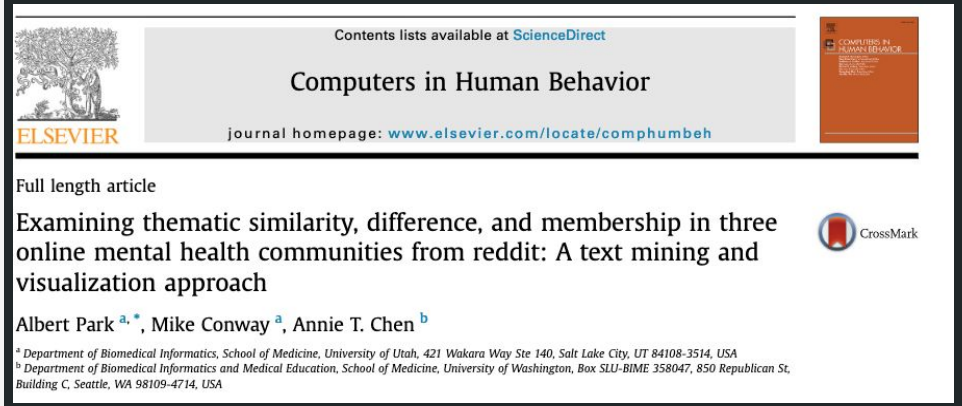
Abstract I conduct an experiment which examines the impact of group norm promotion and social sanctioning on racist online harassment. Racist online harassment de-mobilizes the minorities it targets, and the open, unopposed expression of racism in a public forum can legitimize racist viewpoints and prime ethnocentrism. I employ an intervention designed to reduce the use of anti-black racist slurs by white men on Twitter. I collect a sample of Twitter users who have harassed other users and use accounts I control ("bots") to sanction the harassers. By varying the identity of the bots between in-group (white man) and out-group (black man) and by varying the number of Twitter followers each bot has, I find that subjects who were sanctioned by a high-follower white male significantly reduced their use of a racist slur. This paper extends findings from lab experiments to a naturalistic setting using an objective, behavioral outcome measure and a continuous 2-month data collection period. This represents an advance in the study of prejudiced behavior.

Other social media sources

- Reddit
  - Pushshift a good first port of call to find data: https://redditsearch.io/
- Facebook (URLs dataset)
  - https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/EIAACS
- VKontakte
  - https://github.com/denisStukal/Rvk
- TikTok
  - https://github.com/benjaminguinaudeau/tiktokr
- YouTube
  - https://pypi.org/project/youtube-data-api/
- Many others...

Uses of Reddit data

Reddit for investigation mental health discussions, Park et al., 2018, *Computers in Human Behavior*.



Contents lists available at ScienceDirect

Computers in Human Behavior

journal homepage: www.elsevier.com/locate/comphumbeh

Full length article

Examining thematic similarity, difference, and membership in three online mental health communities from reddit: A text mining and visualization approach

Albert Park [a, *], Mike Conway [a], Annie T. Chen [b]

[a] Department of Biomedical Informatics, School of Medicine, University of Utah, 421 Wakara Way Ste 140, Salt Lake City, UT 84108-3514, USA
[b] Department of Biomedical Informatics and Medical Education, School of Medicine, University of Washington, Box SLU-BIME 358047, 850 Republican St, Building C, Seattle, WA 98109-4714, USA

Uses of Reddit data

Reddit for investigation of community norm generation, Chandrasekharan et al., 2018, *ACM-HCI*.

