Bringing Big Data to Chinese Politics: Learning From the People's Daily Corpus

Matthew P. Robertson

Australian National University

January 06, 2022

Significance of People's Daily corpus

- Official mouthpiece of the Central Committee of the Chinese Communist Party
- Perhaps most authoritative public source of information about Chinese politics
- Published daily since 1946 - over 2 million articles
- Used by CCP to communicate official policy both to rank-and-file and the public
- Placement, emphasis, and construction of stories and headlines is an indicator of their salience. Given that PD is the voice of the Party, this is an indicator of the significance the Party attaches to topics
- Editor-in-chief is at minister-rank; editorials are often written on instruction of leaders and subject to extensive review

Prior art

- PD corpus has been used to a variety of topics:
  - Foreign policy (Kivimäki 2015)
  - Public health (Dong, Chang, and Chen 2008)
  - Propaganda (G. Wu 1994)
  - Transgender issues and homosexuality (Zhang 2014; Huang 2018)
  - People with disabilities (Ye and Zeldes 2020)
  - Climate change (Wang 2018)
  - Revival of Confucianism (S. Wu 2014)
  - Drug control (Liang and Lu 2013)
  - Military metaphors in describing disease (Yang 2021)
  - Death penalty (Trevaskes, Smith & Robertson, forthcoming)

Two of the biggest uses of the corpus are (1) machine learning project to predict policy shifts, outbreaks, and crackdowns, and (2) a study of the shifting meaning of bureaucratism.

Only those two seem to have featured the full corpus.

Prior art

- Policy Change Index: Chan, Julian Tszkin, and Weifeng Zhong. 2018. "Reading China: Predicting Policy Change with Machine Learning." AEI Economics Working Paper, No. 2018-11

Policy Change Index (PCI)

A Machine Learning Framework to Predict Policy Changes

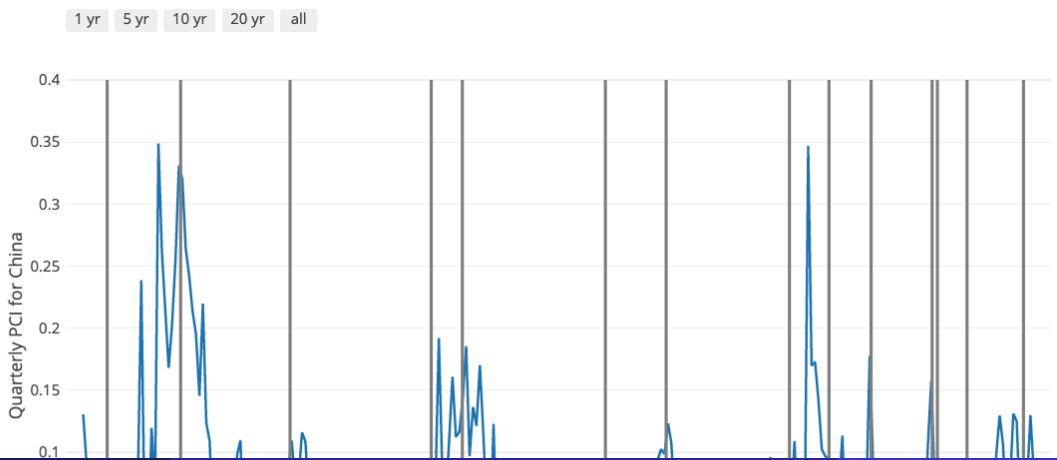
The Policy Change Index (PCI) is a series of open-source machine learning projects that predict authoritarian regimes' major policy moves by "reading" their propaganda publications. The first three projects inducted into the series are about China's policies and based on its official newspaper — the *People's Daily*.

1. **PCI-China**: predicts China's policy changes from 1951 Q1 to the present.
2. **PCI-Crackdown**: predicts how close in time the 2019-20 Hong Kong protests are to a Tiananmen-like crackdown by China.
3. **PCI-Outbreak**: measures the severity of an epidemic outbreak in China, such as COVID-19.

PCI-China | PCI-Crackdown | PCI-Outbreak

PCI-China and major events in China, 1951 Q1 to 2021 Q2

A spike in the PCI-China signals a major policy change, while a vertical bar marks the ground truth of the change labeled by the event. The PCI-China often spikes months before policy changes take place, validating the index's predictive power. Click here to learn more about how it works.



