Common Crawl

A Web Worth of Data: Common Crawl for NLP

Text By The Bay

April 24, 2015

Common Crawl

It's a non-profit that makes

web data

freely accessible to anyone

Common Crawl

Each crawl archive is billions of pages:

February crawl archive is
1.9 billion web pages
~154 terabytes uncompressed

Common Crawl

Released

totally free

without additional intellectual property restrictions

(lives on Amazon Public Data Sets)

Common Crawl File Formats

- WARC
  + Raw HTTP response headers
  + Raw HTTP responses
- WAT
  + HTML head data
  + HTTP header fields
  + Extracted links / script tags
- WET
  + Extracted text

Common Crawl WET

WET (Web Extracted Text) is released in the crawl archive each month

Data attempts to cover widest range of use cases

No distinction between header / navigation / content:

- Does not remove boilerplate
- Does not re-format text as appears in browser

Origins of Common Crawl

Common Crawl founded in 2007
by Gil Elbaz (Applied Semantics / Factual)

Google and Microsoft were the powerhouses

Goal: Democratize and simplify access to
"the web as a dataset"

Open Data and Open Source

Data *powers* the algorithms in our field

How can we have an even playing field for innovation without access to such data?
(Can you replicate work without the data..?)

More data can beat better algorithms
(Banko and Brill, 2001)

Common Crawl for NLP

The web is largely unannotated,
so how are people using it for NLP?

(a) Use extracted text for unsupervised algorithms

(b) Filter it into being semi-annotated or annotated
(big data ⇒ filter ⇒ curated smaller dataset)

Examples of Previous Work

Unsupervised Algorithms

+ N-gram & language models
+ GloVe: Global Vectors for Word Representation

Filtering

+ Web tables for gazetteers
+ Dirt Cheap Web-Scale Parallel Text
+ Extracting US phone numbers

N-gram Counts & Language Models from the Common Crawl

Christian Buck, Kenneth Heafield, Bas van Ooyen

(Edinburgh, Stanford, Owlin BV)

Processed all the text of Common Crawl to produce
975 billion deduplicated tokens

Google N-gram Dataset (Web 1T) consists of
1 billion tokens

N-gram Counts & Language Models from the Common Crawl

Improvement over Google N-grams (2006):

- Inclusion of low count entries
- Deduplication to reduce boilerplate

"Google has shared a deduplicated version ...
in limited contexts, but it was never publicly released."

-- N-gram Counts & Language Models from the Common Crawl (Buck et al.)

N-gram Counts & Language Models from the Common Crawl

"The advantages of structured text do not outweigh the extra computing power needed to process them."

-- N-gram Counts & Language Models from the Common Crawl (Buck et al.)

N-gram Counts & Language Models from the Common Crawl

English (23TB), German (1.02TB), Spanish (986GB), French (912GB), Japanese (577GB), Russian (537GB), Polish (334GB), Italian (325GB) ...

Only 0.14% of the corpus was Finnish, yet yielded a useful corpus of 47GB.

42 languages with >10GB

73 languages with >1GB

N-gram & Language Models

Sentence level deduplication led to a removal of 80% of the English corpus, lower for other languages (in line with Bergsma et al. (2010))

Before preprocessing (English): 23.62 TB

After preprocessing (English): 5.14 TB

(59 billion lines, 975 billion tokens)

N-gram & Language Models

Substantial improvement in perplexity

Corpus		Perplexity include	Perplexity OOVs
Europarl		620.58	1902
United Nations		484.47	863
Giga Fr-En (English)		303.08	355
Common Crawl parallel		299.43	418
News Commentary		696.20	2568
News	2009	158.32	346
	2010	172.41	394
	2011	149.38	275
	2012	139.59	235
	2013	113.74	122
All interpolated		93.81	46
This work		**58.55**	**5**

Table 5: Perplexities on English newstest 2014.

N-gram & Language Models

"...even though the web data is quite noisy even limited amounts give improvements."

	BLEU			
	2012	Δ	2013	Δ
Baseline	35.8		30.9	
+ 50M lines	36.3	0.5	31.5	0.6
+ 100M lines	36.5	0.7	31.5	0.6
+ 200M lines	36.6	0.8	31.8	0.9
+ 400M lines	37.0	1.2	31.8	0.9
+ 800M lines	37.3	1.6	31.8	0.9
+ 1.3B lines	37.7	1.9	32.0	1.1

Table 9: Machine Translation performance for English-Spanish on newstest 2012/2013 using increasing amounts of data for the additional language model.