The Internet's Hidden Rules: An Empirical Study of Reddit Norm Violations at Micro, Meso, and Macro Scales

ESHWAR CHANDRASEKHARAN, Georgia Institute of Technology, USA
MATTIA SAMORY, Virginia Tech, USA
SHAGUN JHAVER, Georgia Institute of Technology, USA
HUNTER CHARVAT, University of Michigan, USA
AMY BRUCKMAN, Georgia Institute of Technology, USA
CLIFF LAMPE, University of Michigan, USA
JACOB EISENSTEIN, Georgia Institute of Technology, USA
ERIC GILBERT, University of Michigan, USA

Norms are central to how online communities are governed. Yet, norms are also emergent, arise from interaction, and can vary significantly between communities—making them challenging to study at scale. In this paper, we study community norms on Reddit in a large-scale, empirical manner. Via 2.8M comments removed by moderators of 100 top subreddits over 10 months, we use both computational and qualitative methods to identify three types of norms: *macro* norms that are universal to most parts of Reddit; *meso* norms that are shared across certain groups of subreddits; and *micro* norms that are specific to individual, relatively unique subreddits. Given the size of Reddit's user base—and the wide range of topics covered by different subreddits—we argue this represents the first large-scale study of norms across disparate online communities. In other words, these findings shed light on what Reddit values, and how widely-held those values are. We conclude by discussing implications for the design of new and existing online communities.

CCS Concepts: • **Human-centered computing** → *Empirical studies in collaborative and social computing*;

Additional Key Words and Phrases: online communities; community norms; moderation; mixed methods.

ACM Reference Format:
Eshwar Chandrasekharan, Mattia Samory, Shagun Jhaver, Hunter Charvat, Amy Bruckman, Cliff Lampe, Jacob Eisenstein, and Eric Gilbert. 2018. The Internet's Hidden Rules: An Empirical Study of Reddit Norm Violations at Micro, Meso, and Macro Scales. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW, Article 32 (November 2018), 25 pages. https://doi.org/10.1145/3274301

Other social media sources

- Reddit
  - Pushshift a good first port of call to find data: https://redditsearch.io/
- Facebook (URLs dataset)
  - https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/EIAACS
- VKontakte
  - https://github.com/denisStukal/Rvk
- TikTok
  - https://github.com/benjaminguinaudeau/tiktokr
- YouTube
  - https://pypi.org/project/youtube-data-api/
- Many others...

Uses of Facebook data

Facebook for investigating how campaigns attract supporters, Bail et al., 2017, *American Sociological Review*.

AMERICAN SOCIOLOGICAL ASSOCIATION

American Sociological Review
2017, Vol. 82(6) 1188–1213

DOI: 10.1177/0003122417733673
journals.sagepub.com/home/asr
SAGE

Channeling Hearts and Minds: Advocacy Organizations, Cognitive-Emotional Currents, and Public Conversation

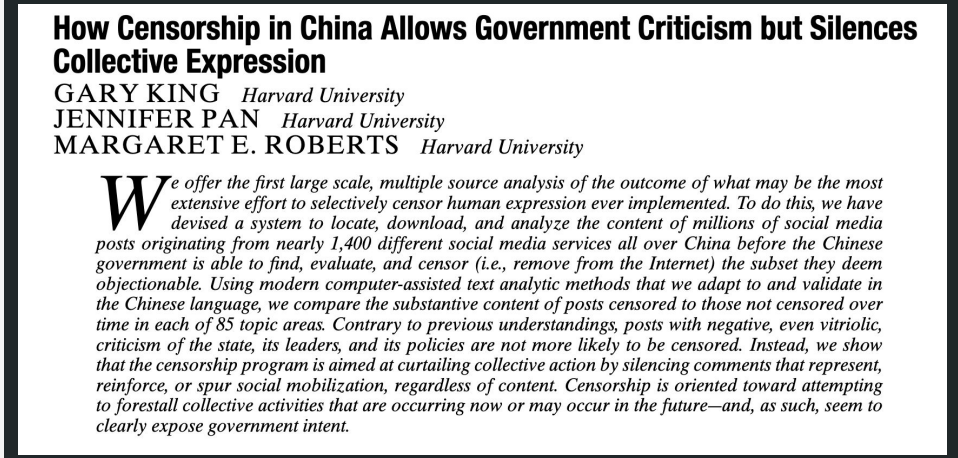
Christopher A. Bail,[a] Taylor W. Brown,[a] and Marcus Mann[a]

Other social media sources

- Reddit
  - Pushshift a good first port of call to find data: https://redditsearch.io/
- Facebook (URLs dataset)
  - https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/EIAACS
- Weibo
  - https://pypi.org/project/weibopy/
- TikTok
  - https://github.com/benjaminguinaudeau/tiktokr
- YouTube
  - https://pypi.org/project/youtube-data-api/
- Many others...

