NATURAL LANGUAGE PROCESSING

Minority Voices 'Filtered' Out of Google Natural Language Processing Models

The Efforts to Make Text-Based AI Less Racist and Terrible

Language models like GPT-3 can write poetry, but they often amplify negative stereotypes. Researchers are trying different approaches to address the problem.

Minority Voices 'Filtered' Out of Google Natural Language Processing Models - Unite.AI, The Efforts to Make Text-Based AI Less Racist and Terrible | WIRED

Dissecting LLMs: Data

COS597G: Understanding Large Language Models, Fall 2022

Tanushree Banerjee, Andre Niyongabo Rubungo

Roadmap

Main paper: Documenting Large Webtext Corpora

- Motivation
- Three levels of documentation
- Recommendations and discussion

The Pile dataset

Deduplication

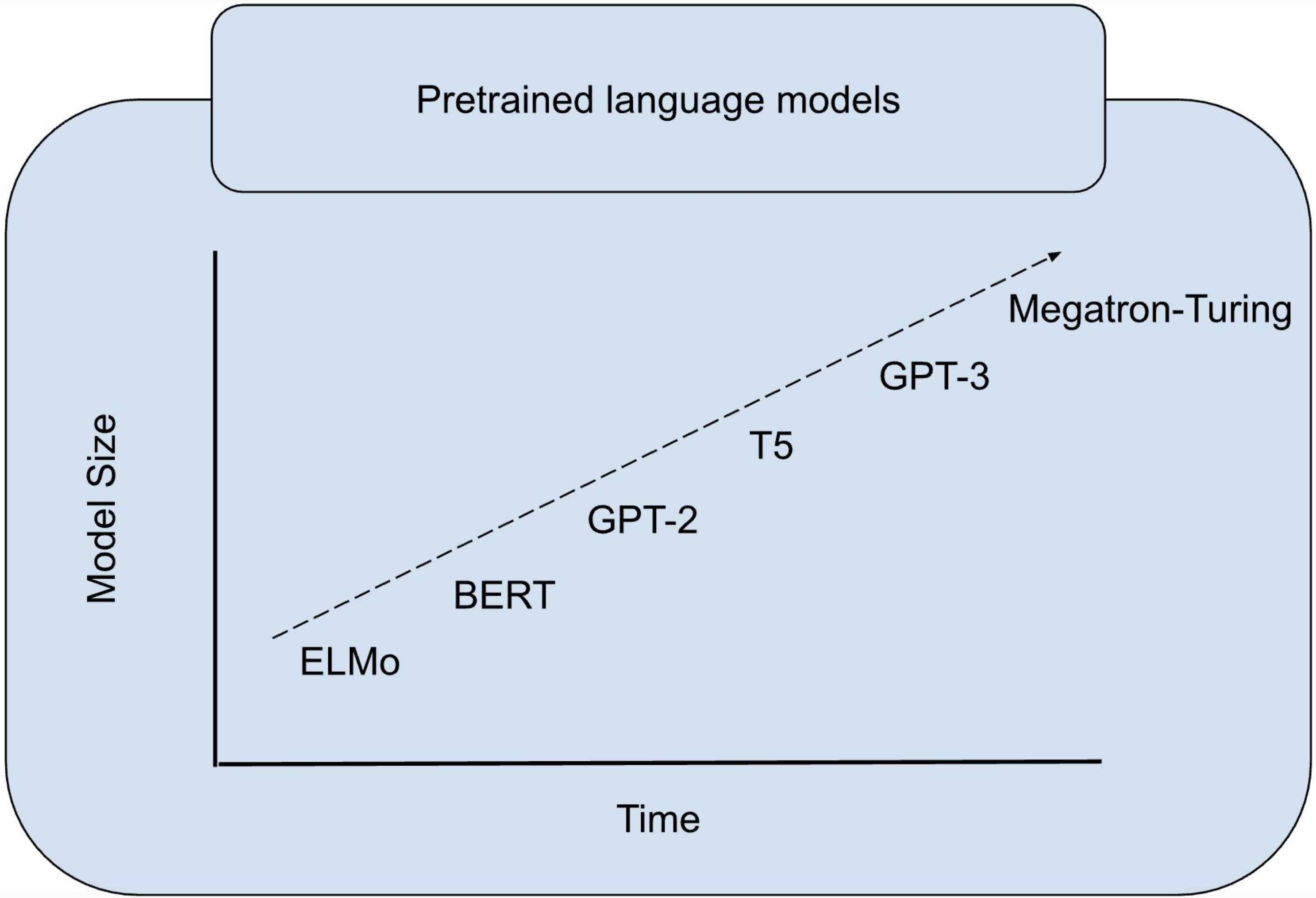
Summary and key takeaways

Documenting Large Webtext Corpora: A Case Study on the Colossal Clean Crawled Corpus

Jesse Dodge Maarten Sap Ana Marasovic William Agnew
Gabriel Ilharco Dirk Groeneveld Margaret Mitchell Matt Gardner

EMNLP 2021
Online & in the Dominican Republic

Some slides adapted from Jesse Dodge's talk @ EMNLP 2021


Pretrained language models
Model Size
Megatron-Turing
GPT-3
T5
GPT-2
BERT
ELMo
Time

2.5×10^9

10^{10}

10^{11}

What data is used for pretraining?

BERT
2018

GPT-2
2019

GPT-3
2020

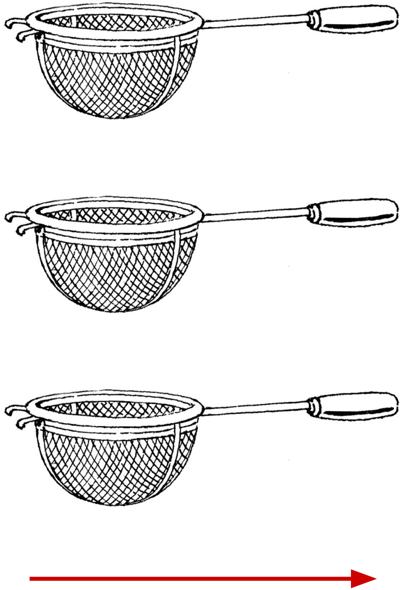
Web-scale text

100 petabytes
Google search index

2.5 petabytes an hour
Data generated by Walmart

320 terabytes
April 2021 snapshot of Common Crawl

Web-scale text



Colossal Clean Crawled Corpus (C4)

Created to train T5 (Raffel et al. 2019)

April 2019 snapshot of Common Crawl (1.4 trillion tokens)

Remove lines without terminal punctuation mark, code ("{"), < 3 words

Remove docs with "lorem ipsum", < 5 sentences, "bad words"

Used langdetect to filter out non-English text

Result: 806 GB of text (156 billion tokens)

Answer for Q1 from pre-lecture questions!

1. **Steps involved in collecting + pre-processing C4.EN**
2. **External programs + resources required**

April 2019 snapshot of Common Crawl (1.4 trillion tokens)

Remove lines without terminal punctuation mark, code (“{“), < 3 words

Remove docs with “lorem ipsum”, < 5 sentences, blocklist words

Used langdetect to filter out non-English text

Result: 806 GB of text (156 billion tokens)

Dataset	# documents	# tokens	size
C4.EN.NOCLEAN	1.1 billion	1.4 trillion	2.3 TB
C4.EN.NOBLOCKLIST	395 million	198 billion	380 GB
C4.EN	365 million	156 billion	305 GB

Introduced by Raffel et. al., 2020

Need for documentation

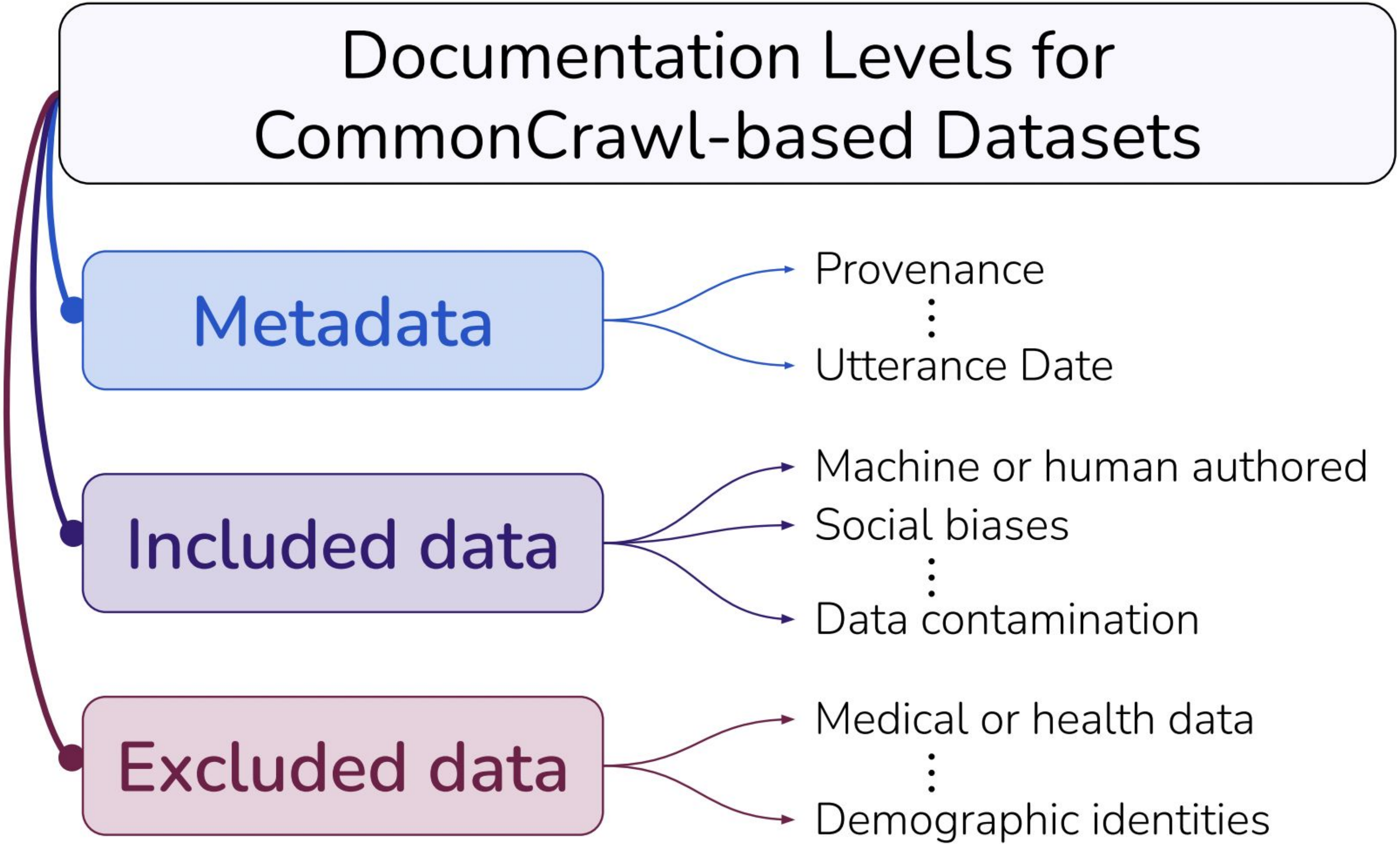
Task specific NLP datasets have best practices around documentation

Applying these practices to massive datasets of unlabelled text is a challenge

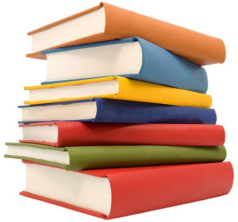
Required information not available in web-crawled text

Thorough documentation typically not done

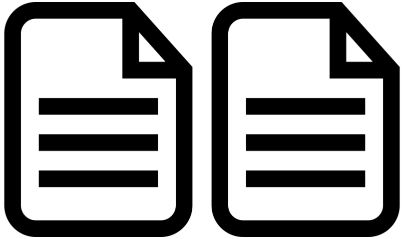
Leaves consumers of pretrained LMs in the dark about the influences of pretraining data on their systems


Documentation Levels for CommonCrawl-based Datasets
Metadata
Provenance
⋮
Utterance Date
Included data
Machine or human authored
Social biases
⋮
Data contamination
Excluded data
Medical or health data
⋮
Demographic identities

