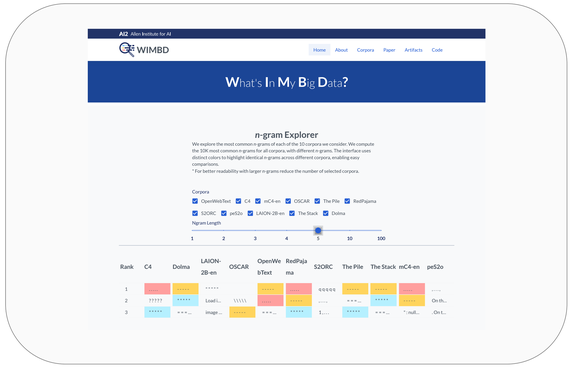
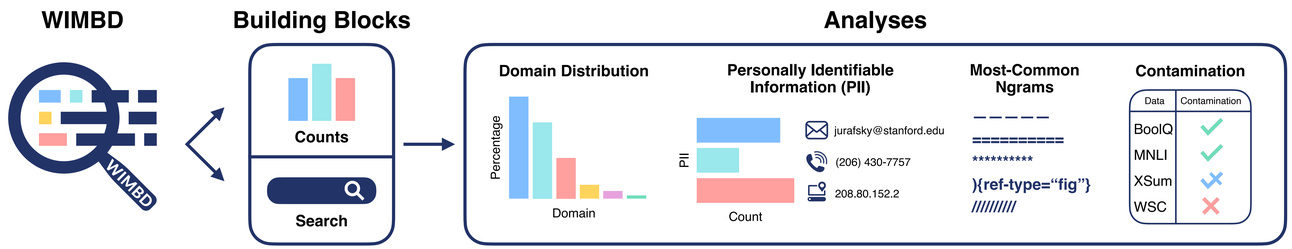
What's In My Big Data?

Yanai Elazar, Akshita Bhagia, Ian Magnusson, Abhilasha Ravichander, Dustin Schwenk, Alane Suhr, Pete Walsh, Dirk Groeneveld, Luca Soldaini, Sameer Singh, Hanna Hajishirzi, Noah A. Smith, Jesse Dodge

Demo





(1) Motivation

- Datasets are the foundation of ML models
- To understand model behavior, we must understand their underlying data
- How do we analyze the contents of terabytes of unstructured text data?!

(2) The Platform

- Search
- Counts
  - Tokens
  - Domains
  - ...

elasticsearch

hadoop MapReduce

```
from wimbd.es import count_documents_containing_phrases

count_documents_containing_phrases("c4", "artificial intelligence")
# 6,065,714
```

(3) Datasets & Analyses

Corpus	Model	Size (GB)	# Documents
OpenWebText	GPT-2*	41.2	8,005,939
C4	T5	838.7	364,868,892
mC4-en	umT5	14,694.0	3,928,733,374
OSCAR	BLOOM*	3,327.3	431,584,362
The Pile	GPT-J/Neo & pythia	1,369.0	210,607,728
RedPajama	LLaMA*	5,602.0	930,453,833
S2Orc	SciBERT*	692.7	11,241,499
peS2o	-	504.3	8,242,162
LAION-2B-en	Stable Diffusion*	570.2	2,319,907,827
The Stack	StarCoder*	7,830.8	544,750,672

1. Data Statistics
a. High-level statistics
b. Internet domains distribution
c. Dates distribution

2. Data Quality
a. Common n-grams
b. Duplicates

3. Community/Society Measurements
a. Contamination
b. PII

4. Cross-data Analysis
a. Distributional similarity
b. Overlapping documents

More analyses in the paper!