Prior art

- Ding, I., & Thompson-Brusstar, M. (2021). The Anti-Bureaucratic Ghost in China's Bureaucratic Machine. The China Quarterly, 248(S1), 116-140

The China Quarterly

Article contents

Abstract
The Spirit of the Ghost
The Churning of the Machine
Dead Letter or Living Faith?
Conclusion
Footnotes
References

The Anti-Bureaucratic Ghost in China's Bureaucratic Machine

Published online by Cambridge University Press: **10 November 2021**

Iza Ding and Michael Thompson-Brusstar
Show author details

Article | Figures | Metrics

Save PDF | Share | Cite | Rights & Permissions

Abstract

The Chinese Communist Party's (CCP) ideology, rooted in its foundational struggles, explicitly denounces "bureaucratism" (*guanliaozhuyi*) as an intrinsic ailment of bureaucracy. Yet while the revolutionary Party has blasted bureaucratism, its revolutionary regime has had to find a way to coexist with bureaucracy, which is a requisite for effective governance. An anti-bureaucratic ghost thus dwells in the machinery of China's bureaucratic state. We analyse the CCP's anti-bureaucratism through two steps. First, we perform a historical analysis of the Party's anti-bureaucratic ideology, teasing out its substance and emphasizing its roots in and departures from European Marxism and Leninism. Second, we trace both the continuity and evolution in the Party's anti-bureaucratic rhetoric, taking an interactive approach that combines close reading with computational analysis of the entire corpus of the *People's Daily* (1947–2020). We find striking endurance as well as subtle shifts in the substance of the CCP's anti-bureaucratic ideology. We show that bureaucratism is an umbrella term that expresses the revolutionary Party's anxiety about losing its popular legitimacy. Yet the substance of the Party's concern evolved from commandism and revisionism under Mao, to corruption and

This project's contribution

The corpus...

- does not seem to have been used to study elite politics and regime power dynamics
- has not been made widely available in a format that is accessible to other researchers

Primary motivation of project is to demonstrate the utility of the dataset while making it widely accessible for further work.

PD corpus is a rich source of data and other researchers are likely to come up with questions I've never thought of.

Technical challenges in collection

- Recent data (last ~12 months) is publicly available
- Historical data behind subscription/paywall
- Rate limits on access volume
- Browser emulation is slow & RAM heavy
- 2m articles at a theoretical 10k per day = 200 days

But...

- Of course it's all on taobao

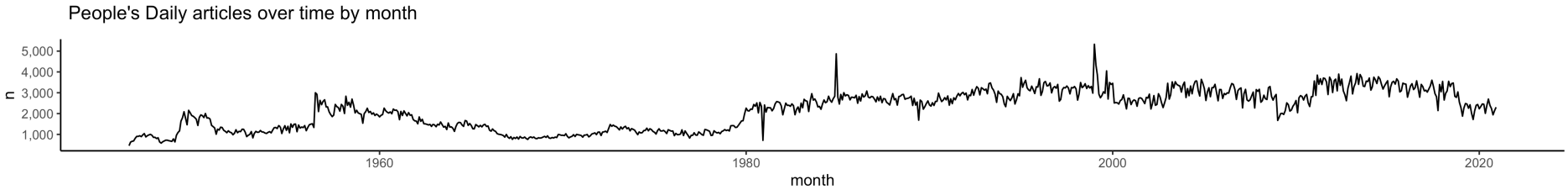
Technical challenges in organization & analysis I

- Some practical challenges to wrangling so much data; flat files don't cut it
- Speed, convenience, robustness: It needs to be put in a database structure
- Went with SQLite; compressed 25gb (2m html files) to 3gb
- Streaming cloud backups with Litestream
- Took days to parse the html and insert the records:
  - parallel processing
  - unfamiliarity with SQLite API (locks, write-ahead log, timeouts, everything)
  - constant crashes, used cron to stack jobs one after another
- Extensive validation required just to know if you did what you wanted to do

Technical challenges in organization & analysis II

- Protip: use data.table when dealing with millions of rows
- Protip 2: databases are good. They allow joins, filters and manipulations out of memory
- Whole corpus yet to be tokenized
- Following are simple dictionary counts. But also normalized by volume of text

Shape of dataset



Comparing Xi to Mao

- Dataset useful for illuminating debates on prominence of Xi Jinping versus Mao Zedong

THE AMERICAN SPECTATOR
EDITED BY R. EMMETT TYRRELL, JR.