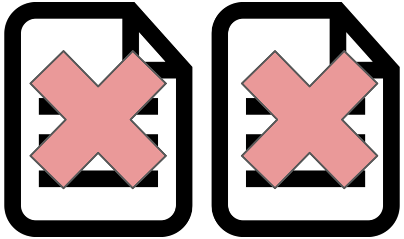
Documentation levels



Metadata



Included data



Excluded data

Corpus level statistics: Metadata

Internet domains and websites

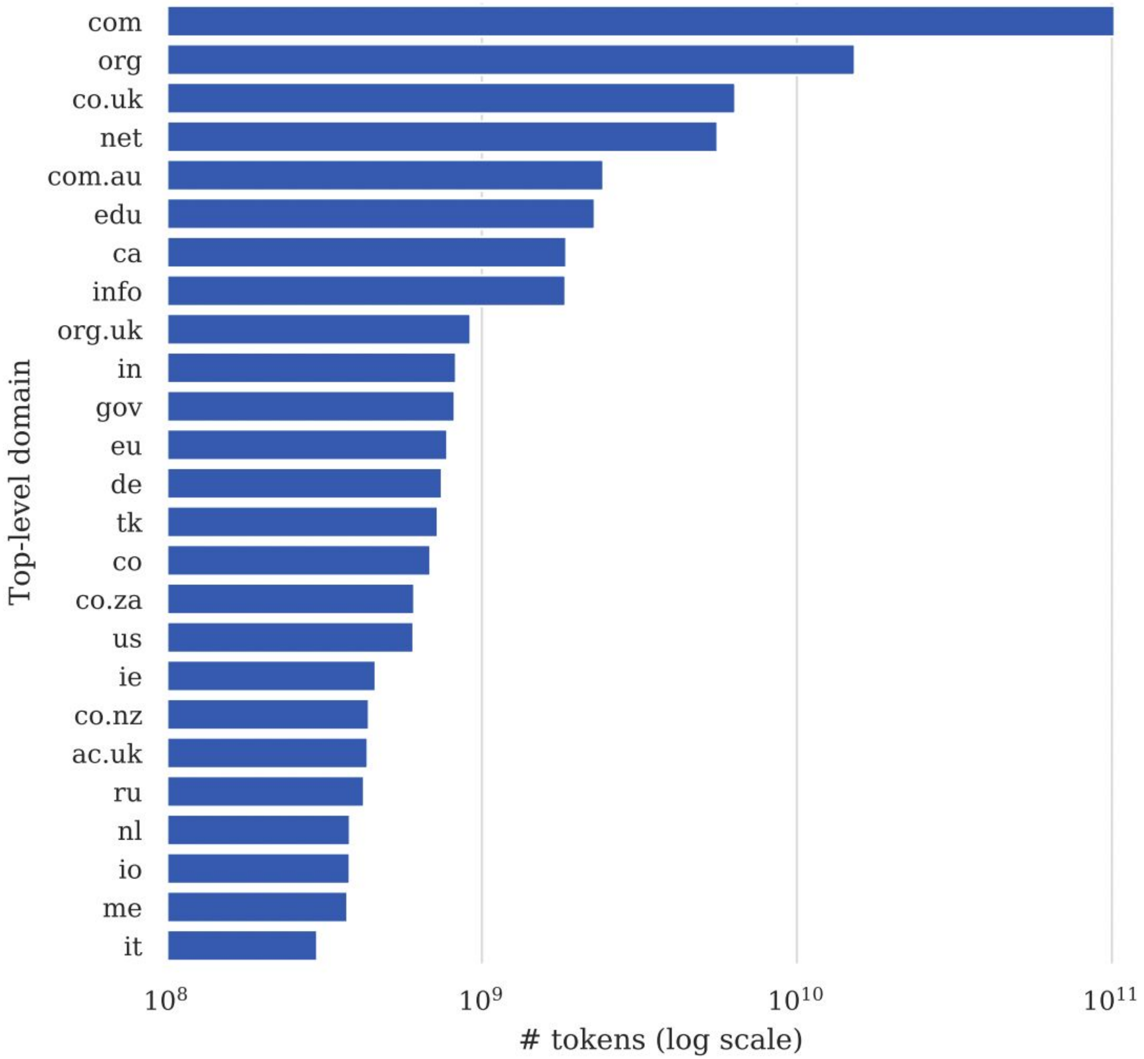
Utterance date

Geolocation

Metadata:
Internet domains

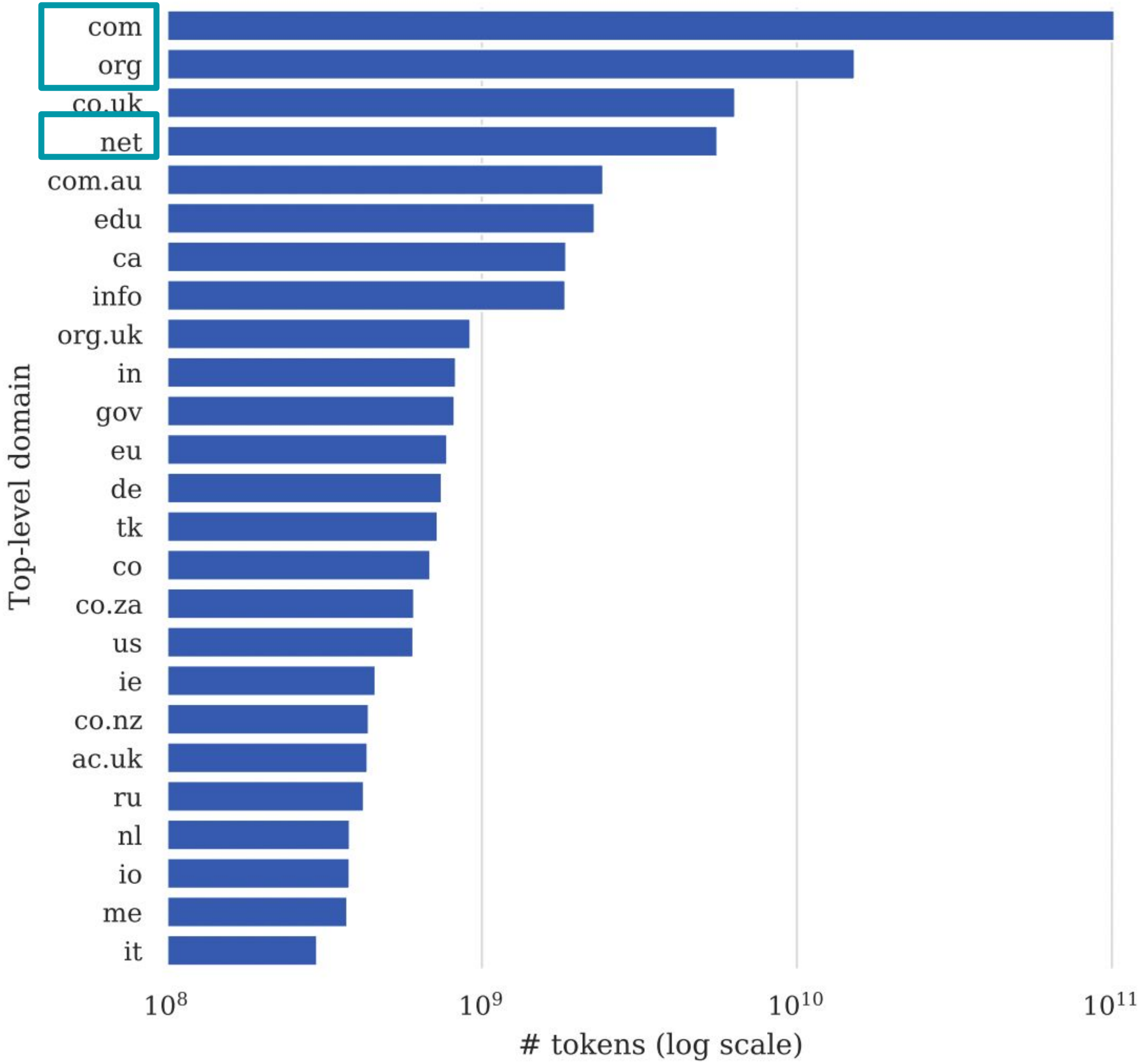
Top-level domains by number of tokens

Tokens measured based on SpaCy tokenizer



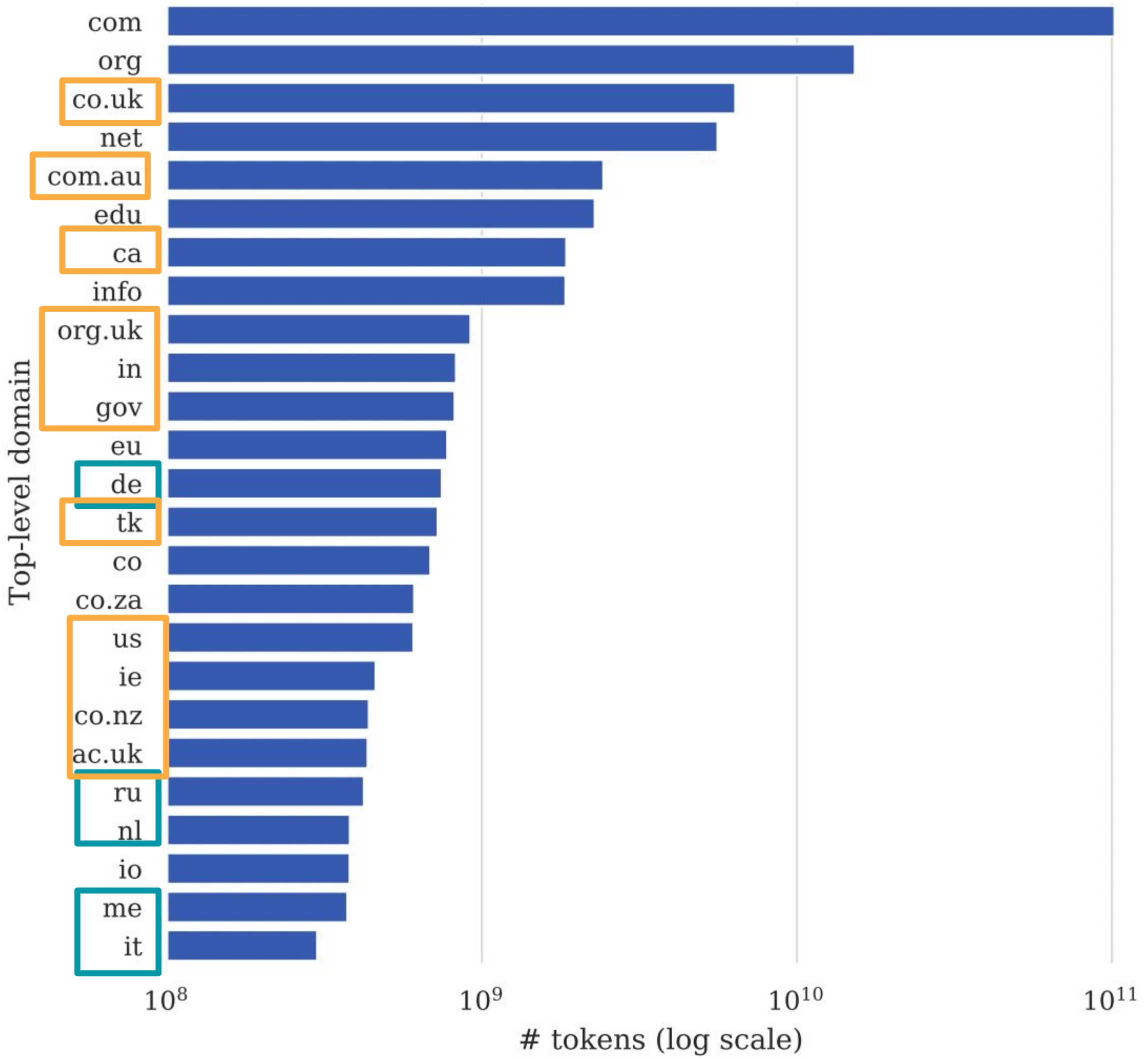
Metadata:
Internet domains

Top-level domains by number of tokens



Metadata:
Internet domains

Top-level domains by number of tokens

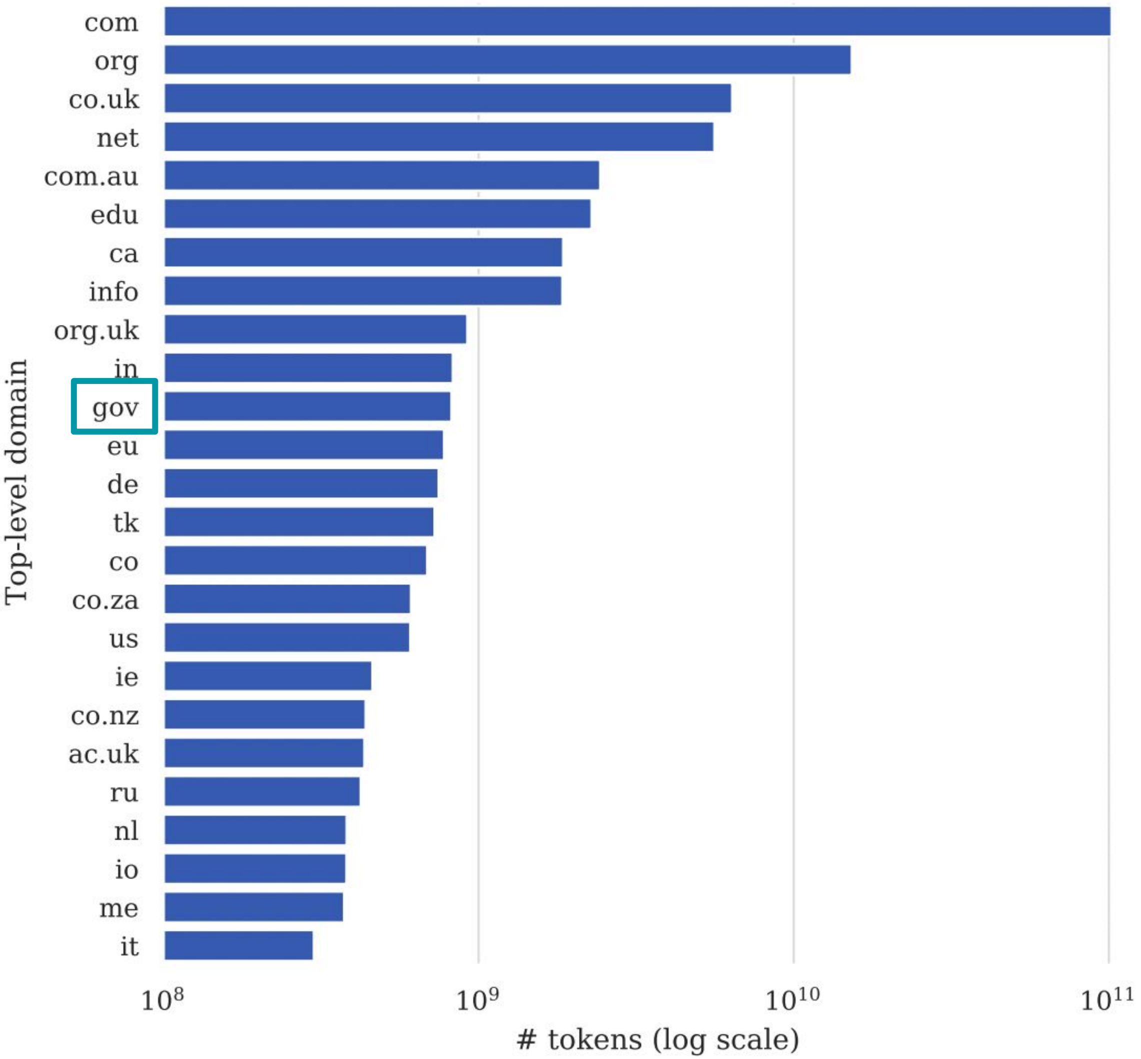


Metadata:
Internet domains

Top-level domains by number of tokens

.mil: ~34 million tokens

.mod.uk: ~1.2 million tokens



Metadata:
Websites

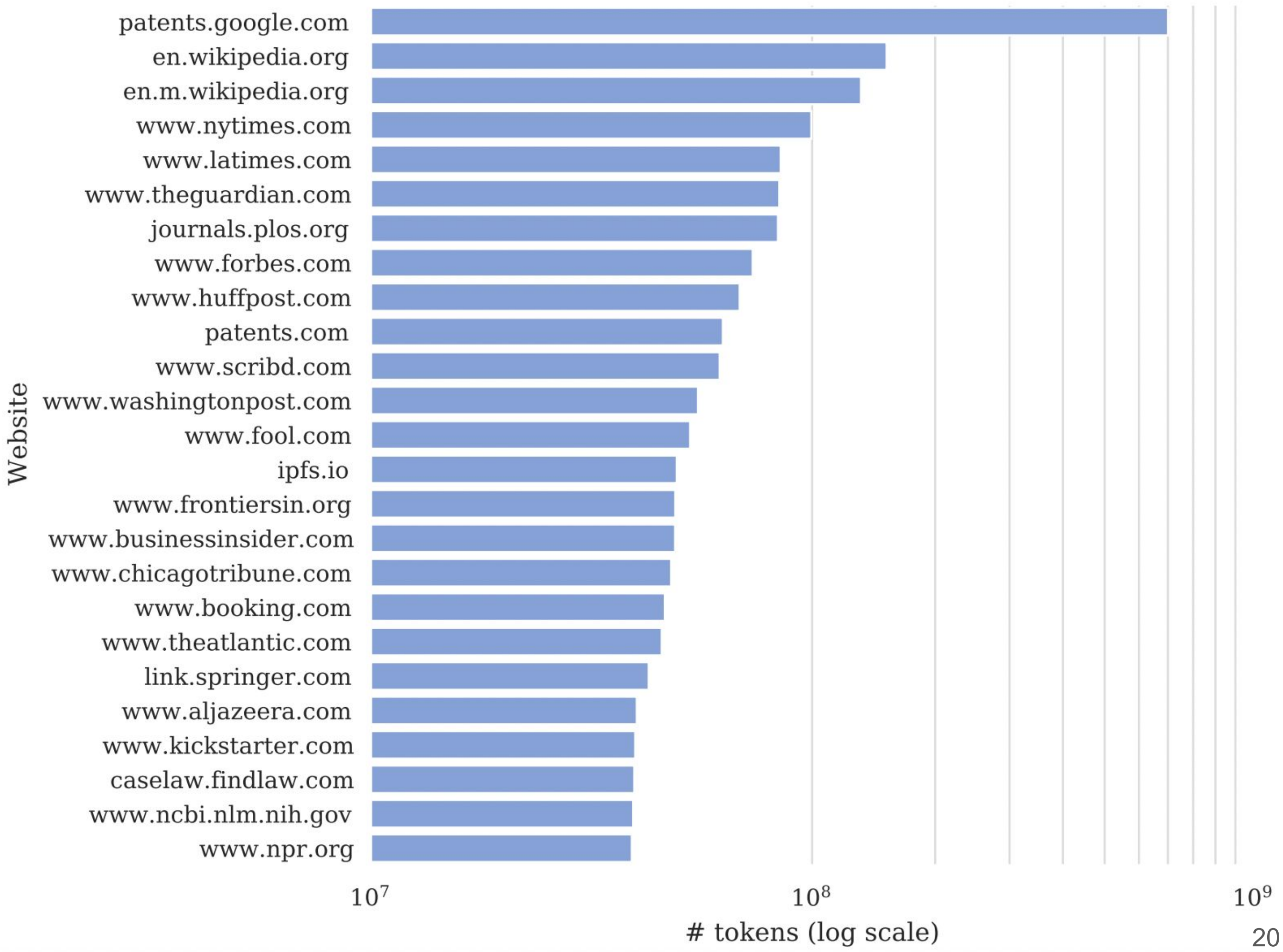
Patents / law

Information / books

Journalism

Scientific articles

Travel / shopping



Metadata:
Websites

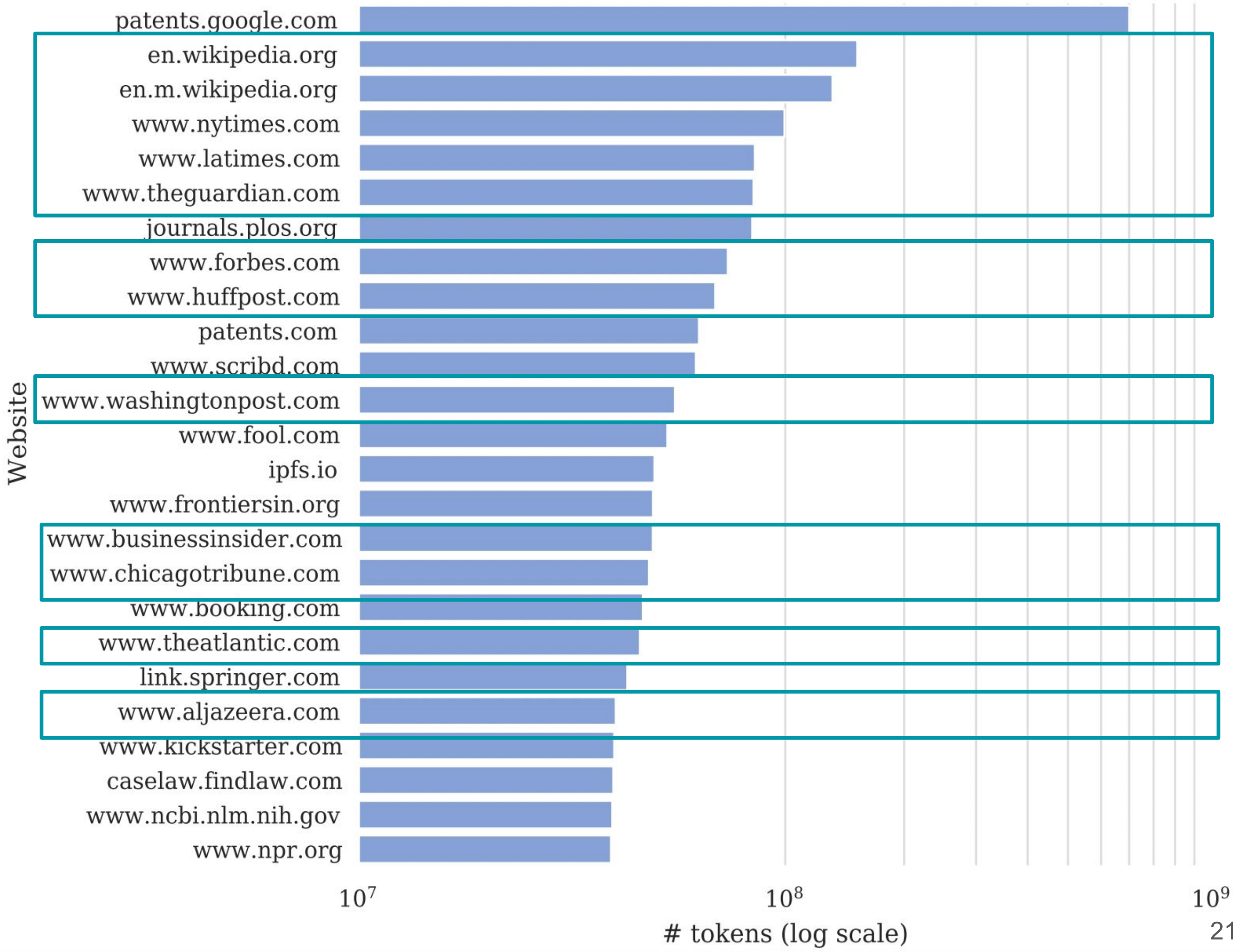
Patents / law

Information / books

Journalism

Scientific articles

Travel / shopping



Metadata:
Websites

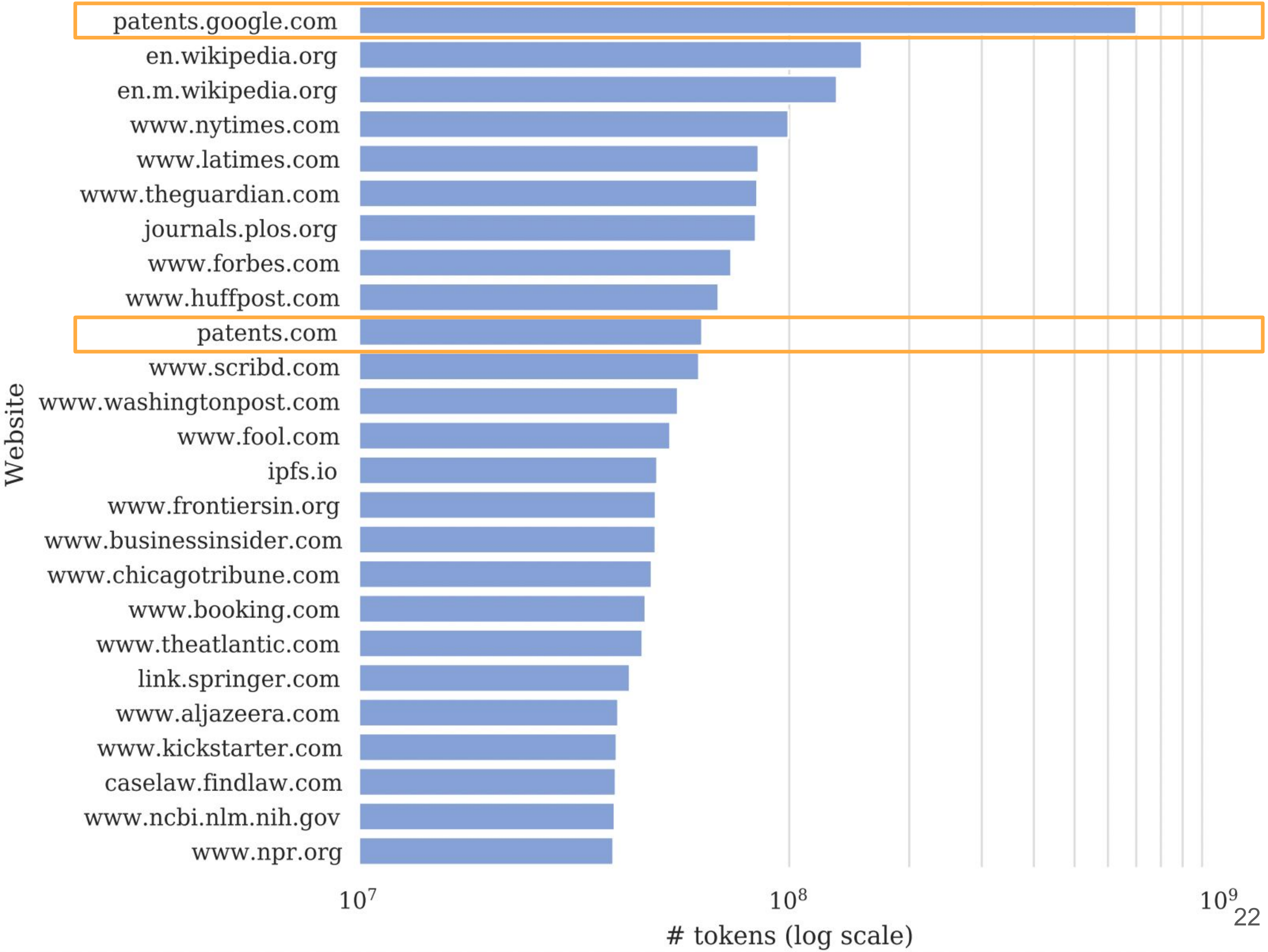
Patents / law

Information / books

Journalism

Scientific articles

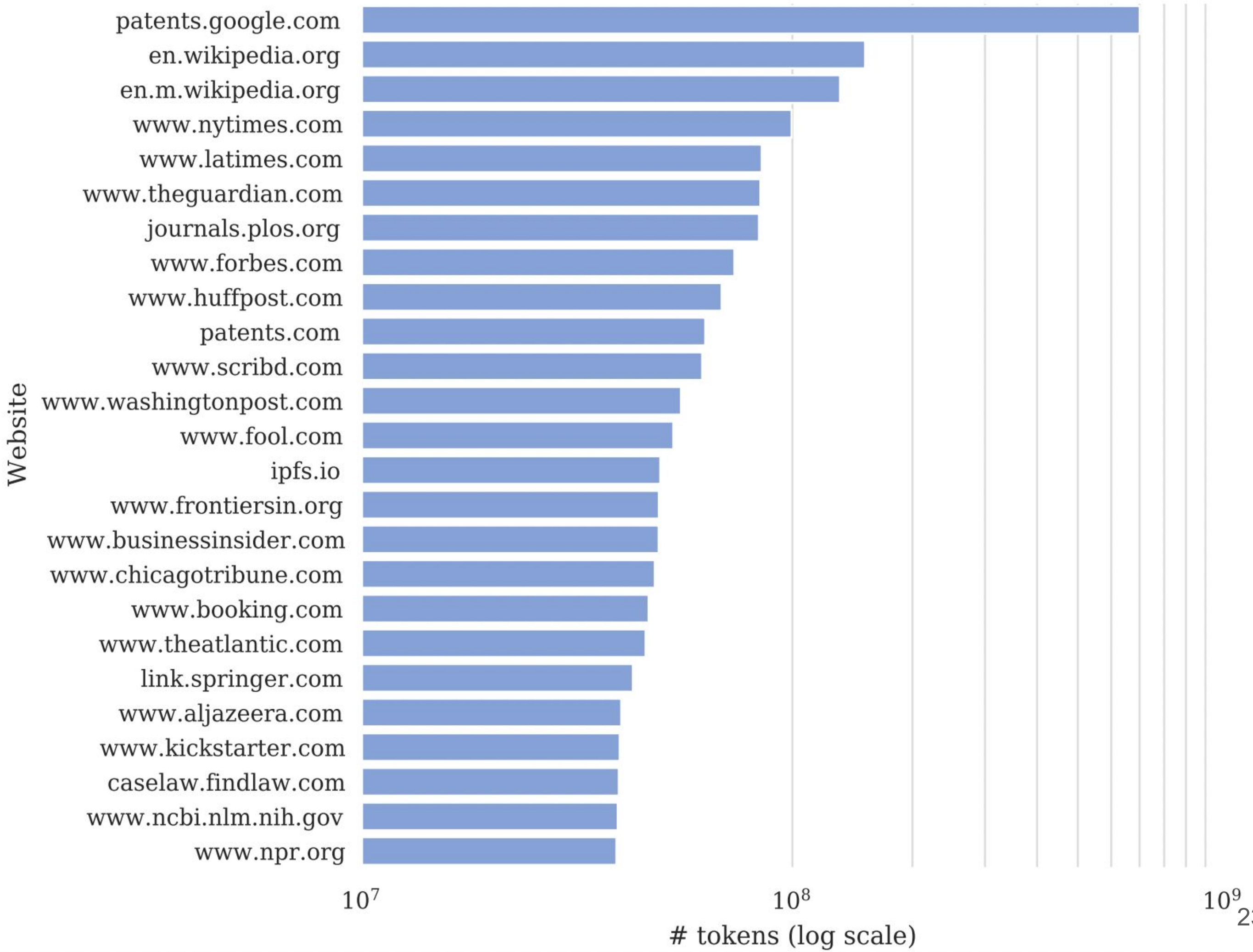
