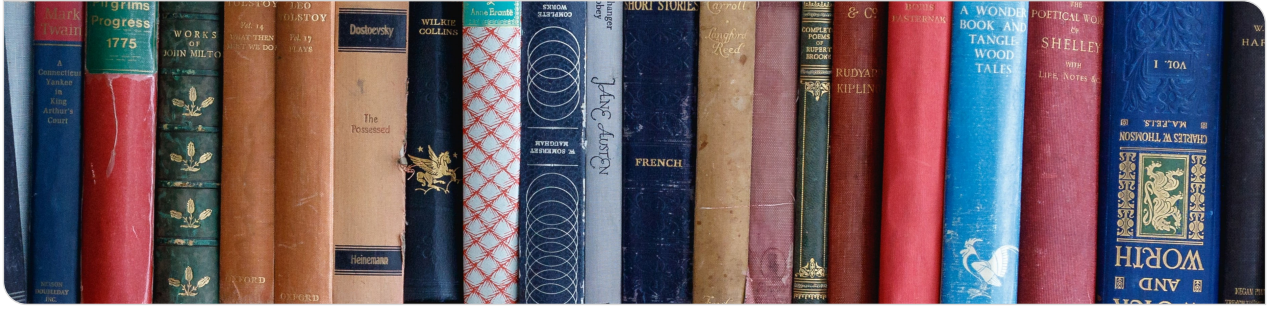
Karlsruhe Institute of Technology

ACM/IEEE JCDL 2024

A Workflow for Efficient and Interactive Analysis of the Google Books Ngram Corpus

Session #15 at The 2024 ACM/IEEE Joint Conference on Digital Libraries

Fabian Richter, Klemens Böhm | December 19, 2024



Introduction

Google scanned large amounts of books and made n-gram counts publicly available:

- Spanning from ~1500 to 2019.
- German 1-gram data: 7.3 GiB – 40,161,066 entries.

Roughly $4 \cdot 10^7$ entries of length $5 \cdot 10^2$:

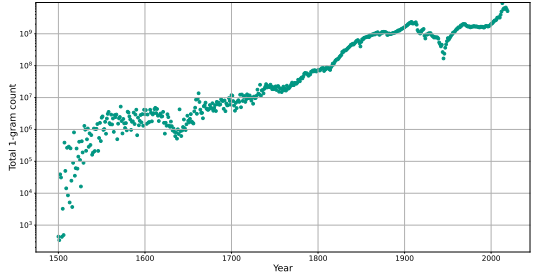
$\approx 2 \cdot 10^{10}$ data points.

But:

Lots of noise and redundancy,

almost 90% of entries are 0.

Still: Large amount of data.



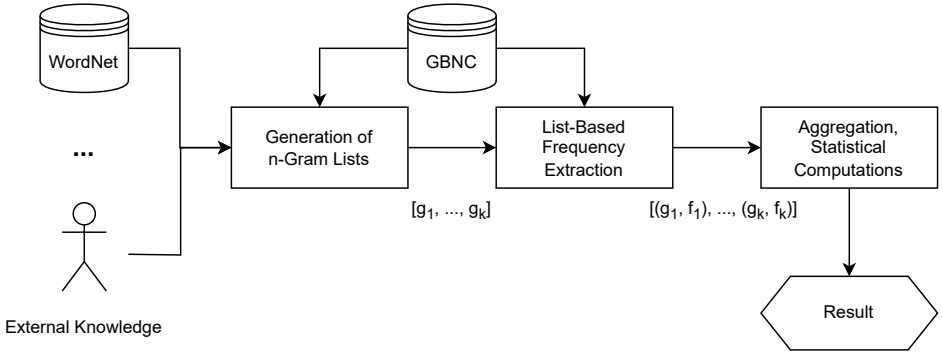
Introduction (continued)