SPECIAL REPORT

Xi Jinping Wants to Become the New Mao

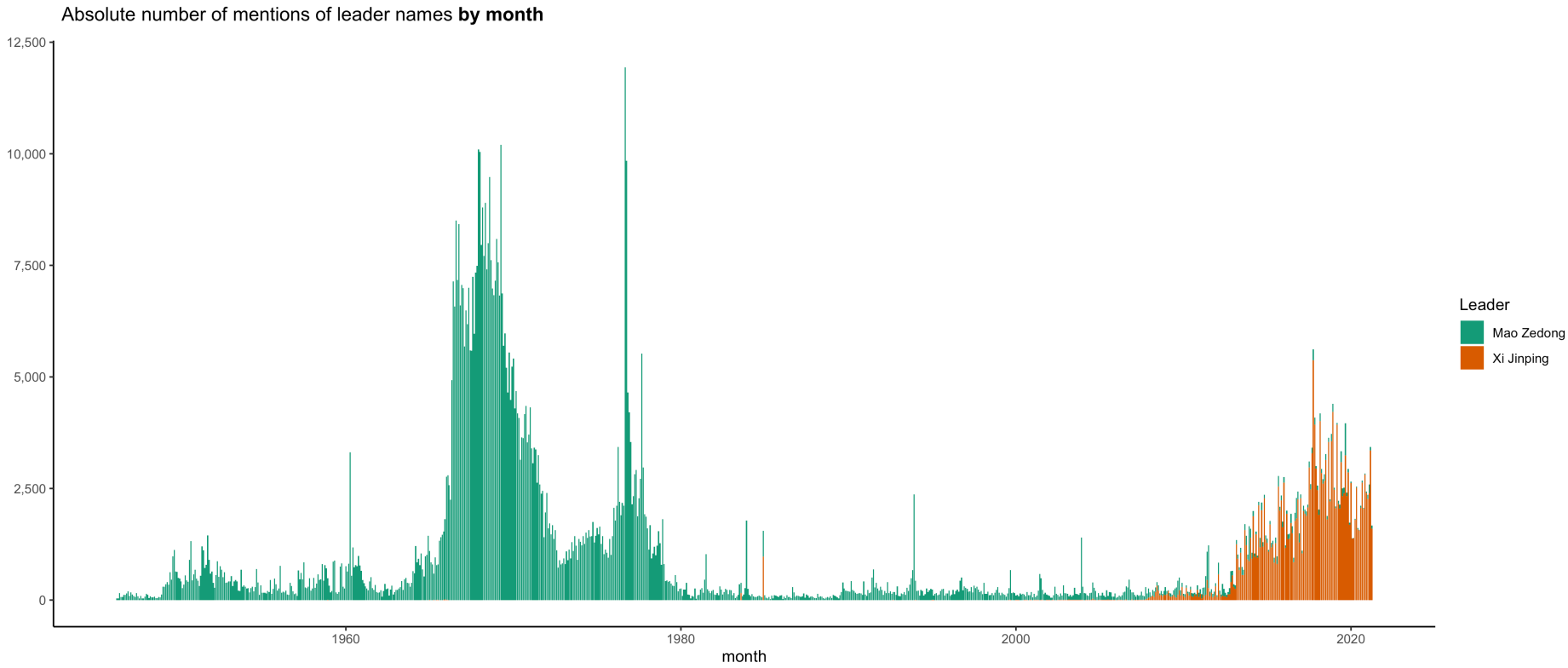
The old Mao almost destroyed China.