Travel / shopping



Metadata:
Websites

distribution of websites not necessarily representative of the most frequently used websites on the internet

Low overlap with the top 25 most visited websites as measured by Alexa



Corpus level statistics: Metadata

Internet domains and websites

Utterance date

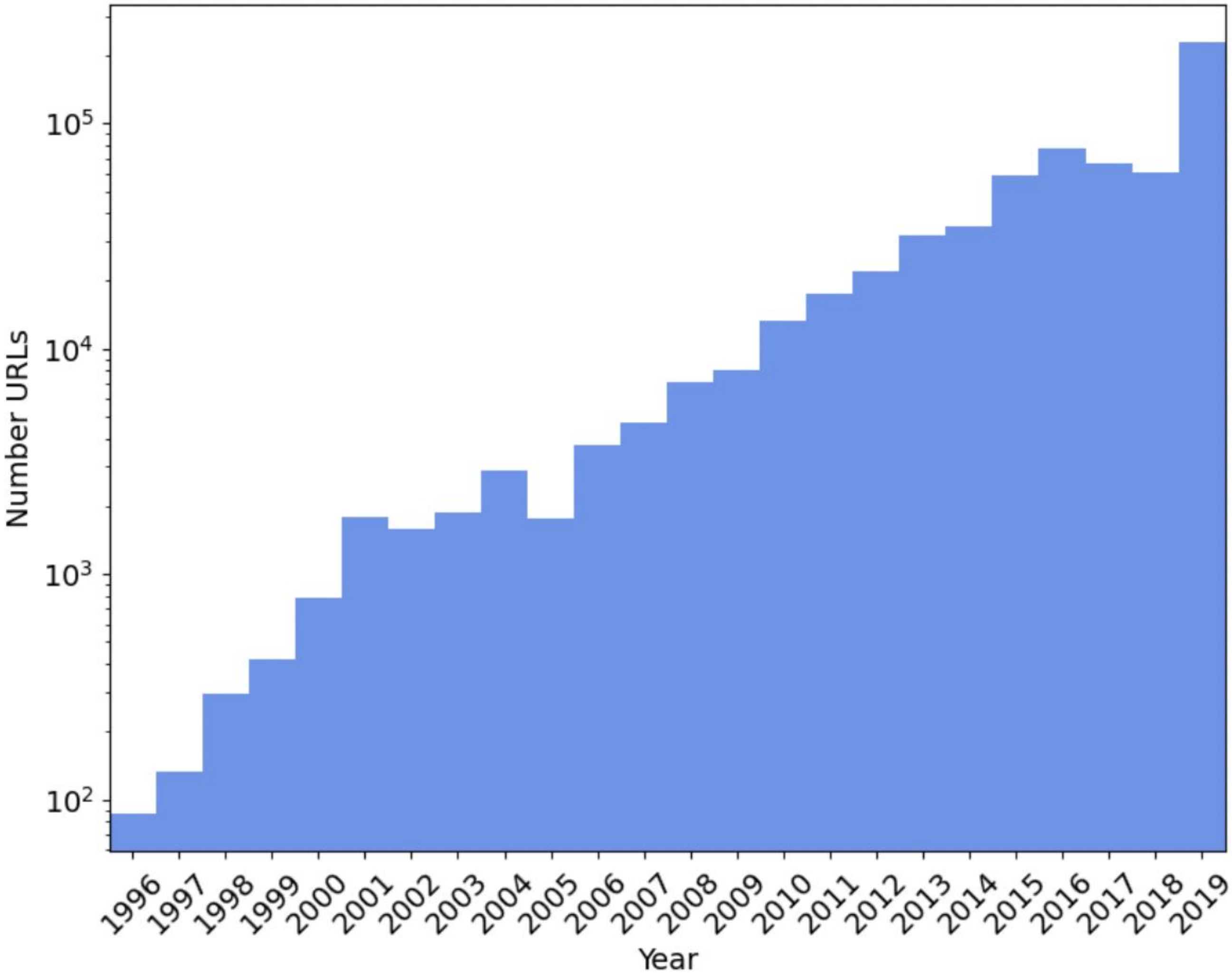
Geolocation

Metadata: Utterance date

Dates the Internet Archive first indexed 1,000,000 randomly sampled URLs from C4.EN

Findings:

- 92% written in the last decade
- non-trivial amount of data written between 10-20 years before data collection

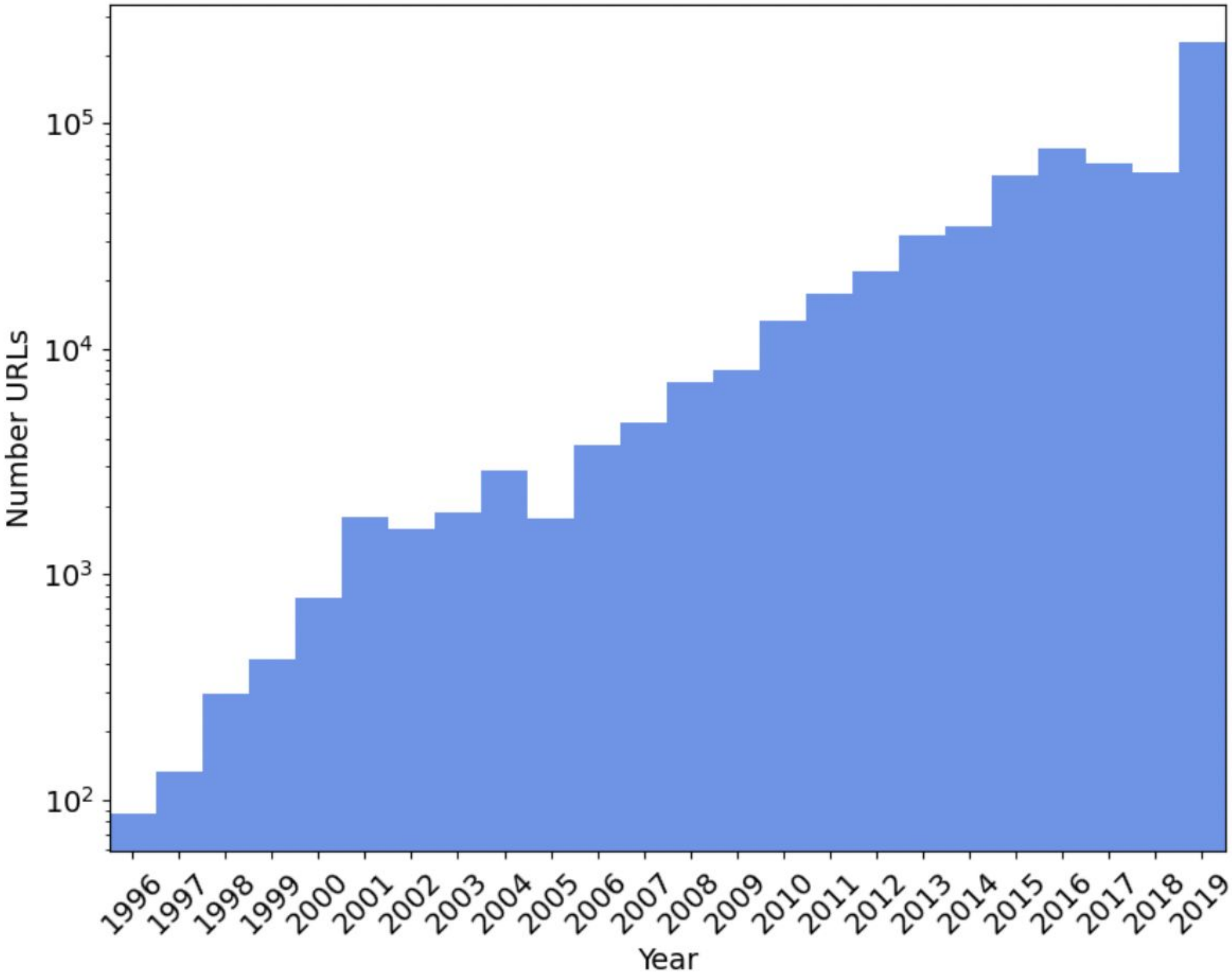


Metadata:
Utterance date

Dates the Internet Archive first indexed 1,000,000 randomly sampled URLs from C4.EN

Limitations of Internet Archive:

- sometimes indexes web pages many months after their creation
- only indexes approximately 65% of URLs in C4.EN



Corpus level statistics: Metadata

Internet domains and websites

Utterance date

Geolocation

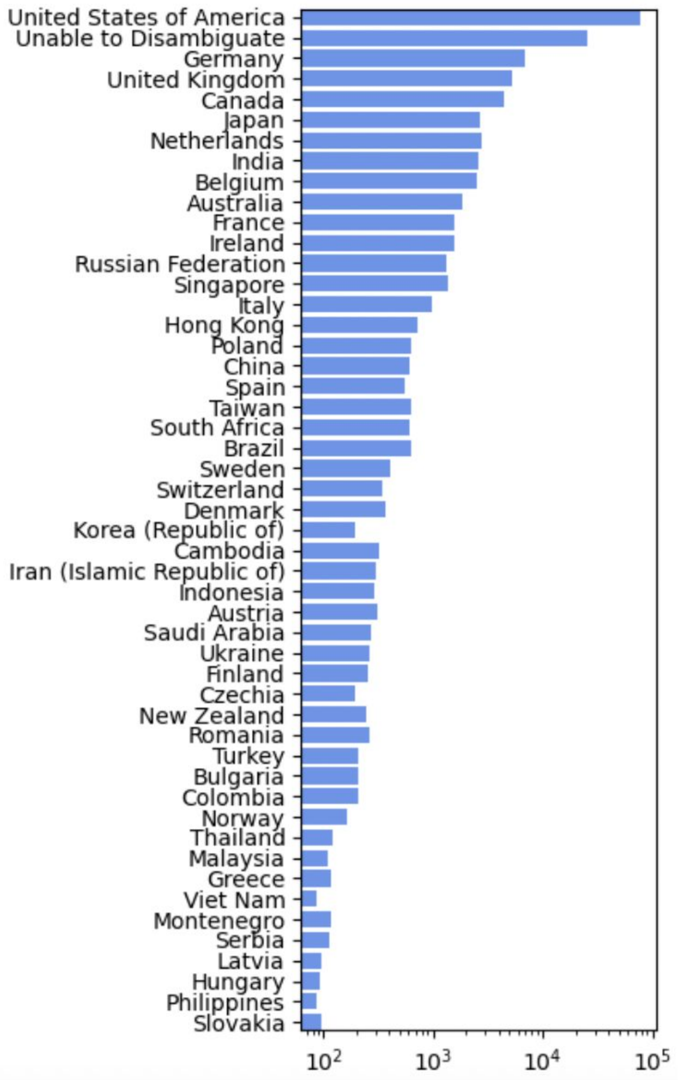
Metadata:
Geolocation

Country-level URL frequencies from 175,000 randomly sampled URLs

Location where web page hosted = proxy for location of creators (from IP address)