Uses of Weibo data

Weibo for understanding government censorship, King et al., 2013, *American Political Science Review*



How Censorship in China Allows Government Criticism but Silences Collective Expression

GARY KING *Harvard University*
JENNIFER PAN *Harvard University*
MARGARET E. ROBERTS *Harvard University*

We offer the first large scale, multiple source analysis of the outcome of what may be the most extensive effort to selectively censor human expression ever implemented. To do this, we have devised a system to locate, download, and analyze the content of millions of social media posts originating from nearly 1,400 different social media services all over China before the Chinese government is able to find, evaluate, and censor (i.e., remove from the Internet) the subset they deem objectionable. Using modern computer-assisted text analytic methods that we adapt to and validate in the Chinese language, we compare the substantive content of posts censored to those not censored over time in each of 85 topic areas. Contrary to previous understandings, posts with negative, even vitriolic, criticism of the state, its leaders, and its policies are not more likely to be censored. Instead, we show that the censorship program is aimed at curtailing collective action by silencing comments that represent, reinforce, or spur social mobilization, regardless of content. Censorship is oriented toward attempting to forestall collective activities that are occurring now or may occur in the future—and, as such, seem to clearly expose government intent.

How do we get the data?

Two main techniques:

- Scraping
- APIs

As well as other options:

- Private agreement
- Purchase

Terms of service...

Terms of service...

1. Internet Research: Ethical Guidelines 3.0 Association of Internet Researchers, 2019: https://aoir.org/reports/ethics3.pdf
2. University of Oxford Internet-Mediated Research, Central University Research Ethics Committee, 2021: https://researchsupport.admin.ox.ac.uk/files/bpg06internet-basedresearchpdf
3. Internet Research Ethics, Stanford Encylopedia of Philosophy, 2021:https://plato.stanford.edu/entries/ethics-internet-research/

Unit selection and sampling

Unit selection

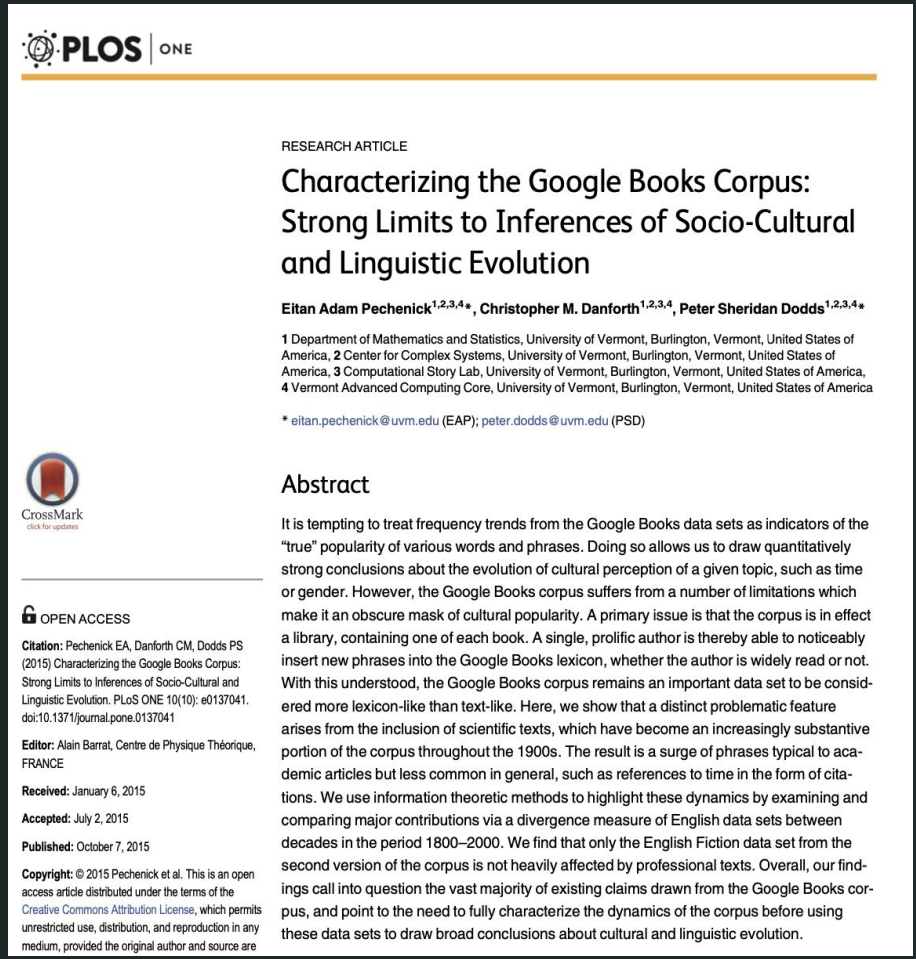
- Who or what do we want to study?
- Over what time frame?