by DOUG BANDOW

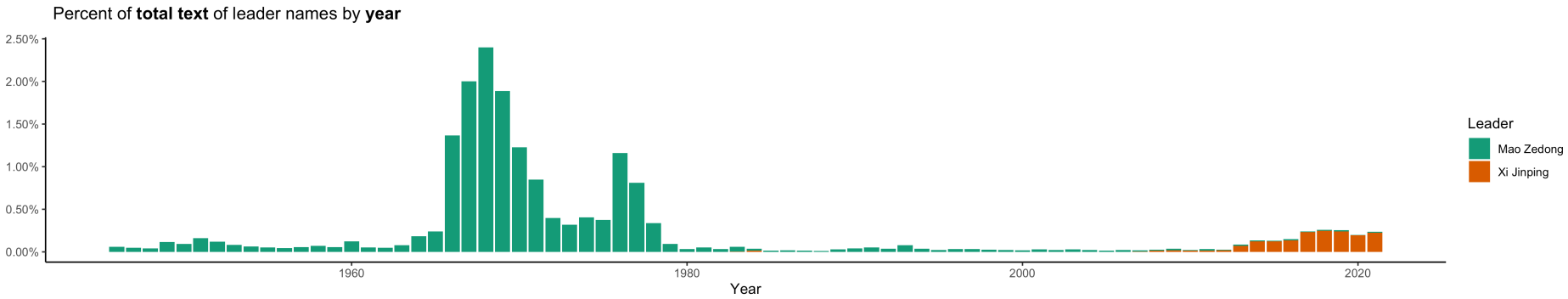
May 26, 2020, 12:58 PM

Comparing Xi to Mao

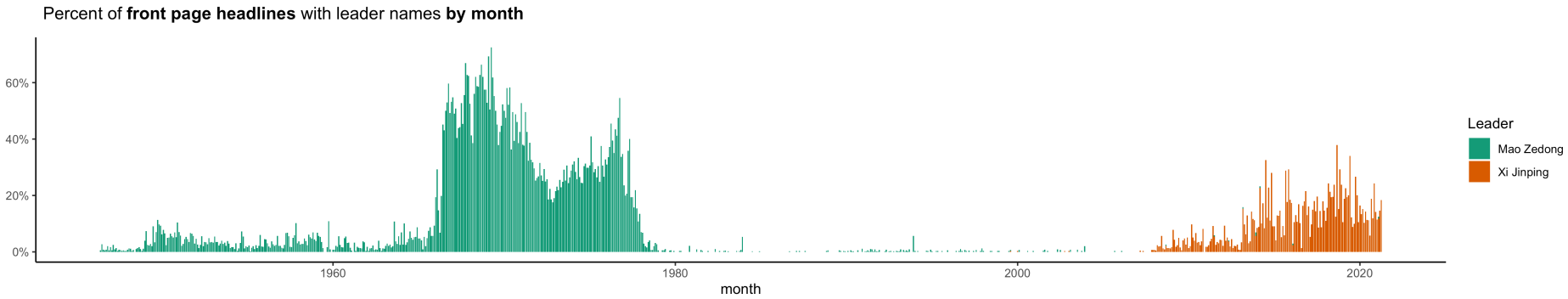
- Is Xi really "the new Mao?"
- Data shows that, at least in People's Daily ink, Mao eclipses Xi; though Xi does eclipse predecessors



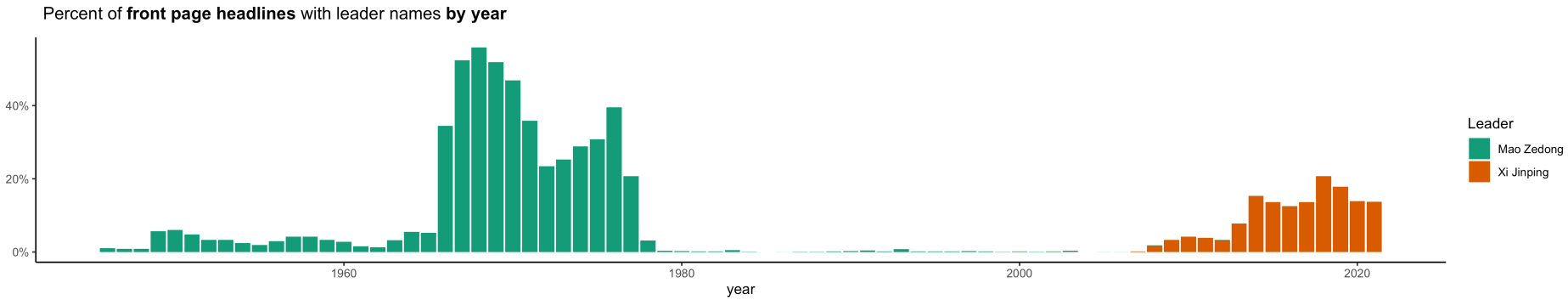
Comparing Xi to Mao



Comparing Xi to Mao



Comparing Xi to Mao



Comparing Xi to Mao

本条数据为：人民日报1966.05.26第1版

把毛主席的书当作我们各项工作的最高指示　——学习工农兵群众对待毛主席著作的态度

《哲学研究》编辑部编

前言

在工农兵群众大学毛主席著作的运动中，不断地涌现出了一批又一批的先进人物。他们坚决听毛主席的话，把毛主席的指示当作自己行动的指南，树立全心全意为人民服务的思想，提高觉悟，改进工作，在阶级斗争、生产斗争和科学实验的革命实践中，取得了一个又一个的胜利。用毛泽东思想武装起来的工农兵群众，是攻无不克、战无不胜的伟大力量。