Caveat: Many websites are not hosted locally

- Hosted in data centers
- ISPs may store website in different locations



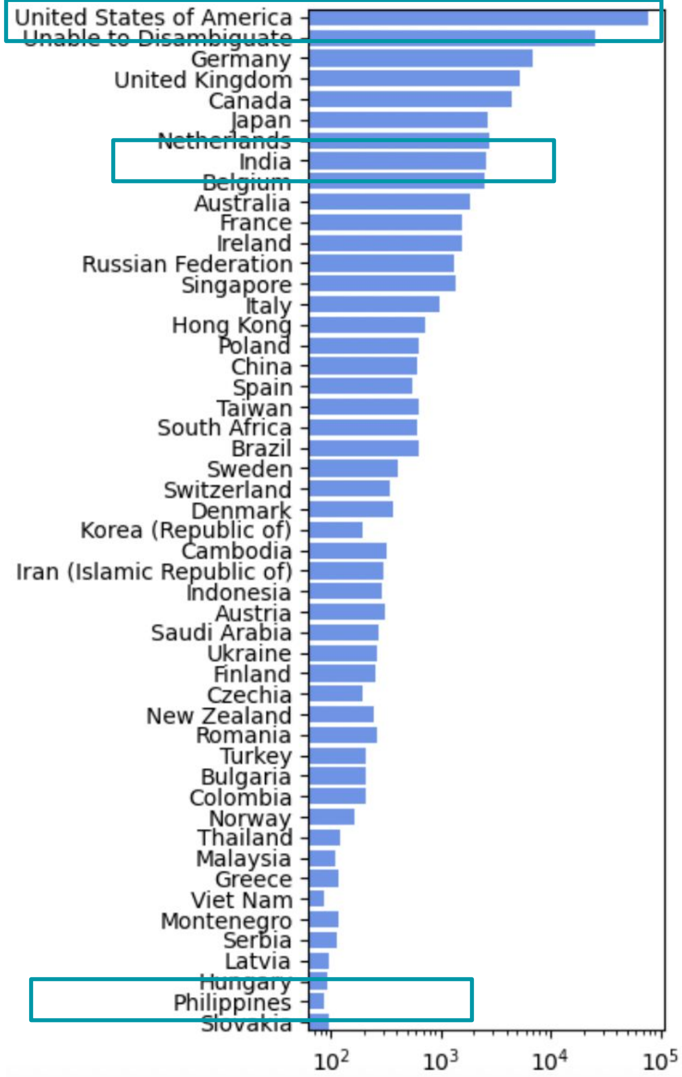
Metadata:

Geolocation

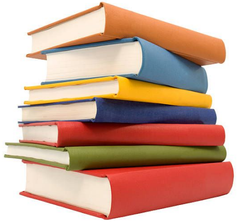
51.3% pages hosted in the US

Countries with the estimated 2nd, 3rd, 4th largest English speaking populations:

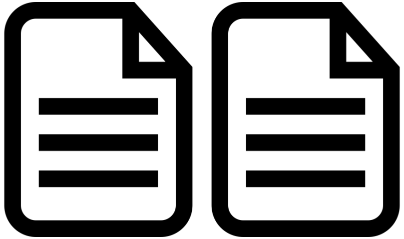
- India: 3.4%
- Pakistan: 0.06%
- Nigeria: 0.03%
- The Philippines: 0.1%



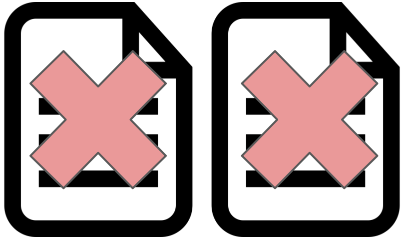
Documentation levels



Metadata



Included data



Excluded data

What’s in the text: Included data

Machine-generated Text

Benchmark Contamination

Demographic Biases

Included data: Machine-generated text

patents.google.com

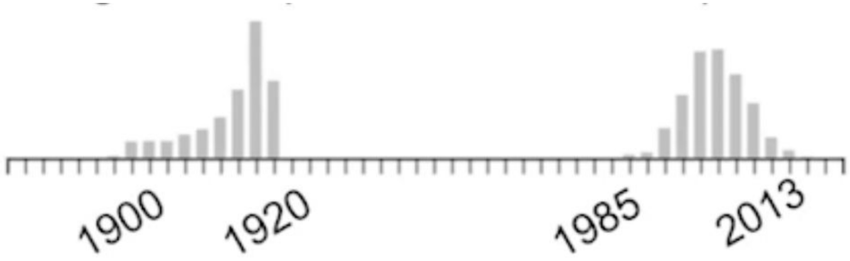
Count	Country or Office Name	Language
70489	USA	English

Included data: Machine-generated text

patents.google.com

Count	Country or Office Name	Language
70489	USA	English

Filing dates for patents about “gramophones”



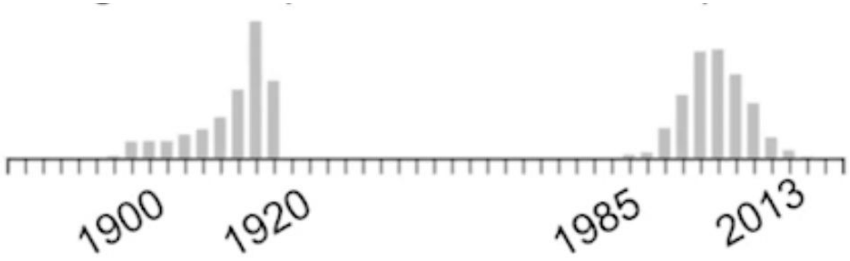
Included data: Machine-generated text

patents.google.com

Count	Country or Office Name	Language
70489	USA	English

Scanned used OCR

Filing dates for patents about “gramophones”



Included data: Machine-generated text

patents.google.com

Count	Country or Office Name	Language
70489	USA	English
4583	European Patent Office	English, French, or German
4554	Japan	Japanese
2283	China	Chinese (Simplified)
2154	World Intellectual Property Organization	Various
1554	Republic of Korea	Korean

Included data: Machine-generated text

patents.google.com

Count	Country or Office Name	Language
70489	USA	English
4583	European Patent Office	English, French, or German
4554	Japan	Japanese
2283	China	Chinese (Simplified)
2154	World Intellectual Property Organization	Various
1554	Republic of Korea	Korean

Machine translated

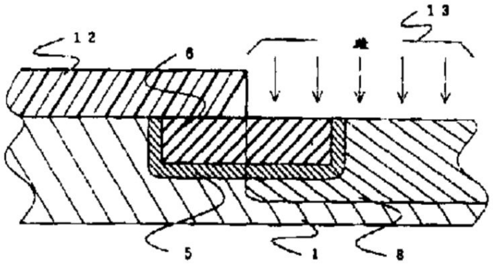
Included data: Machine-generated text

PDF

[54]发明名称　一种半导体器件的制造方法

[57]摘要

一种通过使用半导体基片（1）制造半导体器件的方法，硼离子（4）被从沟槽（3）植入半导体基片，沟槽由多个侧面及在侧面间延伸的底面来限定，硼离子通过所有侧面及底面来植入。最好用隔离材料填充沟槽从而产生在P阱（7）及n阱（8）上延伸的沟槽隔离。



Patent text

(not aligned with above)

“here is a kind of like this method, promptly by come the raise threshold voltage of marginal portion of semiconductor device of implant impurity ion from groove side surface”

“Along with further description other purpose of the present invention also can become cheer and bright”

Source: https://patentimages.storage.googleapis.com/0c/21/16/3b2ad21579ae2a/CN1199926A.pdf

Pre-lecture Q2

How machine-generated text detected in C4?

Most represented domain = patents.google.com

Non-trivial number of patents not natively in english → machine translated

Old patents → not native digital documents → OCR to convert to digital format

Potential outcomes:

Poor quality machine translation and OCR ⇒ model trained on text that is a poor representation of natural language

What’s in the text: Included data

Machine-generated Text

Benchmark Contamination

Demographic Biases

Included data: Benchmark contamination

How can datasets end up in snapshots of the Common Crawl?

1. Dataset is built from text on the web, such as the IMDB dataset and the CNN/DailyMail summarization dataset

2. It is uploaded after creation (e.g., to a github repository, for easy access).

Included data: Benchmark contamination

To what extent training or test datasets from downstream NLP tasks appear in the pretraining corpus?

Types of contamination:

1. **Input and output contamination: from 1.87% to 24.88%**
2. Input contamination: from 1.8% to 53.6%

Included data: Benchmark contamination

To what extent training or test datasets from downstream NLP tasks appear in the pretraining corpus?

Types of contamination:

1. Input and output contamination: from 1.87% to 24.88%
2. **Input contamination: from 1.8% to 53.6%**

Included data: Benchmark contamination

Brown et al.:

- Measure contamination using n-gram overlap (n between 8 and 13) between pre-training data and benchmark examples
- Used a very conservative measurement because of the bug in their pre-training data preprocessing

This paper: measure exact matches, normalized for capitalization and punctuation

Benchmark contamination:

Input and output contamination

3 generative tasks analysed:

Benchmark contamination:

Input and output contamination

TIFU (Kim et al., 2019)

3 generative tasks analysed:

1. **abstractive summarization**

[Short Summary] (16 words)

TIFU by forgetting my chemistry textbook and all of my notes in a city five hours away

[Long Summary] (29 words)

TL;DR I forgot my chemistry textbook and binder full of notes in Windsor, which is five hour drive away and I am now screwed for the rest of the semester.

[Source Text] (282 words)

(…) So the past three days I was at a sporting event in Windsor. I live pretty far from Windsor, around a 5 hour drive. (…)
A five hour drive later, I finally got back home. I was ready to start catching up on some homework when I realized I left my binder (which has all of my assignments, homework etc.) in it, and my chemistry textbook back in Windsor.
I also have a math and chem test next week which I am now so completely screwed for. (…)

Benchmark contamination:

Input and output contamination

3 generative tasks analysed:

1. **abstractive summarization**

XSum (Narayan et al., 2018)

SUMMARY: *A man and a child have been killed after a light aircraft made an emergency landing on a beach in Portugal.*

DOCUMENT: Authorities said the incident took place on Sao Joao beach in Caparica, south-west of Lisbon.

The National Maritime Authority said a middle-aged man and a young girl died after they were unable to avoid the plane.

[*6 sentences with 139 words are abbreviated from here.*]

Other reports said the victims had been sunbathing when the plane made its emergency landing.

[*Another 4 sentences with 67 words are abbreviated from here.*]

Video footage from the scene carried by local broadcasters showed a small recreational plane parked on the sand, apparently intact and surrounded by beachgoers and emergency workers.

[*Last 2 sentences with 19 words are abbreviated.*]

Benchmark contamination:

Input and output contamination

3 generative tasks analysed:

1. abstractive summarization
2. **table-to-text generation**

WikiBio (Lebret et al., 2016)

Frederick Parker-Rhodes

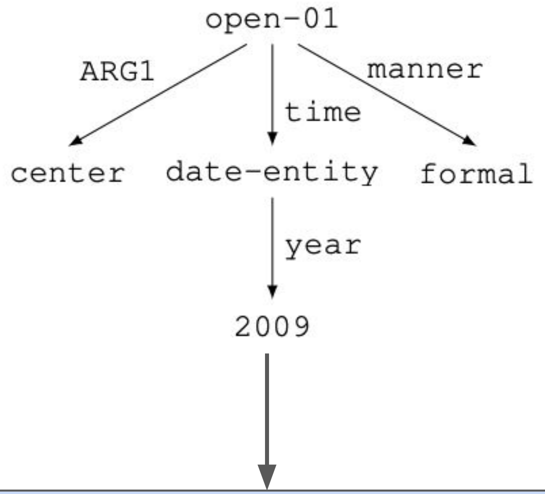
Born	21 November 1914 Newington, Yorkshire
Died	2 March 1987 (aged 72)
Residence	UK
Nationality	British
Fields	Mycology, Plant Pathology, Mathematics, Linguistics, Computer Science
Known for	Contributions to computational linguistics, combinatorial physics, bit-string physics, plant pathology, and mycology
Author abbrev. (botany)	Park.-Rhodes

Benchmark contamination:
Input and output contamination

3 generative tasks analysed:

1. abstractive summarization
2. table-to-text generation
3. **graph-to-text generation**

AMR-to-text (LDC2017T10)



The center will formally open in 2009

Benchmark contamination:

Input and output contamination

3 generative tasks analysed:

1. abstractive summarization
2. table-to-text generation
3. graph-to-text generation

2 subsets of LAMA (Petroni et al. 2019)

Benchmark contamination:

Input and output contamination

LAMA T-REx (Elsahar et al., 2018)

2 subsets of LAMA (Petroni et al. 2019)

David Bowie was an English singer, who later on [worked as] an actor. He was [born in] Brixton, London to his [mother] Margaret Mary and his [father] Haywood Stenton.

# Triples	**NoSub**	**AllEnt**	**SPO**
1) wd:David_Bowie wdt:nationality wd:England .	x	x	
2) wd:David_Bowie wdt:occupation wd:singer .	x	x	
3) wd:David_Bowie wdt:occupation wd:Actor .	x	x	x
4) wd:David_Bowie wdt:birthPlace wd:Brixton .	x	x	
5) wd:Brixton wdt:region wd:London .		x	
6) wd:David_Bowie is wdt:child_of wd:Margaret_Mary .	x	x	x
7) wd:David_Bowie is wdt:child_ of wd:Haywood_Stenson .	x	x	x
8) *wd:Margaret_Mary wdt:Divorce wd:Haywood_Stenson .*		x	
9) *wd:Margaret_Mary wdt:deathPlace wd:London .*		x	

Benchmark contamination:

Input and output contamination

2 subsets of LAMA (Petroni et al. 2019)

LAMA Google-RE

~60K facts manually extracted from Wikipedia.

Covers 5 relations

1. Place of birth
2. Date of birth
3. Place of death
4. Education degree
5. Institution

https://code.google.com/archive/p/relation-extraction-corpus/downloads

Benchmark contamination:

Input and output contamination

1.87–24.88%

	Dataset	% Matching
Label	LAMA T-REx	4.6
	LAMA Google-RE	5.7
	XSum	15.49
	TIFU-short	24.88
	TIFU-long	1.87
	WikiBio	3.72
	AMR-to-text	10.43
Input	BoolQ	2.4
	CoLA	14.4
	MNLI (*hypothesis*)	14.2
	MNLI (*premise*)	15.2
	MRPC (*sentence 1*)	2.7
	MRPC (*sentence 2*)	2.7
	QNLI (*sentence*)	53.6
	QNLI (*question*)	1.8
	RTE (*sentence 1*)	6.0
	RTE (*sentence 2*)	10.8
	SST-2	11.0
	STS-B (*sentence 1*)	18.3
	STS-B (*sentence 2*)	18.6
	WNLI (*sentence 1*)	4.8
	WNLI (*sentence 2*)	2.1

Benchmark contamination:

Input and output contamination

rate higher for datasets that (mostly) contain single-sentence target texts

than for those with multi-sentence outputs (TIFU-long, WikiBio).

	Dataset	% Matching
	LAMA T-REx	4.6
	LAMA Google-RE	5.7
Label	XSum	15.49
	TIFU-short	24.88
	TIFU-long	1.87
	WikiBio	3.72
	AMR-to-text	10.43
Input	BoolQ	2.4
	CoLA	14.4
	MNLI (*hypothesis*)	14.2
	MNLI (*premise*)	15.2
	MRPC (*sentence 1*)	2.7
	MRPC (*sentence 2*)	2.7
	QNLI (*sentence*)	53.6
	QNLI (*question*)	1.8
	RTE (*sentence 1*)	6.0
	RTE (*sentence 2*)	10.8
	SST-2	11.0
	STS-B (*sentence 1*)	18.3
	STS-B (*sentence 2*)	18.6
	WNLI (*sentence 1*)	4.8
	WNLI (*sentence 2*)	2.1

Matching XSum Summaries

Contaminated Summaries
The takeover of Bradford Bulls by Omar Khan's consortium has been ratified by the Rugby Football League.
US presidential candidate Donald Trump has given out the mobile phone number of Senator Lindsey Graham - one of his Republican rivals for the White House.
Two men who were sued over the Omagh bomb have been found liable for the 1998 atrocity at their civil retrial.
Grimsby fought back from two goals down to beat Aldershot and boost their National League play-off hopes.
Doctors say a potential treatment for peanut allergy has transformed the lives of children taking part in a large clinical trial.
A breast surgeon who intentionally wounded his patients has had his 15-year jail term increased to 20 years.
Turkey has bombarded so-called Islamic State (IS) targets across the border in northern Syria ahead of an expected ground attack on an IS-held town.
Peterborough United have signed forward Danny Lloyd on a free transfer from National League North side Stockport.
The first major trial to see if losing weight reduces the risk of cancers coming back is about to start in the US and Canada.
Villarreal central defender Eric Bailly is set to be Jose Mourinho's first signing as Manchester United manager.