	1-grams		2-grams	3-grams	4-grams	5-grams
	#	size	size	size	size	size
Chinese	508,674	58.0 MiB	2.5 GiB	22.7 GiB	16.8 GiB	35.9 GiB
English	**79,080,571**	**12.5 GiB**	**301.6 GiB**	**3.1 TiB**	**2.7 TiB**	**7.4 TiB**
English Fiction	4,608,325	896.5 MiB	29.2 GiB	281.9 GiB	240.9 GiB	613.8 GiB
English (UK)	18,947,489	3.5 GiB	89.0 GiB	824.0 GiB	685.0 GiB	1.6 TiB
English (US)	49,924,752	7.7 GiB	193.5 GiB	1.9 TiB	1.7 TiB	4.4 TiB
French	23,719,351	4.9 GiB	92.0 GiB	807.0 GiB	673.9 GiB	1.7 TiB
German	40,161,066	7.3 GiB	135.3 GiB	902.2 GiB	628.0 GiB	1.3 TiB
Hebrew	2,809,535	647.4 MiB	8.5 GiB	34.2 GiB	19.2 GiB	29.8 GiB
Italian	12,443,487	2.8 GiB	58.5 GiB	436.8 GiB	310.6 GiB	651.8 GiB
Russian	13,109,370	2.1 GiB	46.5 GiB	275.4 GiB	175.1 GiB	331.7 GiB
Spanish	14,151,756	3.0 GiB	57.8 GiB	451.4 GiB	345.4 GiB	786.8 GiB

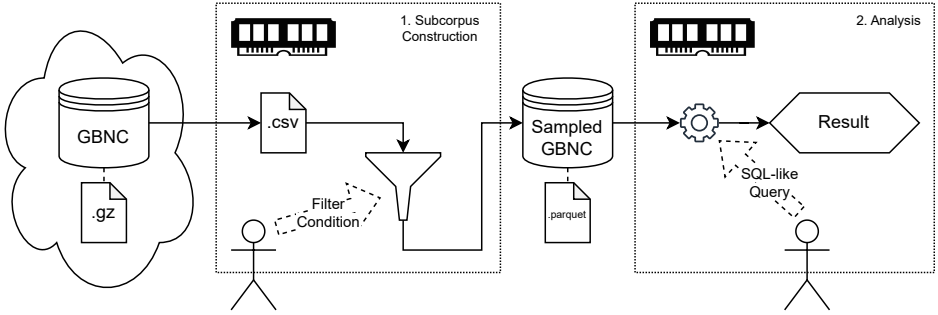
Motivation

		List Generation	Frequency Extraction	Analysis	# of n-grams
medicine	Jurić 2022	✓	✓	simple arithmetic operations	< 50
	Menadue 2020	✓	✓	time series similarity	5
	Teepe, Glase, and Reips 2023	✓	✓	correlation between topics	354 + inflections
philosophy	Haslam and Ye 2019	✓	✓	sum and mean	108
	P. Kesebir and S. Kesebir 2012	✓	✓	peak, nadir, change	60
	Scheffer et al. 2021	✓	✓	principal component analysis, sum	10,000
	Wheeler, McGrath, and Haslam 2019	✓	✓	descriptive statistics, curve fitting	304
social sciences	Caruana-Galizia 2016	✓	✓	correlation with historic events	6
	Madsen and Slåtten 2022	✓	–	–	10
	Willems 2013	✓	–	–	50
	Younes and Reips 2019	✓	✓	correlation with time	60 + inflections

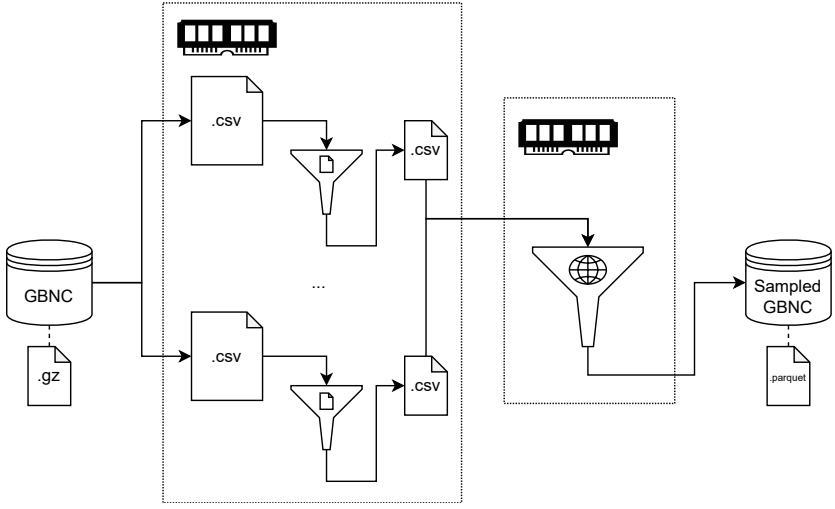
Motivation (continued)



Workflow


1. Subcorpus Construction
2. Analysis
GBNC
.gz
.csv
Filter Condition
Sampled GBNC
.parquet
Result
SQL-like Query

Workflow (continued)



Common Filter Conditions

Dictionary-based

Given a list (or dictionary) of n-grams, extract their specific frequencies.