Sampling

- From where is the data coming?
- Are there any biases in the data generating process?

Considerations when using Google Books data

Google Books and sampling biases, Pechenik et al., 2015, *PLoS One*.



PLOS | ONE

RESEARCH ARTICLE

Characterizing the Google Books Corpus: Strong Limits to Inferences of Socio-Cultural and Linguistic Evolution

Eitan Adam Pechenick[1,2,3,4]*, **Christopher M. Danforth**[1,2,3,4], **Peter Sheridan Dodds**[1,2,3,4]*

1 Department of Mathematics and Statistics, University of Vermont, Burlington, Vermont, United States of America, **2** Center for Complex Systems, University of Vermont, Burlington, Vermont, United States of America, **3** Computational Story Lab, University of Vermont, Burlington, Vermont, United States of America, **4** Vermont Advanced Computing Core, University of Vermont, Burlington, Vermont, United States of America

* eitan.pechenick@uvm.edu (EAP); peter.dodds@uvm.edu (PSD)

Abstract

It is tempting to treat frequency trends from the Google Books data sets as indicators of the "true" popularity of various words and phrases. Doing so allows us to draw quantitatively strong conclusions about the evolution of cultural perception of a given topic, such as time or gender. However, the Google Books corpus suffers from a number of limitations which make it an obscure mask of cultural popularity. A primary issue is that the corpus is in effect a library, containing one of each book. A single, prolific author is thereby able to noticeably insert new phrases into the Google Books lexicon, whether the author is widely read or not. With this understood, the Google Books corpus remains an important data set to be considered more lexicon-like than text-like. Here, we show that a distinct problematic feature arises from the inclusion of scientific texts, which have become an increasingly substantive portion of the corpus throughout the 1900s. The result is a surge of phrases typical to academic articles but less common in general, such as references to time in the form of citations. We use information theoretic methods to highlight these dynamics by examining and comparing major contributions via a divergence measure of English data sets between decades in the period 1800–2000. We find that only the English Fiction data set from the second version of the corpus is not heavily affected by professional texts. Overall, our findings call into question the vast majority of existing claims drawn from the Google Books corpus, and point to the need to fully characterize the dynamics of the corpus before using these data sets to draw broad conclusions about cultural and linguistic evolution.

CrossMark click for updates

OPEN ACCESS

Citation: Pechenick EA, Danforth CM, Dodds PS (2015) Characterizing the Google Books Corpus: Strong Limits to Inferences of Socio-Cultural and Linguistic Evolution. PLoS ONE 10(10): e0137041. doi:10.1371/journal.pone.0137041

Editor: Alain Barrat, Centre de Physique Théorique, FRANCE

Received: January 6, 2015

Accepted: July 2, 2015

Published: October 7, 2015

Considerations when using Twitter data

Twitter and issues of representativeness, Mellon and Prosser, 2017, *Research and Politics*.