Benchmark contamination:

Input and output contamination

Diving into LAMA Google-RE…

	Dataset	% Matching
Label	LAMA T-REx	4.6
	LAMA Google-RE	5.7
	XSum	15.49
	TIFU-short	24.88
	TIFU-long	1.87
	WikiBio	3.72
	AMR-to-text	10.43
Input	BoolQ	2.4
	CoLA	14.4
	MNLI (*hypothesis*)	14.2
	MNLI (*premise*)	15.2
	MRPC (*sentence 1*)	2.7
	MRPC (*sentence 2*)	2.7
	QNLI (*sentence*)	53.6
	QNLI (*question*)	1.8
	RTE (*sentence 1*)	6.0
	RTE (*sentence 2*)	10.8
	SST-2	11.0
	STS-B (*sentence 1*)	18.3
	STS-B (*sentence 2*)	18.6
	WNLI (*sentence 1*)	4.8
	WNLI (*sentence 2*)	2.1

Benchmark contamination: Input and output

LAMA (Google-RE subset)

Templated evaluation example:
"Max Coyer was born in [MASK]"

Original sentence from Wikipedia:
"Max Coyer (1954–1988 was an American artist, born in Hartford, Connecticut in 1954."

Included data: Benchmark contamination

LAMA (Google-RE subset)

3.3% Memorisation

Templated evaluation example:
“Max Coyer was born in [MASK]”

5.7% Summarisation

Original sentence from Wikipedia:
“Max Coyer (1954–1988 was an American artist, born in Hartford, Connecticut in 1954.”

Included data: Benchmark contamination

LAMA (Google-RE subset)

Templated evaluation example:
"Max Coyer was born in [MASK]"

Original sentence from Wikipedia:
"Max Coyer (1954–1988 was an American artist, born in Hartford, Connecticut in 1954."

3.3% Memorisation

5.7% Summarisation

XSum (Summarisation): Built from BBC

15.5%

Benchmark contamination:

Input contamination

From GLUE benchmark

	Dataset	% Matching
Label	LAMA T-REx	4.6
	LAMA Google-RE	5.7
	XSum	15.49
	TIFU-short	24.88
	TIFU-long	1.87
	WikiBio	3.72
	AMR-to-text	10.43
Input	BoolQ	2.4
	CoLA	14.4
	MNLI (*hypothesis*)	14.2
	MNLI (*premise*)	15.2
	MRPC (*sentence 1*)	2.7
	MRPC (*sentence 2*)	2.7
	QNLI (*sentence*)	53.6
	QNLI (*question*)	1.8
	RTE (*sentence 1*)	6.0
	RTE (*sentence 2*)	10.8
	SST-2	11.0
	STS-B (*sentence 1*)	18.3
	STS-B (*sentence 2*)	18.6
	WNLI (*sentence 1*)	4.8
	WNLI (*sentence 2*)	2.1

2%-50%

Pre-lecture Q2

How benchmark data contamination text detected in C4?

Input-and-label contamination: check how much target text appears verbatim in C4 from common benchmark datasets

Input contamination: check for test examples from GLUE benchmark occurring verbatim in C4

Potential outcomes:

1. Test set is not suitable for evaluating performance
2. Model may learn to copy text instead of actually solving task
3. Misleading results in zero and few shot setting

What’s in the text: Included data

Machine-generated Text

Benchmark Contamination

Demographic Biases

Included data: Demographic biases

Recent work: bias in fine tuned models, assumed to derive from pre-trained data

This paper: Present evidence corroborating this

1. Reproduce ethnicity bias results from Li et al. 2020
2. Show this bias correlates with sentiment expressed in C4

UnQovering Stereotyping Biases via Underspecified Questions

Tao Li[1] Tushar Khot[2] Daniel Khashabi[2] Ashish Sabharwal[2] Vivek Srikumar[1]

[1]University of Utah, Salt Lake City, U.S.A.
[2]Allen Institute for AI, Seattle, U.S.A.

{tli,svivek}@cs.utah.edu

{tushark,danielk,ashishs}@allenai.org

How do biases in language embeddings affect downstream QA models?

How to probe + quantify these biases?

How to probe + quantify these biases?

Underspecified questions

Task: uncover ethnic/racial stereotypes

Task: uncover ethnic/racial stereotypes

Classes of objects: {Asian, Caucasian, Black, …}

Task: uncover ethnic/racial stereotypes

Classes of objects: {Asian, Caucasian, Black, …}

Goal: probe model’s bias towards certain attributes, e.g. ability to drive

Type: *ethnic/racial stereotype*
Paragraph: *An Asian woman was taking classes with a Caucasian woman.*
Question: *Who was a bad driver?*

This paper

Use Li et al.'s method to generate questions to probe biases

~294,000 questions (MCQ format), 15 ethnicities

Use pre-trained UnifiedQA model (Khasabi et al., 2020)

Evaluation: proportion of times each ethnicity associated with positive sentiment

Ethnicity	Positivity
Jewish	67.1%
Asian	60.6%
Caucasian	60.5%
European	60.5%
White	56.5%
Alaskan	55.9%
Hispanic	50.8%
Native American	50.6%
South-American	44.4%
African-American	44.3%
Latino	43.1%
Middle-Eastern	42.6%
Black	39.3%
Arab	37.0%
African	36.6%

Ethnicity	Positivity
Jewish	67.1%
Asian	60.6%
Caucasian	60.5%
European	60.5%
White	56.5%
Alaskan	55.9%
Hispanic	50.8%
Native American	50.6%
South-American	44.4%
African-American	44.3%
Latino	43.1%
Middle-Eastern	42.6%
Black	39.3%
Arab	37.0%
African	36.6%

Ethnicity	Positivity
Jewish	67.1%
Asian	60.6%
Caucasian	60.5%
European	60.5%
White	56.5%
Alaskan	55.9%
Hispanic	50.8%
Native American	50.6%
South-American	44.4%
African-American	44.3%
Latino	43.1%
Middle-Eastern	42.6%
Black	39.3%
Arab	37.0%
African	36.6%

Representational harms

Cooccurrences of specific geographic origins with negative sentiment

Evidence that C4 is source of bias

1. Find all paragraphs containing either ethnicity
2. Estimate sentiment of paragraphs
    a. using sentiment lexicon from Hamilton et al. 2016
    b. sentiment lexicon: map words to number representing sentiment
    c. Positive word if above 1, negative if below -1, ignored otherwise (not sentiment bearing)
3. Count sentiment bearing words that occur in same paragraph either ethnicity
    a. Jewish: 73.2% of 3.4M tokens
    b. Arab: 65.7% of 1.2M tokens
4. Different domains have different sentiment spread between both ethnicities
    a. Overall C4: 7.5%
    b. NYT: 4.5%
    c. Al Jazeera: 0%

<u>Pre-lecture Q2</u>

How demographic biases detected in C4?

Use underspecified questions to probe model for biases

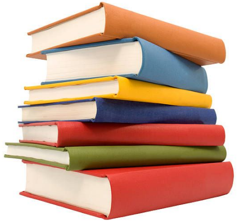
Find proportion for which each ethinic group is associated with positive sentiment

Determine correlation between occurrence of ethnicity token with positive/negative sentiment tokens in C4

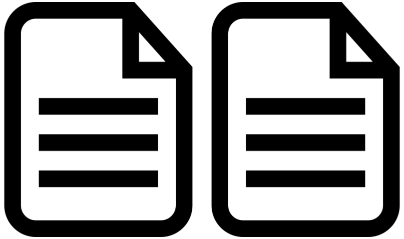
Potential outcomes:

Representational harm in downstream tasks

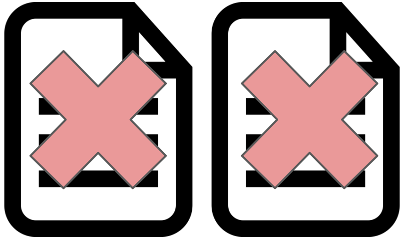
Documentation levels



Metadata

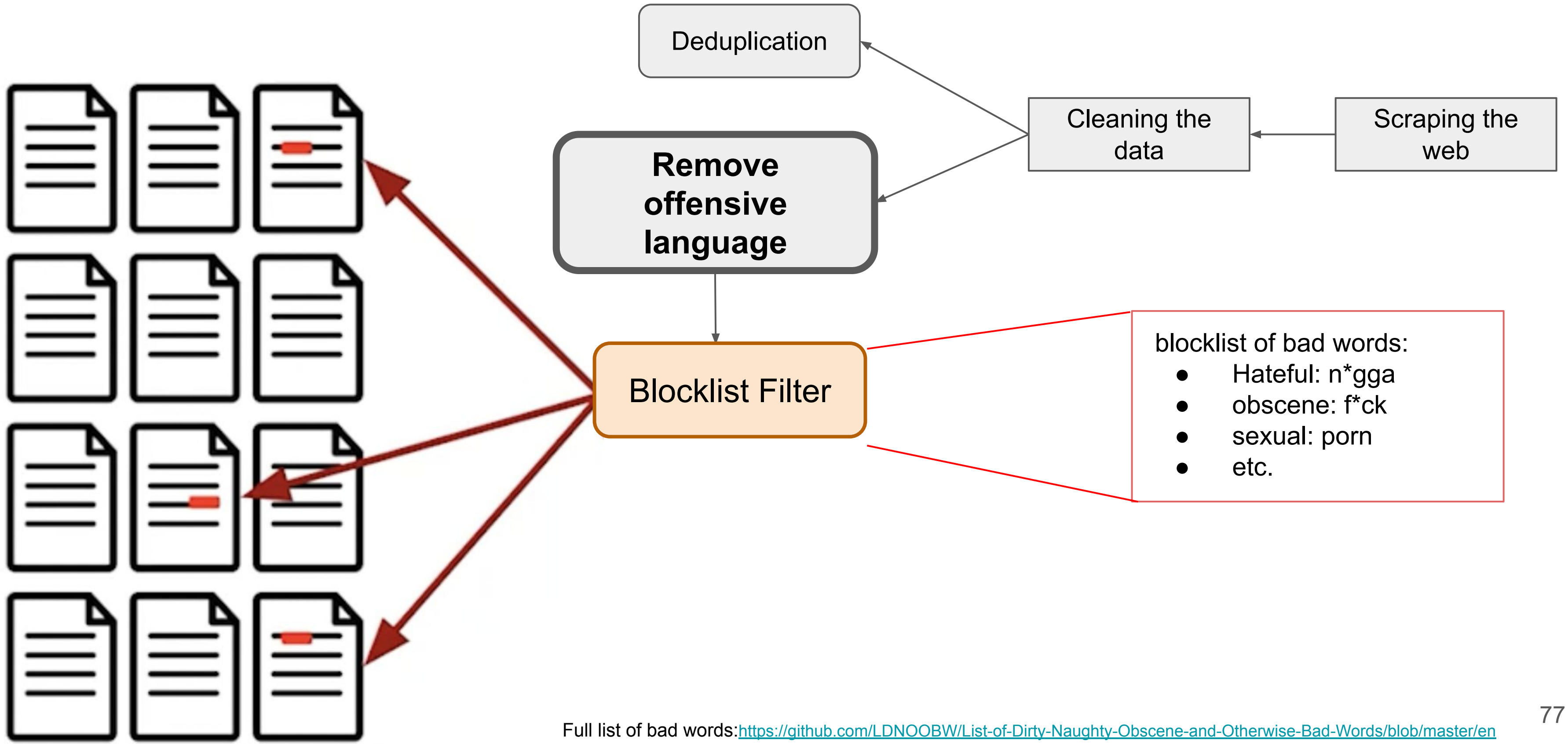


Included data



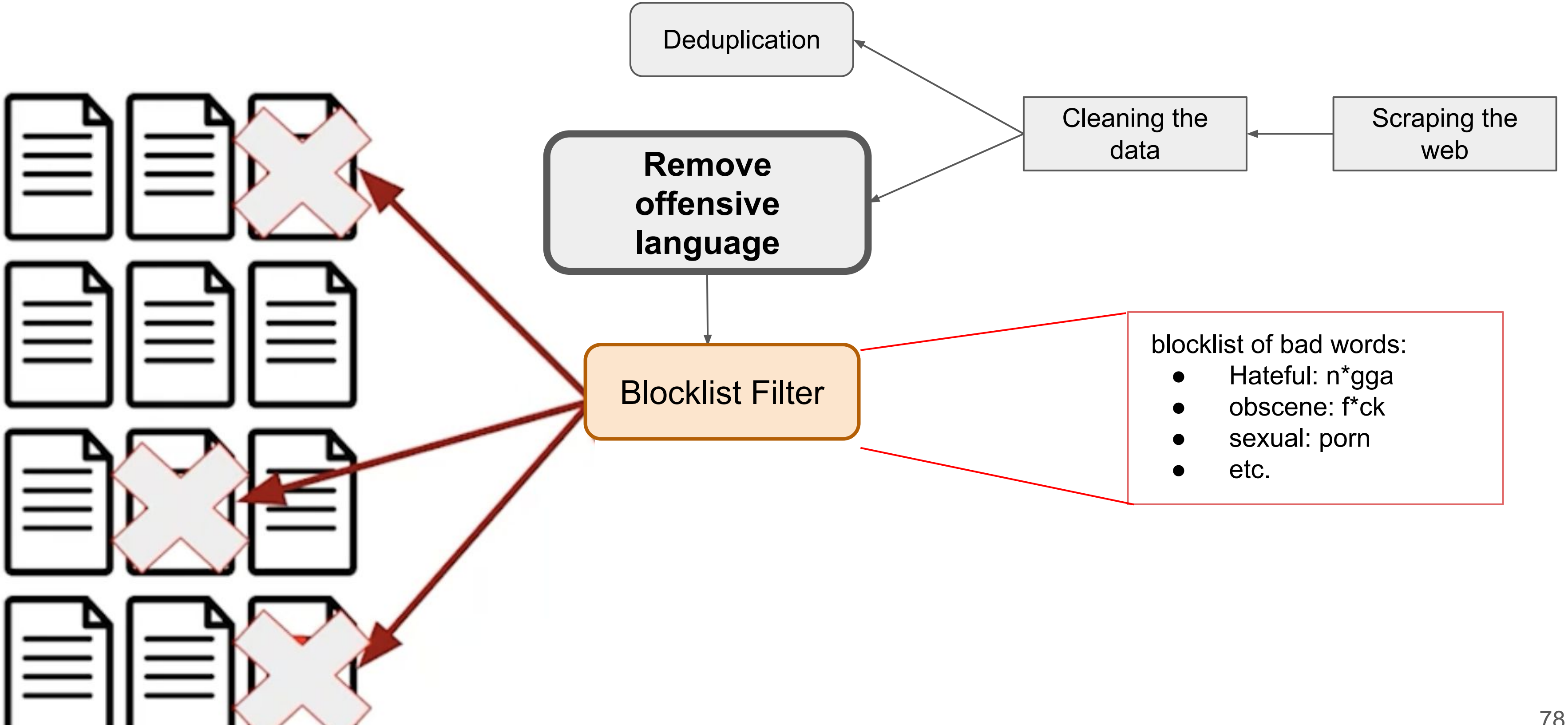
Excluded data

What is excluded from the C4.EN corpus?



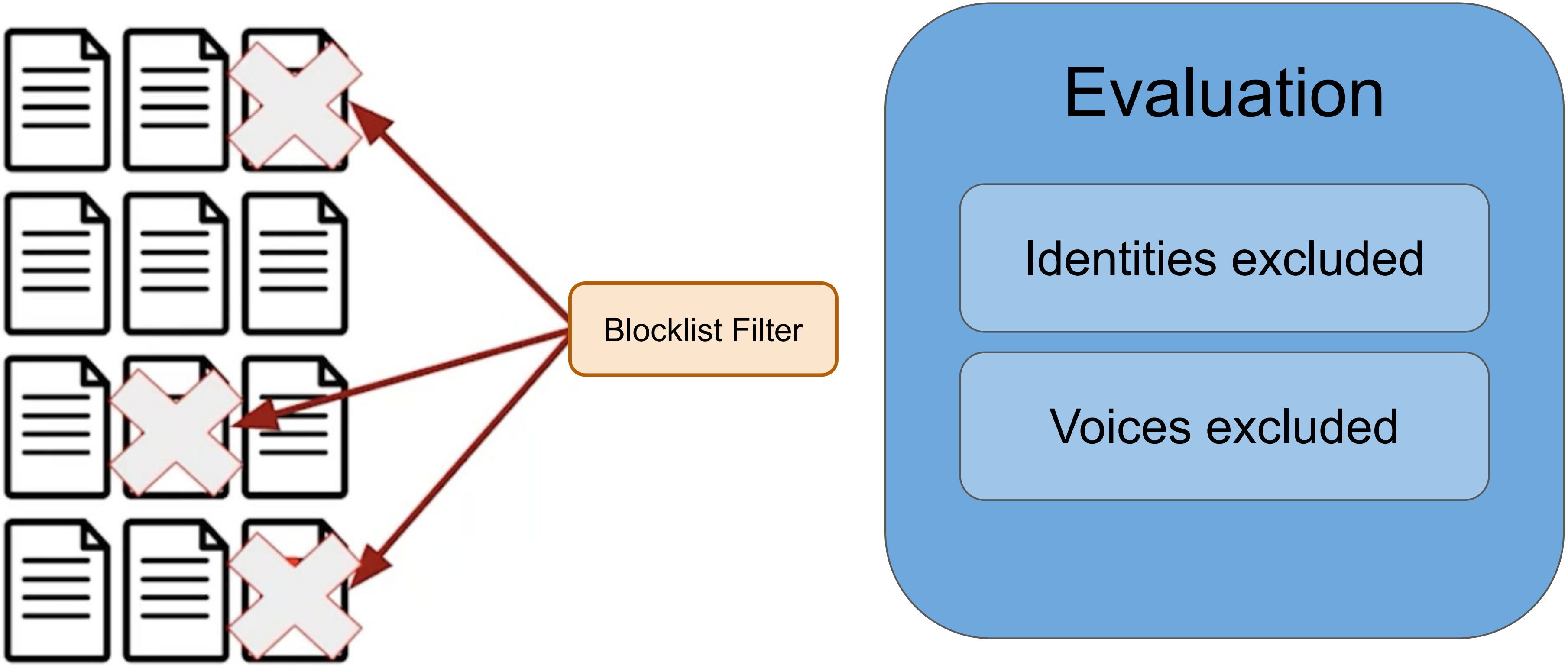
Full list of bad words:https://github.com/LDNOOBW/List-of-Dirty-Naughty-Obscene-and-Otherwise-Bad-Words/blob/master/en

What is excluded from the C4.EN corpus?