工农兵群众对待毛主席著作的根本态度，就是带着深厚的无产阶级感情读毛主席的书，就是坚决把毛主席的书当作各项工作的最高指示，真正做到“毛主席怎样说的，我就怎样做”。他们把毛主席的书比作“粮食”、“武器”、“方向盘”和“闹革命的指路明灯”。他们说：“毛主席的话就是最大的政治”，“毛主席的话就是真理”，“千条万条听毛主席的话是第一条”，“千难万难有了毛泽东思想就不难”，“一次又一次地读毛主席的书，就一次又一次地认识到毛主席思想的正确和伟大，一次又一次地加深了对毛主席著作的阶级感情”。他们带着问题学，活学活用，特别在“用”字上狠下工夫，学得好，用得好。

对待毛主席著作采取什么态度，这不是一般的学习态度问题，而是阶级立场问题，是世界观的问题。

毛主席的书，是天才地、全面地、系统地、创造性地发展了的马克思列宁主义，是对中国革命和建设经验的科学总结，是对国际共产主义运动正反两方面经验的科学总结，是被实践反复证明了的伟大的革命的科学真理。毛泽东思想是当代马克思列宁主义的顶峰，是最高最活的马克思列宁主义，是战胜现代修正主义和资产阶级思想的强大武器。是不是听毛主席的话，是不是照着毛主席指引的共产主义方向前进，是不是坚决遵照毛主席的指示办事，是不是把毛主席的书当作各项工作的最高指示，这是革命还是反革命的分水岭，是马克思列宁主义还是修正主义的分水岭。

“读毛主席的书，听毛主席的话，照毛主席的指示办事”，这是关系到巩固无产阶级专政、防止资本主义复辟的大问题，是关系到我们党和国家的命运和前途的大问题，也是关系到世界革命前途的大问题。

把毛主席的书当作各项工作的最高指示，就是坚决服从无产阶级革命的利益，就是坚决服从马克思列宁主义的科学真理。听毛主席的话，就是坚持真理，就是坚持无产阶级革命。毛泽东思想是认清敌我、分辨是非的锐利武器，是辨别真马克思主义还是假马克思主义的准绳。在我们的一切工作中，不论在什么时候和什么情况下，我们都要坚决地按照毛主席的指示办事。凡是毛主席指示的，我们就要坚决拥护，坚决照办；凡是违背毛主席指示的，我们就要坚决抵制，坚决反对。

听毛主席的话，在一切工作中突出政治、以阶级斗争为纲，就要坚决反对对毛主席的话采取阳奉阴违的恶劣态度，就要对一小撮反党反社会主义分子打着“红旗”反红旗，披着马克思列宁主义、毛泽东思想的外衣，反对马克思列宁主义、毛泽东思想，进行反党反社会主义的罪恶勾当，彻底揭露，斗争到底。

工农兵群众活学活用毛主席著作，高举毛泽东思想的伟大红旗，积极参加社会主义文化大革命。他们不仅是阶级斗争、生产斗争、科学实验三大革命运动的主力军，而且也是社会主义文化大革命的主力军。他们活学活用毛主席著作，为我们树立了光辉的榜样。

知识分子、学术工作者必须认真地向工农兵学习，深入到工农兵群众中去，同他们一起生活、一起斗争，改造自己的思想感情，真正站到无产阶级的立场上来，真正树立无产阶级世界观，学习工农兵群众对待毛主席著作的态度和学习方法。只有这样，才能把毛泽东思想学到手，在社会主义革命和社会主义建设中发挥自己的作用。

现在，我们从报刊上摘录了工农兵群众谈学习毛主席著作的态度和方法的一部分材料，按问题加以整理，发表在这里，供大家学习和研究。在材料的编选上，可能有不少缺点，希望读者们批评指正。

一、毛主席的书是我们各项工作的最高指示

编者按：工农兵群众怀着浓厚的无产阶级感情，把毛主席的书当作各项工作的最高指示，这就是他们对待毛主席著作的根本态度。他们说：

“毛泽东思想是我们的命根子”。

“毛主席著作好比‘粮食’，‘武器’、‘方向盘’”。

“毛主席的话就是真理，毛泽东思想是我国人民革命的指针”。

“千条万条听毛主席的话是第一条”。

“千难万难有了毛泽东思想就不难”。

……

毛泽东思想是我们战无不胜的伟大旗帜，这是中国几十年的革命经验证明了的。工农兵群众这些生动的、形象的语言，是他们对我国几十年来革命经验的概括，也是他们在当前三大革命斗争的实践中，活学活用毛主席著作的切身体会。