Research Note

Twitter and Facebook are not representative of the general population: Political attitudes and demographics of British social media users

Research and Politics
July-September 2017: 1–9

DOI: 10.1177/2053168017720008
journals.sagepub.com/home/rap
SAGE

Jonathan Mellon[1] and Christopher Prosser[2]

Abstract

A growing social science literature has used Twitter and Facebook to study political and social phenomena including for election forecasting and tracking political conversations. This research note uses a nationally representative probability sample of the British population to examine how Twitter and Facebook users differ from the general population in terms of demographics, political attitudes and political behaviour. We find that Twitter and Facebook users differ substantially from the general population on many politically relevant dimensions including vote choice, turnout, age, gender, and education. On average social media users are younger and better educated than non-users, and they are more liberal and pay more attention to politics. Despite paying more attention to politics, social media users are less likely to vote than non-users, but they are more likely to support the left leaning Labour Party when they do vote. However, we show that these apparent differences mostly arise due to the demographic composition of social media users. After controlling for age, gender, and education, no statistically significant differences arise between social media users and non-users on political attention, values or political behaviour.

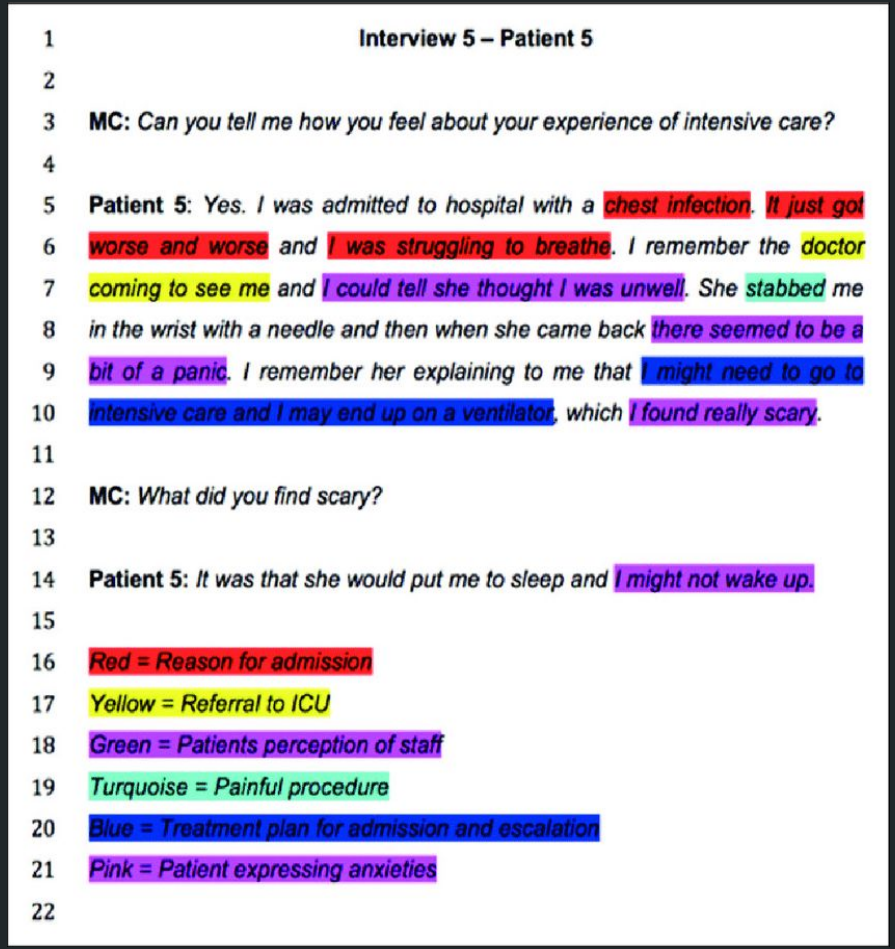
In summary…

- Unit selection:
  - What and where are we measuring?
- Sampling:
  - How is the data being generated?
- Developing the RQ:
  - Given the unit of analysis and way in which sample was generated….
    - What can we ask of these data?