(4) Results

(i) Most common n-grams

OpenWebText

n-gram	Count
----------	3.4M
. . . . .	1.05M
= = = = = = = = = =	830K
* * * * * * * * * *	595K
##########	302K
amp ; amp ; amp ; amp ; amp ;	278K
; amp ; amp ; amp ; amp ; amp	265K
— — — — — — — — — —	249K
… … … … … … … … … …	88.1K
~ ~ ~ ~ ~ ~ ~ ~ ~ ~	83.3K

C4

n-gram	Count
? ? ? ? ? ? ? ? ? ?	9M
. . . . .	7.27M
----------	4.41M
* * * * * * * * * *	3.87M
! ! ! ! ! ! ! ! ! !	1.91M
. You can follow any responses to this entry through	784K
❓❓❓❓❓❓❓❓❓❓	753K
You can follow any responses to this entry through the	752K
can follow any responses to this entry through the RSS	752K
follow any responses to this entry through the RSS 2.0	748K

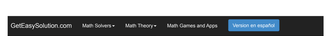
LAION-2B-en

n-gram	Count
----------	1.65M
❓❓❓❓❓❓❓❓❓❓❓	1.43M
. . . . .	1.15M
\ \ \ \ \ \ \ \ \ \	809K
< br / > < br / > < br	797K
/ > < br / > < br / >	796K
br / > < br / > < br /	796K
> < br / > < br / > <	576K
l Price : 1 Credit (USD $ 1)	437K
vector l Price : 1 Credit (USD $ 1	437K

The Stack

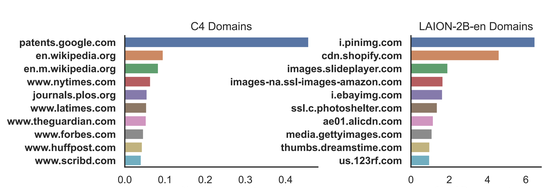
n-gram	Count
----------	4.29B
* * * * * * * * * *	3.87B
0 0 0 0 0 0 0 0 0 0	2.75B
= = = = = = = = = =	2.62B
, " resolved " : " https : / /	1.46B
" , " resolved " : " https : /	1.46B
" resolved " : " https : / / registry.npmjs.org	1.42B
resolved " : " https : / / registry.npmjs.org /	1.42B
, , , , , , , , , ,	1B
. tgz " , " integrity " : " sha512	938M

(ii) Internet-domain distribution

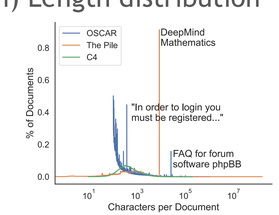


What is 2.343 percent of 280000 - step by step solution

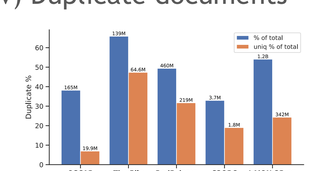
Domain	Corpus	Rank	Tokens	% of All Tokens
www.geteasysolution.com	C4	473082	49,859	0.000032%
www.geteasysolution.com	RedPajama	472159	49,859	0.000023%
www.geteasysolution.com	mC4-en	1658921	156,174	0.0000056%



(iii) Length distribution



(iv) Duplicate documents



(v) Personally Identifiable Information (PII)

Corpus	Email Addresses Count	Email Addresses Prec.	Phone Numbers Count	Phone Numbers Prec.	IP Addresses Count	IP Addresses Prec.
OpenWebText	364K	*99*	533K	*87*	70K	*54*
OSCAR	62.8M	*100*	107M	*91*	3.2M	*43*
C4	7.6M	*99*	19.7M	*92*	796K	*56*
mC4-en	201M	*92*	4B	*66*	97.8M	*44*
The Pile	19.8M	*43*	38M	*65*	4M	*48*
RedPajama	35.2M	*100*	70.2M	*94*	1.1M	*30*
S2ORC	630K	*100*	1.4M	*100*	0K	*0*
peS2o	418K	*97*	227K	*31*	0K	*0*
LAION-2B-en	636K	*94*	1M	*7*	0K	*0*
The Stack	4.3M	*53*	45.4M	*9*	4.4M	*55*

(vi) Benchmark contamination