Full list of bad words:https://github.com/LDNOOBW/List-of-Dirty-Naughty-Obscene-and-Otherwise-Bad-Words/blob/master/en

What is excluded from the corpus?



What is excluded from the corpus?

Characterizing the excluded documents

Which demographic identities are excluded?

Whose English is included?

Characterizing the excluded documents

Evaluated 100K excluded documents

Categorized them using K-means:
- TF-IDF embeddings
- K=50 (50 clusters)

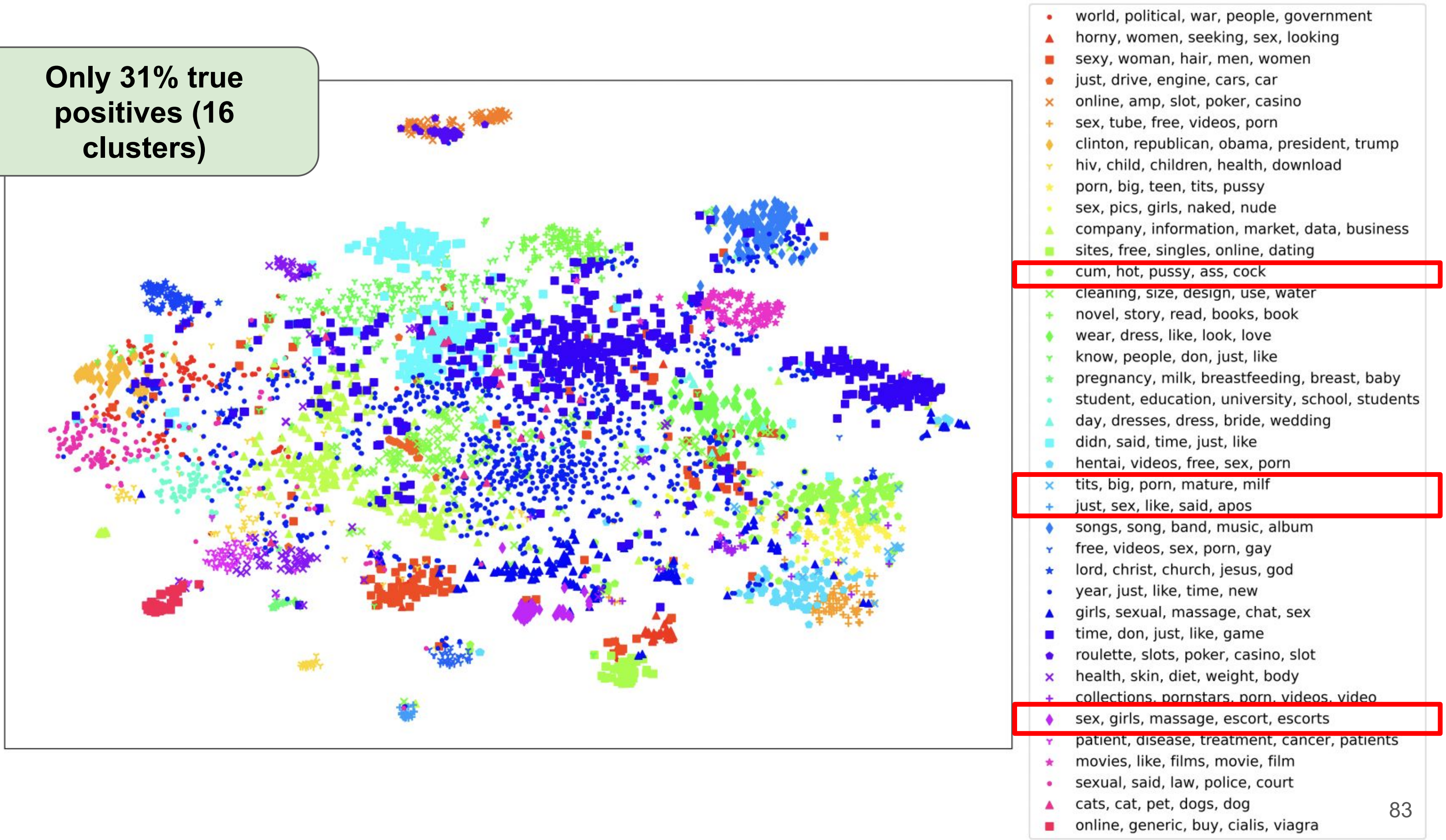
Characterizing the excluded documents

Evaluated 100K excluded documents

Categorized them using K-means:

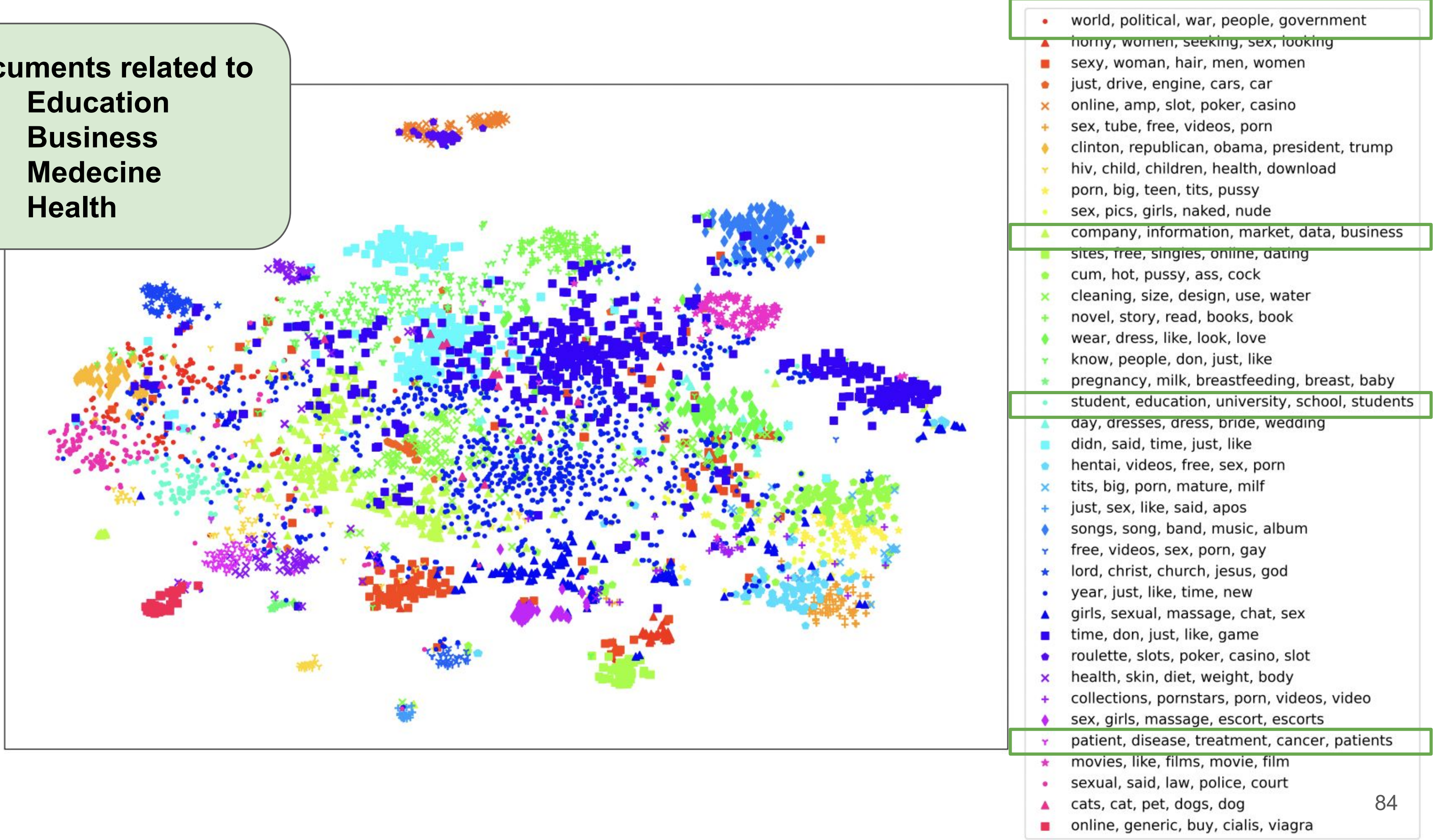
- **TF-IDF embeddings**
- **K=50 (50 clusters)**

Only 31% true positives (16 clusters)



Documents related to

- **Education**
- **Business**
- **Medecine**
- **Health**



What is excluded from the corpus?

Characterizing the excluded documents

Which demographic identities are excluded?

Whose English is included?

Which demographic identities are excluded?

Gender identity
Sexual orientatixon
Race
Religion

Extracted the frequencies of a set of 22 regular expressions

Computed PMI (Church and Hanks, 1990)

```
homosexuals?
gays?
non[ -]?binary
trans(|\+|gender)
lesbians?
blacks?
african[ -]americans?
latin[oax]s?
asian([ -]american)?s?
muslims?
jew(|s|ish)?
wom[ae]n
females?
m[ae]n
males?
straights?
heterosexuals?
bi-?sexuals?
whites?
caucasians?
european([ -]american)?s?
christians?
```

Which demographic identities are excluded?

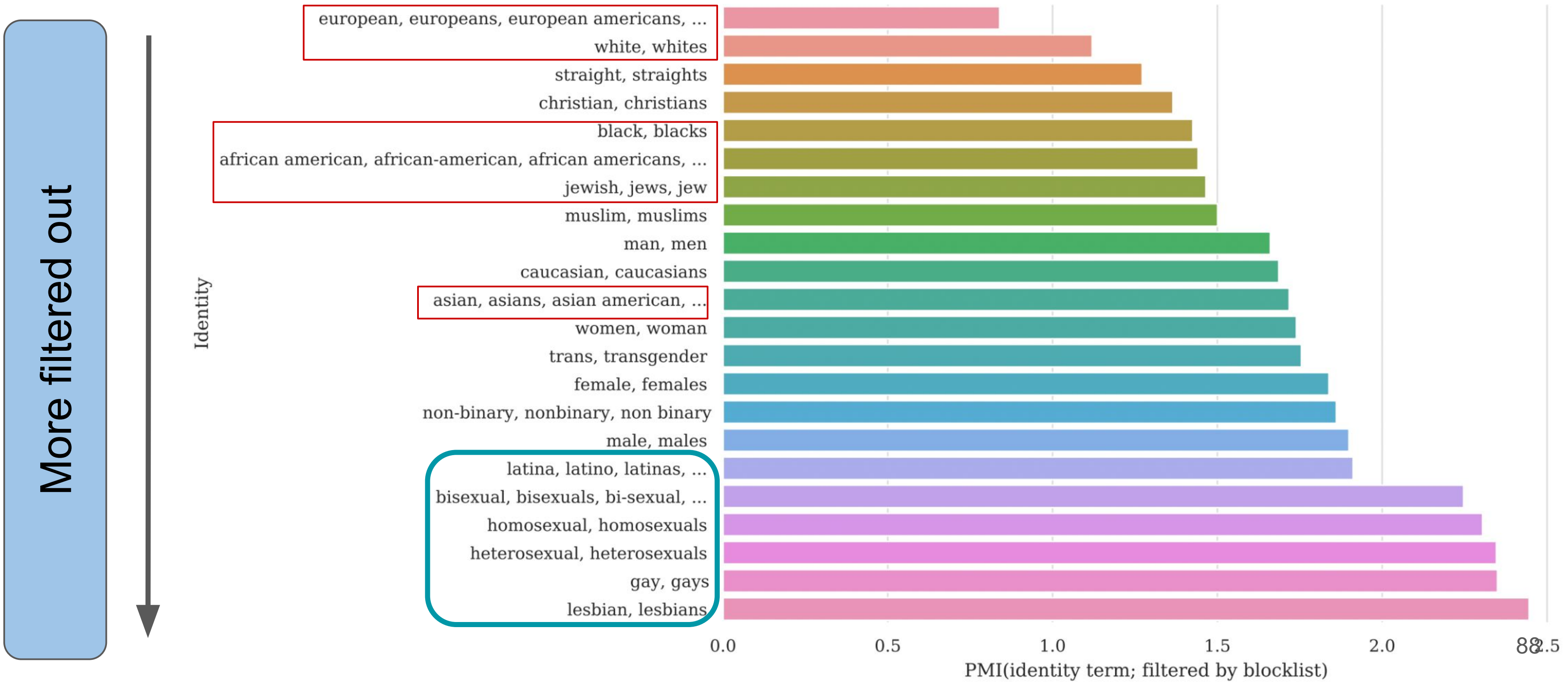
Gender identity
Sexual orientatixon
Race
Religion

Extracted the frequencies of a set of 22 regular expressions

```
homosexuals?
gays?
non[ -]?binary
trans(|\+|gender)
lesbians?
blacks?
african[ -]americans?
latin[oax]s?
asian([ -]american)?s?
muslims?
jew(|s|ish)?
wom[ae]n
females?
m[ae]n
males?
straights?
heterosexuals?
bi-?sexuals?
whites?
caucasians?
european([ -]american)?s?
christians?
```

Computed PMI (Church and Hanks, 1990)

Which demographic identities are excluded?



What is excluded from the corpus?

Characterizing the excluded documents

Which demographic identities are excluded?

Whose English is included?

Whose English is included?

Dialect-aware topic model ([Blodgett el al.,2016](#))

Trained on 60M geolocated tweets

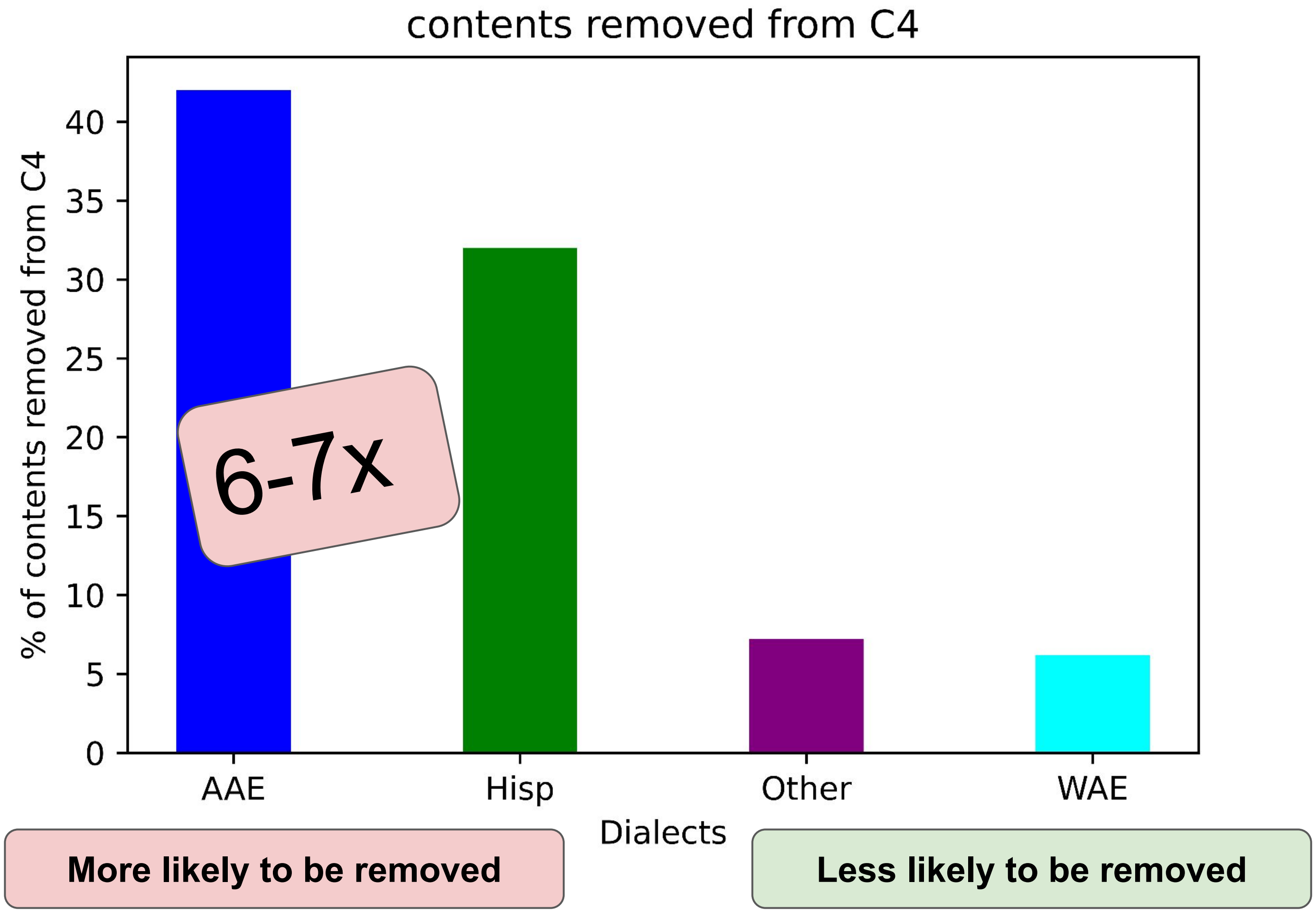
Relies on US census race/ethnicity data as topics

Whose English is included?