Cleaning and preparing the data

Coding texts

Image taken from: Qualitative Research in Critical Care, Charlesworth and Foëx, 2016, *Journal of the Intensive Care Society*



Interview 5 – Patient 5

MC: *Can you tell me how you feel about your experience of intensive care?*

Patient 5: *Yes. I was admitted to hospital with a chest infection. It just got worse and worse and I was struggling to breathe. I remember the doctor coming to see me and I could tell she thought I was unwell. She stabbed me in the wrist with a needle and then when she came back there seemed to be a bit of a panic. I remember her explaining to me that I might need to go to intensive care and I may end up on a ventilator, which I found really scary.*

MC: *What did you find scary?*

Patient 5: *It was that she would put me to sleep and I might not wake up.*

Red = Reason for admission

Yellow = Referral to ICU

Green = Patients perception of staff

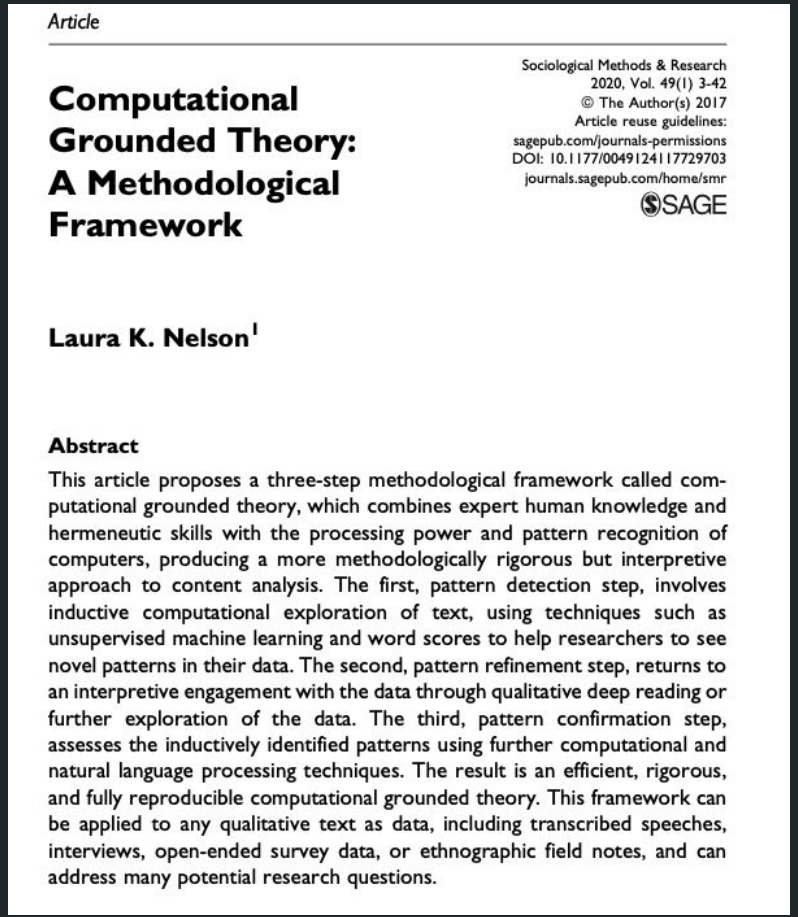
Turquoise = Painful procedure

Blue = Treatment plan for admission and escalation

Pink = Patient expressing anxieties

Combining qualitative and quantitative in analysis of text

Computational Grounded Theory: A Methodological Framework, Nelson, Laura K., 2020, *Sociological Methods & Research*.



Article

Sociological Methods & Research
2020, Vol. 49(1) 3-42

DOI: 10.1177/0049124117729703
journals.sagepub.com/home/smr
SAGE

Computational Grounded Theory: A Methodological Framework

Laura K. Nelson[1]