毛泽东思想是我们的命根子

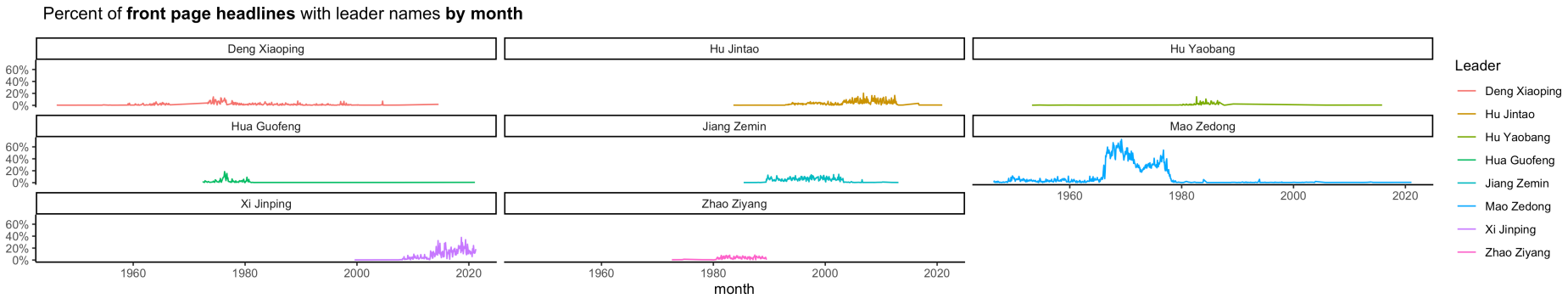
毛泽东思想是当代马克思列宁主义的顶峰，是最高最活的马克思列宁主义，是劳动人民翻身求解放的武器，是我们革命和建设的指针，是反对帝国主义和现代修正主义和各国反动派最锐利的思想武器，是世界人民革命的灯塔，也是抚育我们青年一代健康成长的阳光雨露。

我这个苦孩子，就是在毛泽东思想哺育下成长的。在旧社会，我家没有地没有房，爸爸从小就给地主干活，终年劳动，仍是两手空空，维持不了一家的生活。结果，还换来一身重病，临死时，连一口白米稀饭也没有吃上。我们母子只好东跑西撞讨饭为生。到了晚上，妈妈把我搂在怀里，冰冷的眼泪滴在我的脸上。就这样盼星星盼月亮，终于盼来了救星共产党和伟大领袖毛主席，把我从火海中拯救出来。我有了吃有了穿，党还把我送进学校念书，使我由一个穷娃娃变成了中学生。一九六三年，我光荣地当上了人民解放军战士，进一步受到了毛泽东思想的熏陶和抚育，使我不断地提高了无产阶级觉悟，连续三年评为五好战士，入了团，入了党。