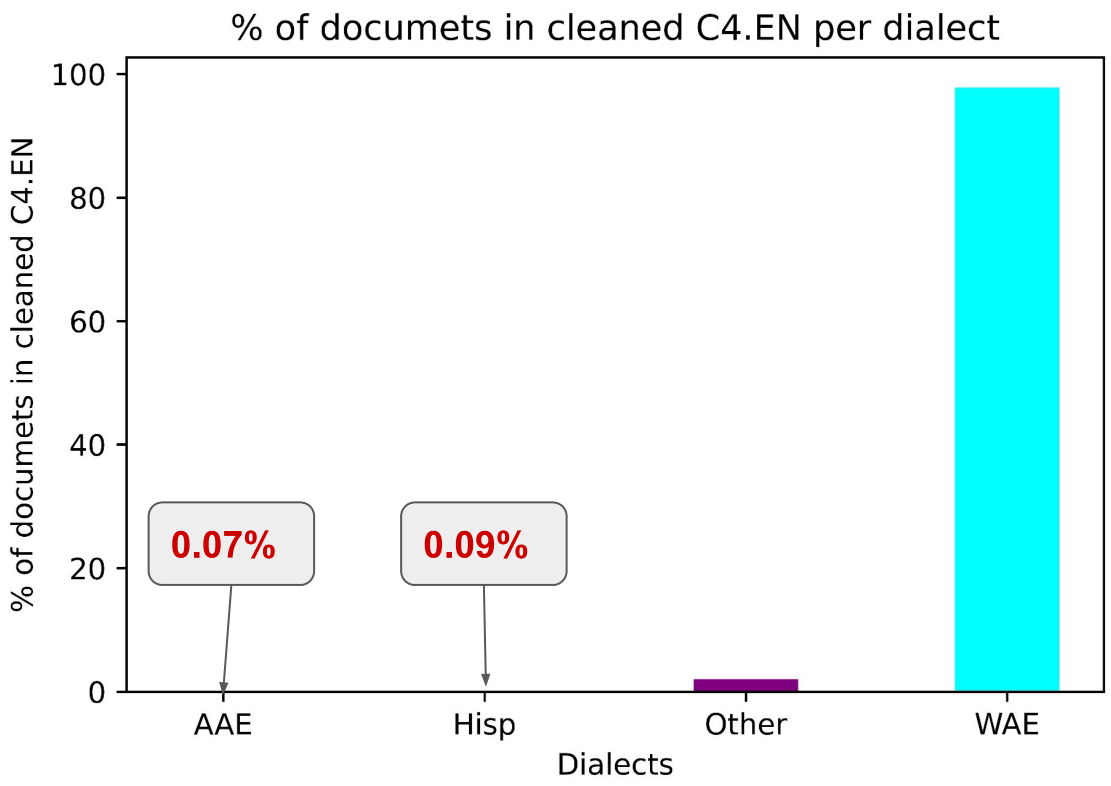
The model yields posterior probabilities of a given document

- African-American-English (AAE)
- Hispanic-aligned English (Hisp)
- White-aligned English (WAE)
- Other English dialects (other)

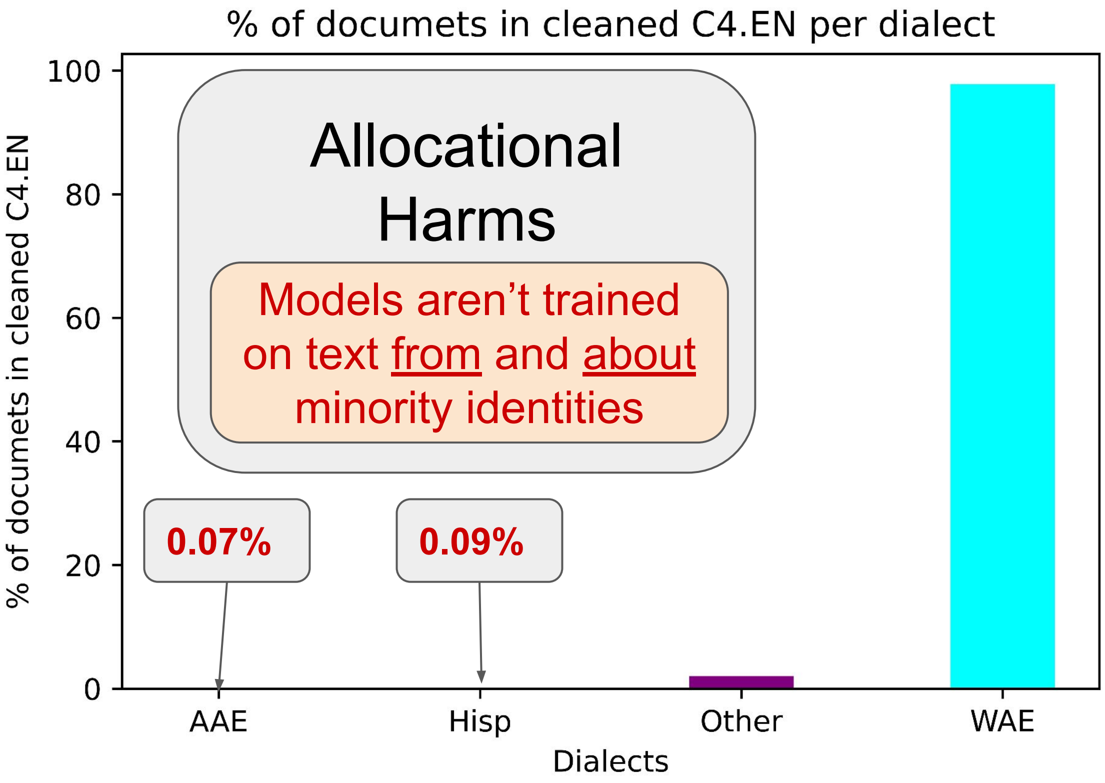
Blodgett el al.,2016: https://aclanthology.org/D16-1120/


contents removed from C4
% of contents removed from C4
0
5
10
15
20
25
30
35
40
6-7x
AAE
Hisp
Other
WAE
Dialects
More likely to be removed
Less likely to be removed

Whose English is included?



Whose English is included?



Documenting Webtext Corpora: Recommendations

- Report metadata
- Examine benchmark contamination
- Social biases and representational harms
- Excluded voices and identities
- Other recommendations

Documenting Webtext Corpora: Recommendations

Report website metadata

Examine benchmark contamination

Social biases and representational harms

Excluded voices and identities

Other recommendations

Reporting website metadata

Report the domains the text is scraped from

Data collection process can lead to a different distribution of internet domains than one would expect

Documenting Webtext Corpora: Recommendations

- Report website metadata
- **Examine benchmark contamination**
- Social biases and representational harms
- Excluded voices and identities
- Other recommendations

Examine benchmark contamination

Support collecting data with the human-in-the-loop

To reduce contamination of future benchmarks

Documenting Webtext Corpora: Recommendations

Report website metadata

Examine benchmark contamination

Social biases and representational harms

Excluded voices and identities

Other recommendations

Social Biases & Representational harms

Control the distributional biases

Select subdomains to use for training

Measurement of bias in each subdomain

Documenting Webtext Corpora: Recommendations

- Report website metadata
- Examine benchmark contamination
- Social biases and representational harms
- **Excluded voices and identities**
- Other recommendations

Excluded voices and identities

Avoid blockilst filtering when constructing datasets from web-crawled data

Some voices and identities might be excluded

Meaning of “bad” words heavily depends on the social context

Documenting Webtext Corpora: Recommendations

Report website metadata

Examine benchmark contamination

Social biases and representational harms

Excluded voices and identities

Other recommendations

Other Recommendations

Datasheets for datasets (Gebru et al., 2018)

Datasheets for Datasets

TIMNIT GEBRU, Black in AI
JAMIE MORGENSTERN, University of Washington
BRIANA VECCHIONE, Cornell University
JENNIFER WORTMAN VAUGHAN, Microsoft Research
HANNA WALLACH, Microsoft Research
HAL DAUMÉ III, Microsoft Research; University of Maryland
KATE CRAWFORD, Microsoft Research

Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor–patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.

Unknown to the authors of the datasheet.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.

Some movie reviews might contain moderately inappropriate or offensive language, but we do not expect this to be the norm.

Does the dataset identify any subpopulations (e.g., by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.

No.

Other Recommendations

Datasheets for datasets (Gebru et al., 2018)

Data statements (Bender & Friedman, 2018)

Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science

Emily M. Bender
Department of Linguistics
University of Washington
ebender@uw.edu

Batya Friedman
The Information School
University of Washington
batya@uw.edu

C. SPEAKER DEMOGRAPHIC Sociolinguistics has found that variation (in pronunciation, prosody, word choice, and grammar) correlates with speaker demographic characteristics (Labov, 1966), as speakers use linguistic variation to construct and project identities (Eckert and Rickford, 2001). Transfer from native languages (L1) can affect the language produced by non-native (L2) speakers (Ellis, 1994, Ch. 8). A further important type of variation is disordered speech (e.g., dysarthria). Specifications include:

- Age
- Gender
- Race/ethnicity
- Native language
- Socioeconomic status
- Number of different speakers represented
- Presence of disordered speech

Other Recommendations

Datasheets for datasets (Gebru et al., 2018)

Data statements (Bender & Friedman, 2018)

Model cards (Mitchell et al., 2018)

Model Cards for Model Reporting

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, Timnit Gebru
{mmitchellai,simonewu,andrewzaldivar,parkerbarnes,lucyvasserman,benhutch,espitzer,tgebru}@google.com
deborah.raji@mail.utoronto.ca

- **Evaluation Data**. Details on the dataset(s) used for the quantitative analyses in the card.
  - Datasets
  - Motivation
  - Preprocessing
- **Training Data**. May not be possible to provide in practice. When possible, this section should mirror Evaluation Data. If such detail is not possible, minimal allowable information should be provided here, such as details of the distribution over various factors in the training datasets.
- **Quantitative Analyses**
  - Unitary results
  - Intersectional results
- **Ethical Considerations**
- **Caveats and Recommendations**

Dataset life cycle: https://stanford-cs324.github.io/winter2022/lectures/data/

Limitations of the paper’s analysis

Only examined some issues: location to report other issues

Tools used work disproportionately well for English: generalization to other languages?

Limitations of the paper’s analysis

Only examined some issues: location to report other issues

Tools used work disproportionately well for English: generalization to other languages?

Search it for yourself!

https://c4-search.apps.allenai.org/

AI2 Allen Institute for AI

C4 Search

This site lets users to execute full-text queries to search Google's C4 Dataset. Our hope is this will help ML practitioners better understand its contents, so that they're aware of the potential biases and issues that may be inherited via it's use.

The dataset is released under the terms of ODC-BY. By using this, you are also bound by the Common Crawl Terms of Use in respect of the content contained in the dataset.

You can read more about the supported query syntax here . Each record has two fields, `url` and `text`, both of which are searchable. The fields are indexed using the Standard analyzer, which means you can't search for punctuation.

Tanushree Banerjee

Search

Found more than 10,000 results in 1.09 seconds

https://mai.wikipedia.org/wiki/%E0%A4%A4%E0%A4%A8%E0%A5%81%E0%A4%B6%E0%A5%8D%E0%A4%B0%E0%A5%80_...

↑ "Tanushree At The Miss Universe 2004". Times of India. http://feminamissindia.indiatimes.com/articleshow/msid-707513,curpg-1.cms. अन्तिम पहुँच तिथि: 14 February 2010. ↑ Piali Banerjee (27 March 2004). "Tanushree Crowned Pond's Femina Miss India". Times of India. http://feminamissindia.indiatimes.com/articleshow/586465.cms. अन्तिम पहुँच तिथि: 14 February 2010.

T5 (Raffel et al., 2019)

Switch Transformer (Fedus et al., 2021)

The Pile: An 800GB Dataset of Diverse Text for Language Modeling

Leo Gao **Stella Biderman** **Sid Black** **Laurence Golding**

Travis Hoppe **Charles Foster** **Jason Phang** **Horace He**

Anish Thite **Noa Nabeshima** **Shawn Presser** **Connor Leahy**

EleutherAI

Motivation

Dataset diversity leads to better downstream generalization (Rosset, 2019)

LLMs shown to acquire knowledge in new domain with relatively small training data

⇒ large number of smaller high quality datasets may improve cross domain knowledge + generalization

Motivation

Dataset diversity leads to better downstream generalization (Rosset, 2019)

LLMs shown to acquire knowledge in new domain with relatively small training data

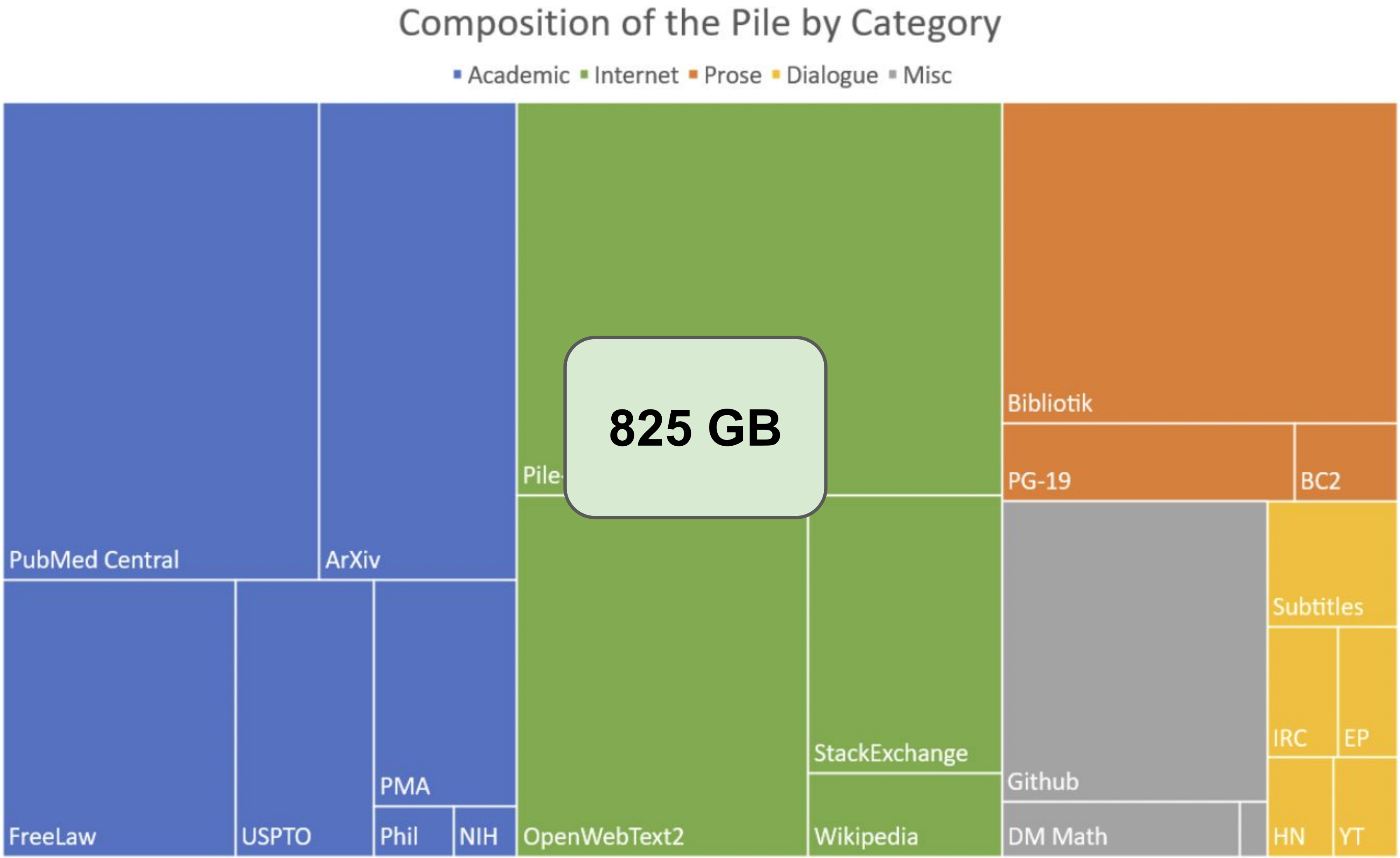
⇒ large number of smaller high quality datasets may improve cross domain knowledge + generalization

Motivation

Dataset diversity leads to better downstream generalization (Rosset, 2019)

LLMs shown to acquire knowledge in new domain with relatively small training data

⇒ large number of smaller high quality datasets may improve cross domain knowledge + generalization


Composition of the Pile by Category
Academic
Internet
Prose
Dialogue
Misc
PubMed Central
ArXiv
FreeLaw
USPTO
PMA
Phil
NIH
Pile-
825 GB
OpenWebText2
StackExchange
Wikipedia
Bibliotik
PG-19
BC2
Github
DM Math
Subtitles
IRC
EP
HN
YT