Abstract

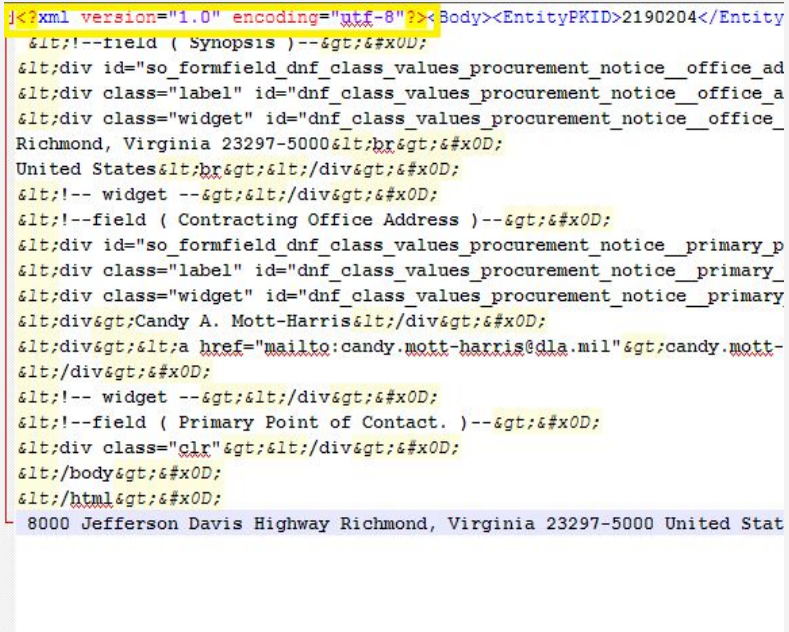
This article proposes a three-step methodological framework called computational grounded theory, which combines expert human knowledge and hermeneutic skills with the processing power and pattern recognition of computers, producing a more methodologically rigorous but interpretive approach to content analysis. The first, pattern detection step, involves inductive computational exploration of text, using techniques such as unsupervised machine learning and word scores to help researchers to see novel patterns in their data. The second, pattern refinement step, returns to an interpretive engagement with the data through qualitative deep reading or further exploration of the data. The third, pattern confirmation step, assesses the inductively identified patterns using further computational and natural language processing techniques. The result is an efficient, rigorous, and fully reproducible computational grounded theory. This framework can be applied to any qualitative text as data, including transcribed speeches, interviews, open-ended survey data, or ethnographic field notes, and can address many potential research questions.

Cleaning the text:

- **UTF-8 encodings**

- **HTML code**

Cleaning the text:



Cleaning the text:

- UTF-8 encodings
- HTML code
- **Preprocessing**

Preprocessing the text:

- Normalization
  - E.g. UKIP ➜ ukip
  - E.g. Ukip ➜ ukip

Preprocessing the text:

- Normalization
  - E.g. UKIP ➜ ukip
  - E.g. Ukip ➜ ukip
- Tokenization
  - “ukip started contesting for power” ➜ “ukip” “started” “contesting” “for” “power”

Preprocessing the text:

- Normalization
  - E.g. UKIP ➜ ukip
  - E.g. Ukip ➜ ukip
- Tokenization
  - “ukip started contesting for power” ➜ “ukip” “started” “contesting” “for” “power”
- Stop word removal
  - “ukip” “started” “contesting” “for” “power” ➜ “ukip” “started” “contesting” “power”

See: https://www.ranks.nl/stopwords for more on stop words

Preprocessing the text:

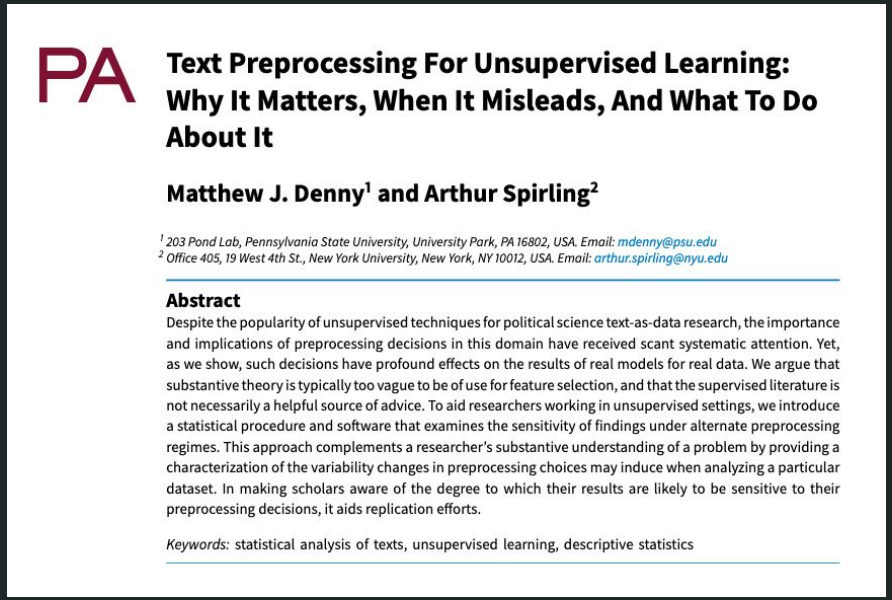
- Normalization
  - E.g. UKIP ➜ ukip
  - E.g. Ukip ➜ ukip
- Tokenization
  - “ukip started contesting for power” ➜ “ukip” “started” “contesting” “for” “power”
- Stop word removal
  - “ukip” “started” “contesting” “for” “power” ➜ “ukip” “started” “contesting” “power”
- Stemming
  - “ukip” “started” “contesting” “power”
  - ➜ “ukip” “started” “contest” “power”

Preprocessing the text:

- Normalization
  - E.g. UKIP ➜ ukip
  - E.g. Ukip ➜ ukip
- Tokenization
  - “ukip started contesting for power” ➜ “ukip” “started” “contesting” “for” “power”
- Stop word removal
  - “ukip” “started” “contesting” “for” “power” ➜ “ukip” “started” “contesting” “power”
- Stemming
  - “ukip” “started” “contesting” “power”
  - ➜ “ukip” “started” “contest” “power”
- Lemmatization
  - “ukip” “started” “contest” “power” ➜ “ukip” “start” “contest” “power”

Text preprocessing and why it matters

Text Preprocessing For Unsupervised Learning: Why It Matters, When It Misleads, And What To Do About It
Denny, Matthew J. and Arthur Spirling, 2018, *Political Analysis*.



PA

Text Preprocessing For Unsupervised Learning: Why It Matters, When It Misleads, And What To Do About It

Matthew J. Denny[1] and Arthur Spirling[2]

[1] *203 Pond Lab, Pennsylvania State University, University Park, PA 16802, USA. Email: mdenny@psu.edu*
[2] *Office 405, 19 West 4th St., New York University, New York, NY 10012, USA. Email: arthur.spirling@nyu.edu*

Abstract

Despite the popularity of unsupervised techniques for political science text-as-data research, the importance and implications of preprocessing decisions in this domain have received scant systematic attention. Yet, as we show, such decisions have profound effects on the results of real models for real data. We argue that substantive theory is typically too vague to be of use for feature selection, and that the supervised literature is not necessarily a helpful source of advice. To aid researchers working in unsupervised settings, we introduce a statistical procedure and software that examines the sensitivity of findings under alternate preprocessing regimes. This approach complements a researcher's substantive understanding of a problem by providing a characterization of the variability changes in preprocessing choices may induce when analyzing a particular dataset. In making scholars aware of the degree to which their results are likely to be sensitive to their preprocessing decisions, it aids replication efforts.

Keywords: statistical analysis of texts, unsupervised learning, descriptive statistics

Hands-on examples

https://github.com/cjbarrie/CTA-Ed

Selecting approaches

- **Identifying trends/themes**
  - **Word counts**
  - **Sentiment analysis/dictionary-based methods**

Selecting approaches

- Identifying trends/themes
  - Word counts
  - Sentiment analysis/dictionary-based methods

- **Computational tools**
  - **Topic models (unsupervised learning)**

Selecting approaches

- Identifying trends/themes
  - Word counts
  - Sentiment analysis/dictionary-based methods
- Computational tools
  - Topic models (unsupervised learning)
- **Advanced computation**
  - **NLP for supervised learning**
  - **Word embedding/word2vec...**
    - **See: https://www.tensorflow.org/tutorials/text/word_embeddings**

Thanks!

christopher.barrie@ed.ac.uk

https://www.cjbarrie.xyz/

@cbarrie