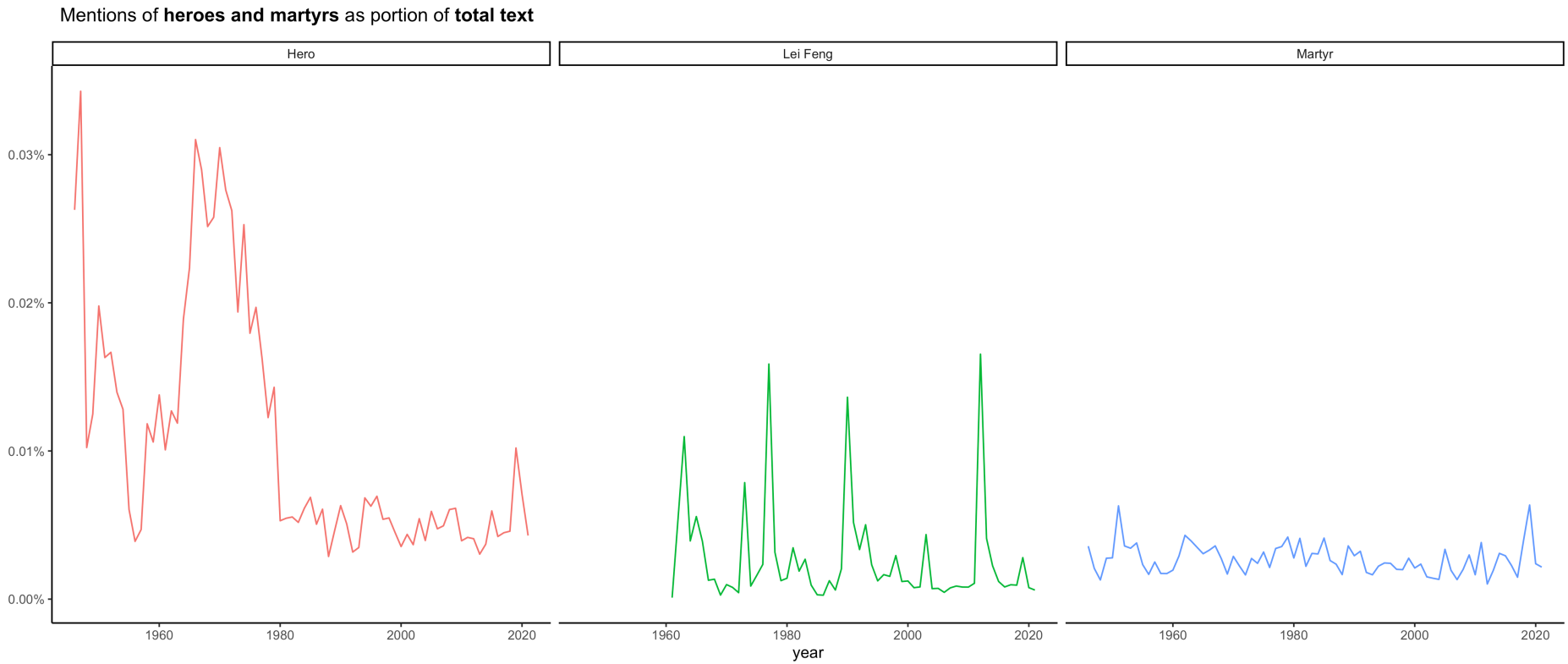
这些经历，使我从内心里深刻地感到，共产党和毛主席是我们穷人的救命恩人，毛主席的书，是为我们穷人谋翻身求解放的，毛泽东思想是我们的命根子。我一定要做到用毛泽东思想统帅一切，推动一切，改造一切，干一辈子革命，学一辈子毛主席著作，一辈子按毛主席指示办事。我要加紧努力读毛主席的书，听毛主席的话，按毛主席的指示办事，做毛主席的好战士。