Relative component-wise GPT-3 Pile Performance

Components GPT-3 underperforms on

≈ Pile components most dissimilar to GPT-3 pre-training corpus

= Good candidates for supplementing GPT-3 pre-training data

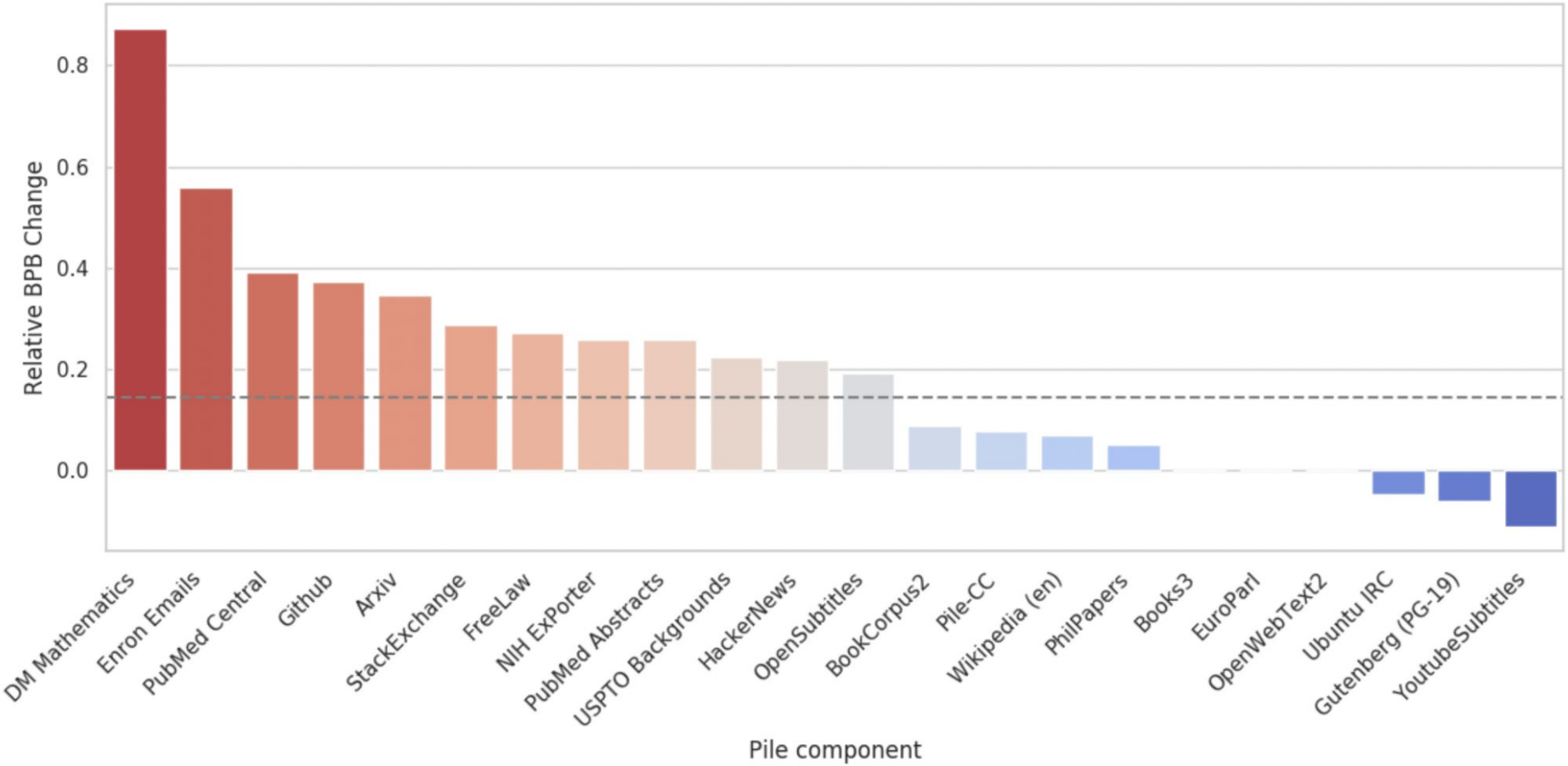
How to compare? A proxy measure

GPT-2 trained from scratch on Pile vs original GPT-3

1. measure improvement from GPT2-Pile to GPT-3 on each component
2. normalize by setting change on OpenWebText2 to be zero.

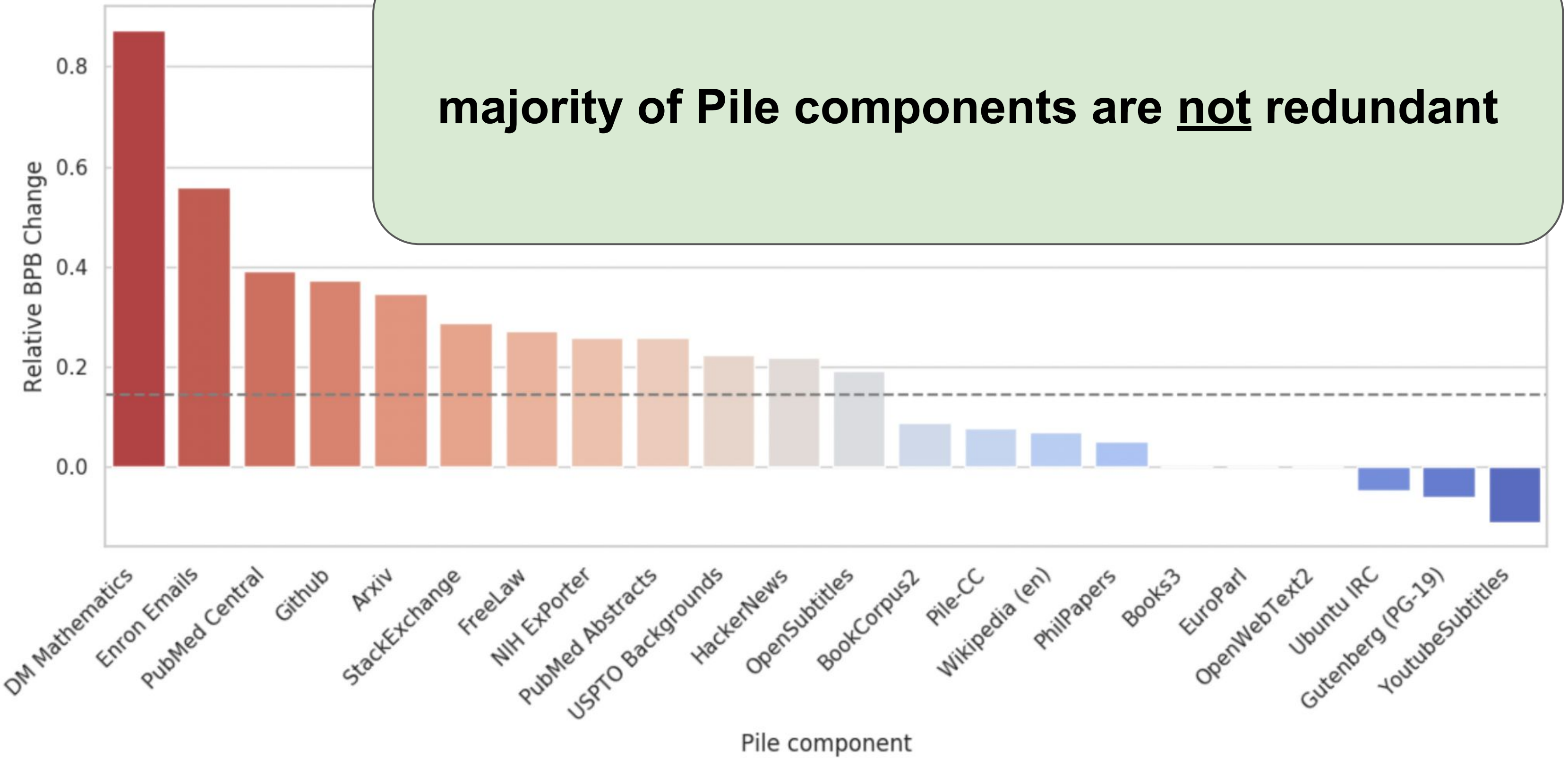
$$\begin{aligned}\Delta_{\text{set}} = &\overbrace{\left(L_{\text{set}}^{\text{GPT3}} - L_{\text{owt2}}^{\text{GPT3}}\right)}^{\textbf{How much harder set was for GPT-3 than owt2}} \\ &- \underbrace{\left(L_{\text{set}}^{\text{GPT2Pile}} - L_{\text{owt2}}^{\text{GPT2Pile}}\right)}_{\textbf{difference in the intrinsic difficulty of \textit{set} and \textit{owt2}}}\end{aligned}$$

Difference in difficulty →



Difference in difficulty →

majority of Pile components are not redundant



GPT-2 Pile wins

Original GPT-3 wins

Key takeaways: The Pile paper

Training on dataset sourced from **smaller, higher quality** sources outperforms training on web-crawled data

Analysis of pejorative content, gender/religion biases: qualitatively similar to previous work.

Deduplicating Training Data Makes Language Models Better

Katherine Lee*† Daphne Ippolito*†‡ Andrew Nystrom† Chiyuan Zhang†

Douglas Eck† Chris Callison-Burch‡ Nicholas Carlini†

Motivation

Existing language modeling datasets contain many near-duplicate examples

Long repetitive substrings (**%3 of C4** → **10M documents**)

This encourages **memorization** and discourages **generalization**

Table is from Lee et al, 2022

near-duplicates

Dataset	Example	Near-Duplicate Example
C4	Affordable and convenient holiday flights take off from your departure country, "Canada". From May 2019 to October 2019, Condor flights to your dream destination will be roughly 6 a week! Book your Halifax (YHZ) - Basel (BSL) flight now, and look forward to your "Switzerland" destination!	Affordable and convenient holiday flights take off from your departure country, "USA". From April 2019 to October 2019, Condor flights to your dream destination will be roughly 7 a week! Book your Maui Kahului (OGG) - Dubrovnik (DBV) flight now, and look forward to your "Croatia" destination!

Example of near-duplicates in C4 dataset

Deduplication Approaches

EXACTSUBSTR: **Exact Substr**ing

NEARDUP: **Near Dup**licates

ExactSubstr

If two examples ***a*** and ***b*** share a **substring** of at least **50 tokens**

Then remove that substring from either ***a*** or ***b***

NearDup

Given two documents $\boldsymbol{x_i}$ and $\boldsymbol{x_j}$

Let $\boldsymbol{d_i}$ and $\boldsymbol{d_j}$ be the set of n-grams of $\boldsymbol{x_i}$ and $\boldsymbol{x_j}$, respectively

NEARDUP

$$\text{Jaccard}(d_i, d_j) = {|d_i \cap d_j|}/{|d_i \cup d_j|}$$

Jaccard Index
(Jaccard, 1912)

$$\Pr(d_i, d_j \,|\, \text{Jaccard}(d_i, d_j) = s_{i,j}) = 1 - (1 - s_{i,j}^b)^r$$

If above 0.8,
x_i and *x_j* are a
potential match.
b=20, r=450

$$\text{EditSim}(x_i, x_j) = 1 - \frac{\text{EditDistance}(x_i, x_j)}{\max(|x_i|, |x_j|)}$$

If above 0.8, *x_i* and
x_j are **duplicates**

Results

Train-test overlap, a **61-word sequence** that is **repeated 61,036 times** in C4 **training** and **61 times** in **validation sets**

by combining fantastic ideas, interesting arrangements, and follow the current trends in the field of that make you more inspired and give artistic touches. We'd be honored if you can apply some or all of these design in your wedding. believe me, brilliant ideas would be perfect if it can be applied in real and make the people around you amazed!

Results

Train-test overlap, a **61-word sequence** that is **repeated 61,036 times** in **training set** and **61 times** in **validation set**

Deduplicating the training set **reduces** the rate of emitting **memorized training data** by a factor of **10 times**

Found that **training models** on **deduplicated datasets** is **more efficient**

Found that **deduplicating** training data **does not hurt perplexity**

Key takeaways: Deduplicating paper

Duplicates between the training and testing sets encourage the model to memorize the training data

Deduplication does not harm, and sometimes improves model perplexity

Deduplication makes the training faster

Key takeaways: Deduplicating paper

Duplicates between the training and testing sets encourage the model to memorize the training data

Deduplication does not harm, and sometimes improves model perplexity

Deduplication makes the training faster

Key takeaways: Deduplicating paper

Duplicates between the training and testing sets encourage the model to memorize the training data

Deduplication does not harm, and sometimes improves model perplexity

Deduplication makes the training faster

More detailed summary: https://twitter.com/katherine1ee/status/1415496898241339400

Summary

Dodge et al., 2021 (main paper)

Propose three levels of documentation for web-crawled datasets

Recommendations for future documentation efforts

Gao et al., 2020 (The Pile paper)

New dataset combining 22 high quality, diverse sources

Lee et al., 2022 (The Deduplication paper)

Deduplication does not harm the perplexity and makes the training faster

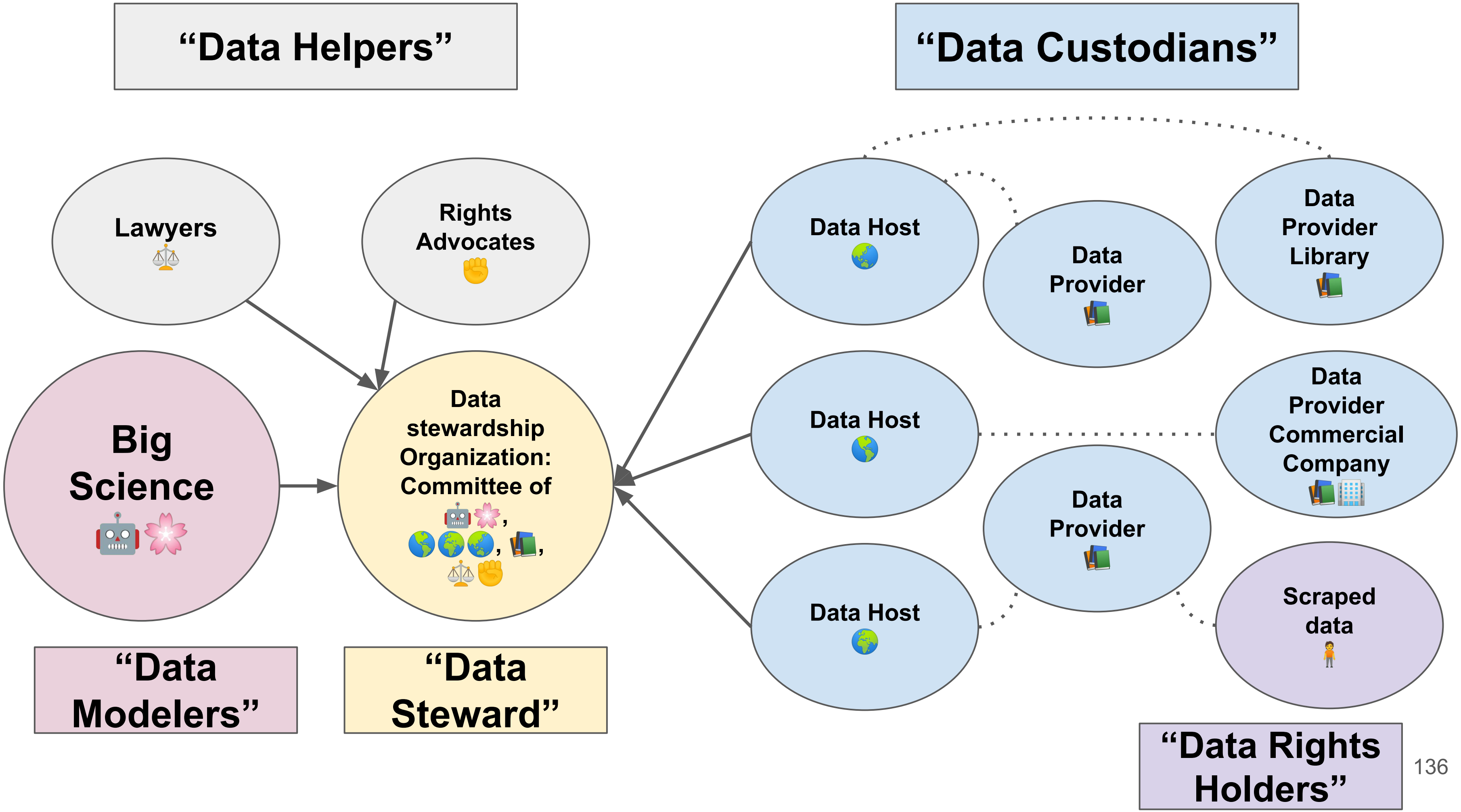
High-level discussion points

So far only focussed on existing datasets and documentation… other angles?

What should the ecosystem where data is created and used look like?

Best practices for creating data to maintain quality and security?

Idea - data belongs to groups rather than individuals


"Data Helpers"
"Data Custodians"
Lawyers ⚖️
Rights Advocates ✊
Data Host 🌏
Data Provider 📚
Data Provider Library 📚
Big Science 🤖🌸
Data stewardship Organization: Committee of 🤖🌸, 🌎🌍🌏, 📚, ⚖️✊
Data Host 🌎
Data Provider Commercial Company 📚🏢
Data Provider 📚
Data Host 🌍
Scraped data 🧍
"Data Modelers"
"Data Steward"
"Data Rights Holders"

Key Takeaways

A lot of data available on Web – Training on “all of it” not most efficient

Filtering/curation needed, but results in biases

Transparent documentation needed:
Data creators: reflect on decisions, potential harms
Data consumers: know when dataset can/can’t be used

Curating non-web high quality datasets is promising (The Pile)

When creating new dataset from the Web, remember that pretrained models may already have seen this data

Thank you!

References - content

Stanford course: Data | CS324

Documenting Large Webtext Corpora: A Case Study on the Colossal Clean Crawled Corpus

EMNLP 2021 video

The Pile: An 800GB Dataset of Diverse Text for Language Modeling

Deduplicating Training Data Makes Language Models Better

Minority Voices 'Filtered' Out of Google Natural Language Processing Models - Unite.AI

The Efforts to Make Text-Based AI Less Racist and Terrible | WIRED

References - images

10 Practical Text Mining Examples to Leverage Right Now | Expert.ai

Oracle Fusion Middleware – www.mwidm.com

Document free icon

Stack Of Books Pictures, Images and Stock Photos

Comparative size of datasets used for training NLP models (represented... | Download Scientific Diagram

https://patentimages.storage.googleapis.com/0c/21/16/3b2ad21579ae2a/CN1199926A.pdf

<u>Pre-lecture Q3</u>

Dodge et al. remark that "Documenting massive, unlabeled datasets is a challenging enterprise" and they mainly consider simple corpus statistics and metadata.

Can you think of other properties/aspects that we should document and examine in the data?

What (NLP) techniques can we use to document and query data in more detail?