Top-k

Given k, find the k most frequent n-grams in the corpus (usually for one fixed n.

These two conditions cover all the cases we have observed in our literature study.

These filter conditions have two desirable properties:

- they can be evaluated *locally*, with no information exchange between batches; and
- at least in the observed cases, they reduce the corpus size significantly: *one filtering pass* is enough.

Case Study

Haslam and Ye 2019

- Objective: popularity of *psychoanalysis* terms in English and French
- Download size: 13.1 GiB
- Time required: 36 minutes
- Resulting file size: <100 KiB

Scheffer et al. 2021

- Objective: development of most frequent words in English and Spanish
- Download size: 15.5 GiB
- Processing time: 42 minutes
- Resulting file size: <10 MiB

Takeaways

- Information needs are different, but workflows are similar.
- Time requirement is heavily dominated by download times, filtering and analysis are (almost) negligible.
- **Analyzing the GBNC is not (necessarily) a Big Data problem**!

Conclusion

Our Contributions

We have presented

- a categorization of information needs from literature;
- a description of a workflow to identify, retrieve and analyze interesting subsets of the GBNC;
- a case study to highlight our workflow's ability to reproduce existing research results.

Future Work

In the future, we plan to

- apply our workflow to novel research questions, enabling data-driven knowledge discovery; and
- provide a full implementation of the workflow as a *Python* package.

Please contact me at **fabian.richter@kit.edu** if you have any questions or remarks!

References I

- Caruana-Galizia, Paul (2016). “Politics and the German language: Testing Orwell’s hypothesis using the Google N-Gram corpus”. In: *Digital Scholarship in the Humanities* 31.3, pp. 441–456.
- Haslam, Nick and Lotus Ye (2019). “Freudian slip? The changing cultural fortunes of psychoanalytic concepts”. In: *Frontiers in Psychology* 10, p. 468468.
- Jurić, Tado (2022). “Using digital humanities for understanding COVID-19: lessons from digital history about earlier coronavirus pandemic”. In: *medRxiv*, pp. 2022–02.
- Kesebir, Pelin and Selin Kesebir (2012). “The cultural salience of moral character and virtue declined in twentieth century America”. In: *The Journal of Positive Psychology* 7.6, pp. 471–480.
- Madsen, Dag Øivind and Kåre Slåtten (2022). “The possibilities and limitations of using Google Books Ngram Viewer in research on management fashions”. In: *Societies* 12.6, p. 171.
- Menadue, Christopher B. (2020). “Pandemics, epidemics, viruses, plagues, and disease: comparative frequency analysis of a cultural pathology reflected in science fiction magazines from 1926 to 2015”. In: *Social Sciences & Humanities Open* 2.1, p. 100048.

References II

- Scheffer, Marten et al. (2021). “The rise and fall of rationality in language”. In: *Proceedings of the National Academy of Sciences* 118.51, e2107848118.
- Teepe, Gisbert Wilhelm, Edda Magareta Glase, and Ulf-Dietrich Reips (2023). “Increasing digitalization is associated with anxiety and depression: A Google Ngram analysis”. In: *PLOS One* 18.4, e0284091.
- Wheeler, Melissa A., Melanie J. McGrath, and Nick Haslam (2019). “Twentieth century morality: The rise and fall of moral concepts from 1900 to 2007”. In: *PLOS One* 14.2, e0212267.
- Willems, Klaas (2013). “‘Culturomics’ and the representation of the language of the Third Reich in digitized German books”. In: *Interdisciplinary Journal for Germanic Linguistics and Semiotic Analysis* 18.1, pp. 87–99.
- Younes, Nadja and Ulf-Dietrich Reips (2019). “Guideline for improving the reliability of Google Ngram studies: Evidence from religious terms”. In: *PLOS One* 14.3.