（解放军某部班长廖双全）

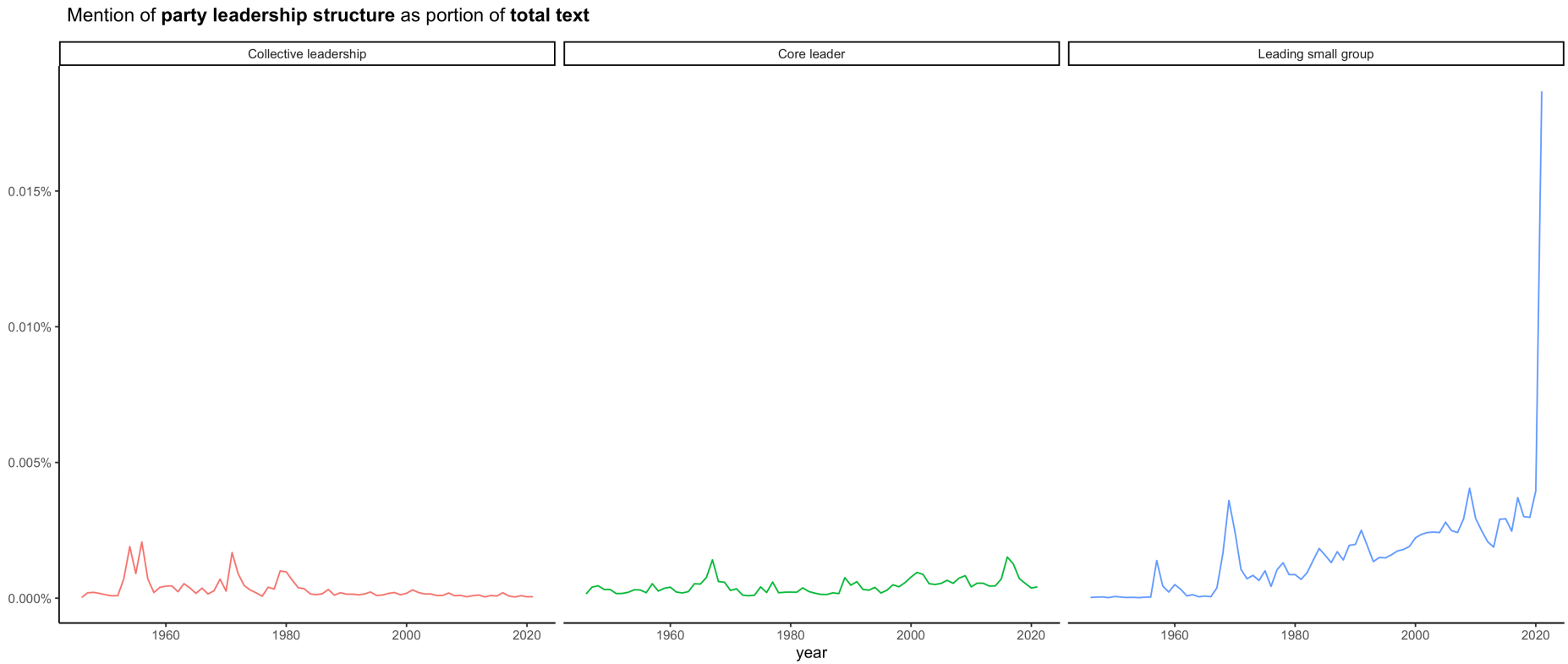
Comparing leaders



Thematic analyses: heroism, martyrdom, collective leadership I



Thematic analyses: heroism, martyrdom, collective leadership II



Purges

- Nope.
- Data insufficient to say anything useful.

To do

- Tokenize text (jieba/spacy)
- Productionize, update
- More analysis to prove out value of the corpus
- Word embeddings?
- Figure out safe way of making dataset accessible to other researchers
  - maybe just the document feature matrix public?
  - put the SQLite files on a cloud bucket and share the key per request?
  - copyright issues at play

Publication plans

- Not ambitious: Simple 'dataset' paper
- Two other exploratory analyses to prove out utility of the data:
  - Policy priorities: Topic models of front page articles (by decade?)
  - Foreign policy: sentiment analysis of discussion of other countries
- Potentially useful dataset as one building block in a large scale comparative analysis of policy in autocratic regimes

References I

Dong, Dong, Tsan-Kuo Chang, and Dan Chen. 2008. "Reporting AIDS and the Invisible Victims in China: Official Knowledge as News in the People's Daily, 1986–2002." Journal of Health Communication 13 (4): 357–74. https://doi.org/10.1080/10810730802063793.

Huang, Yixiong. 2018. "Media Representation of Tongxinglian in China: A Case Study of the People's Daily." Journal of Homosexuality 65 (3): 338–60. https://doi.org/10.1080/00918369.2017.1317475.

Kivimäki, Timo. 2015. "People's Daily and the Reality of the South China Sea Territorial Disputes." Asian Politics & Policy 7 (2): 319–23. https://doi.org/10.1111/aspp.12183.

Liang, Bin, and Hong Lu. 2013. "Discourses of Drug Problems and Drug Control in China: Reports in the People's Daily, 1946–2009." China Information 27 (3): 301–26. https://doi.org/10.1177/0920203x13491387.

References II

Wang, Sidan. 2018. "Dynamic Constructed Climate Change Discourses and Discourse Networks Across Newspapers in China Around Three Critical Policy Moments: A Comparative Study of People's Daily, China Daily, and Southern Weekend." PhD thesis, University of Exeter. https://ore.exeter.ac.uk/repository/handle/10871/33375.

Wu, Guoguang. 1994. "Command Communication: The Politics of Editorial Formulation in the People's Daily." The China Quarterly, no. 137: 194–211. http://www.jstor.org/stable/655694.

Wu, Shufang. 2014. "The Revival of Confucianism and the CCP's Struggle for Cultural Leadership: A Content Analysis of the People's Daily, 2000–2009." Journal of Contemporary China 23 (89): 971–91. https://doi.org/10.1080/10670564.2014.882624.

Yang, Zheng. 2021. "Military Metaphors in Contemporary Chinese Disease Coverage: A Case Study of the People's Daily, 1946–2019." Chinese Journal of Communication 14 (3): 259–77. https://doi.org/10.1080/17544750.2020.1818593.

References III

Ye, Wen, and Geri Alumit Zeldes. 2020. "The Representation of People with Disabilities in an Official Newspaper in China: A Longitudinal Study of the People's Daily from 2003 to 2013." Journal of Disability Policy Studies 31 (1): 26–34. https://doi.org/10.1177/1044207319868783.

Zhang, Qing Fei. 2014. "Transgender Representation by the People's Daily Since 1949." Sexuality & Culture 18 (1): 180–95. https://doi.org/10.1007/s12119-013-9184-3.