N-gram & Language Models

Project data was released at
http://statmt.org/ngrams

- Raw text split by language
- Deduped text split by language
- Resulting language models

GloVe: Global Vectors for Word Representation

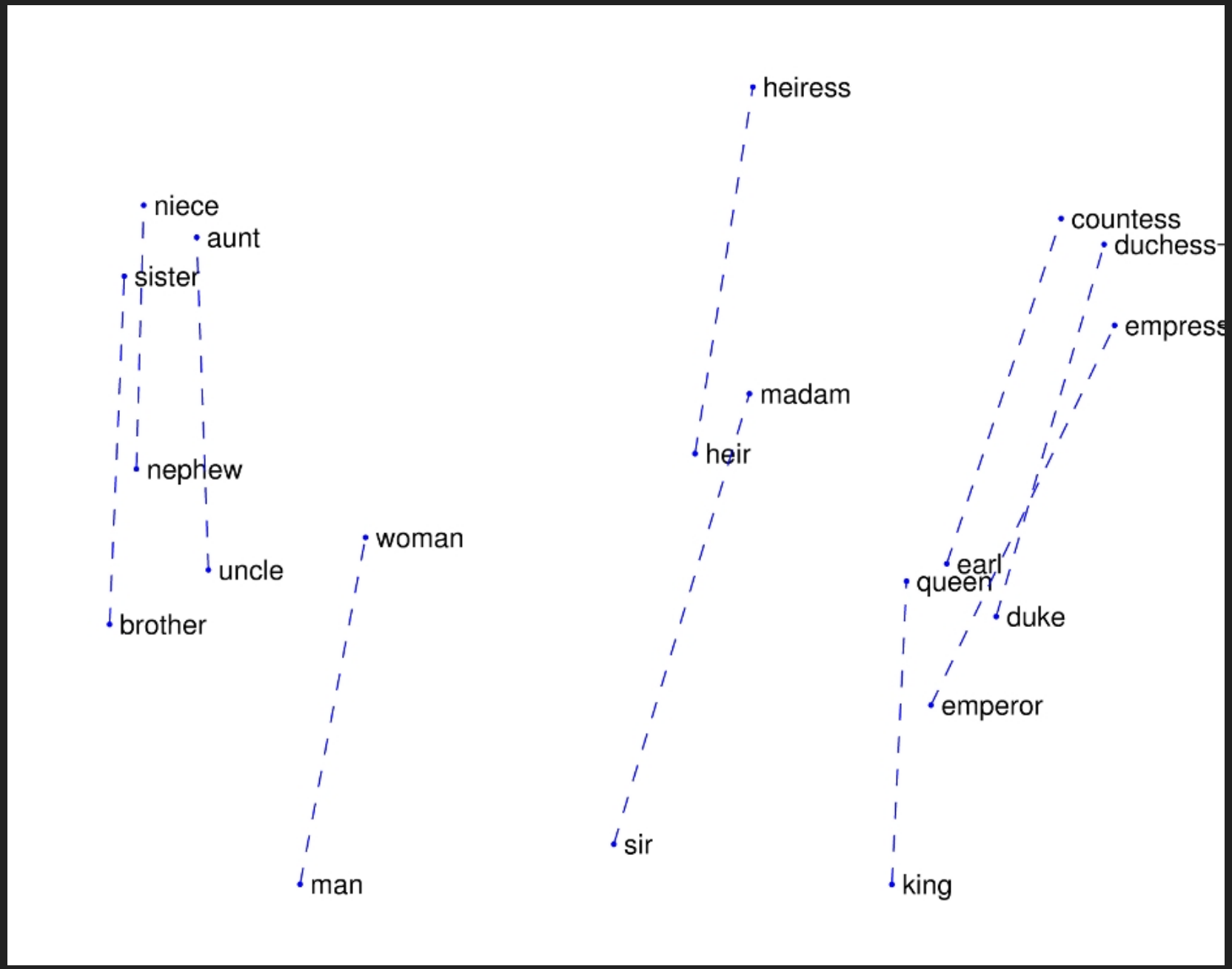
Jeffrey Pennington, Richard Socher, Christopher D. Manning

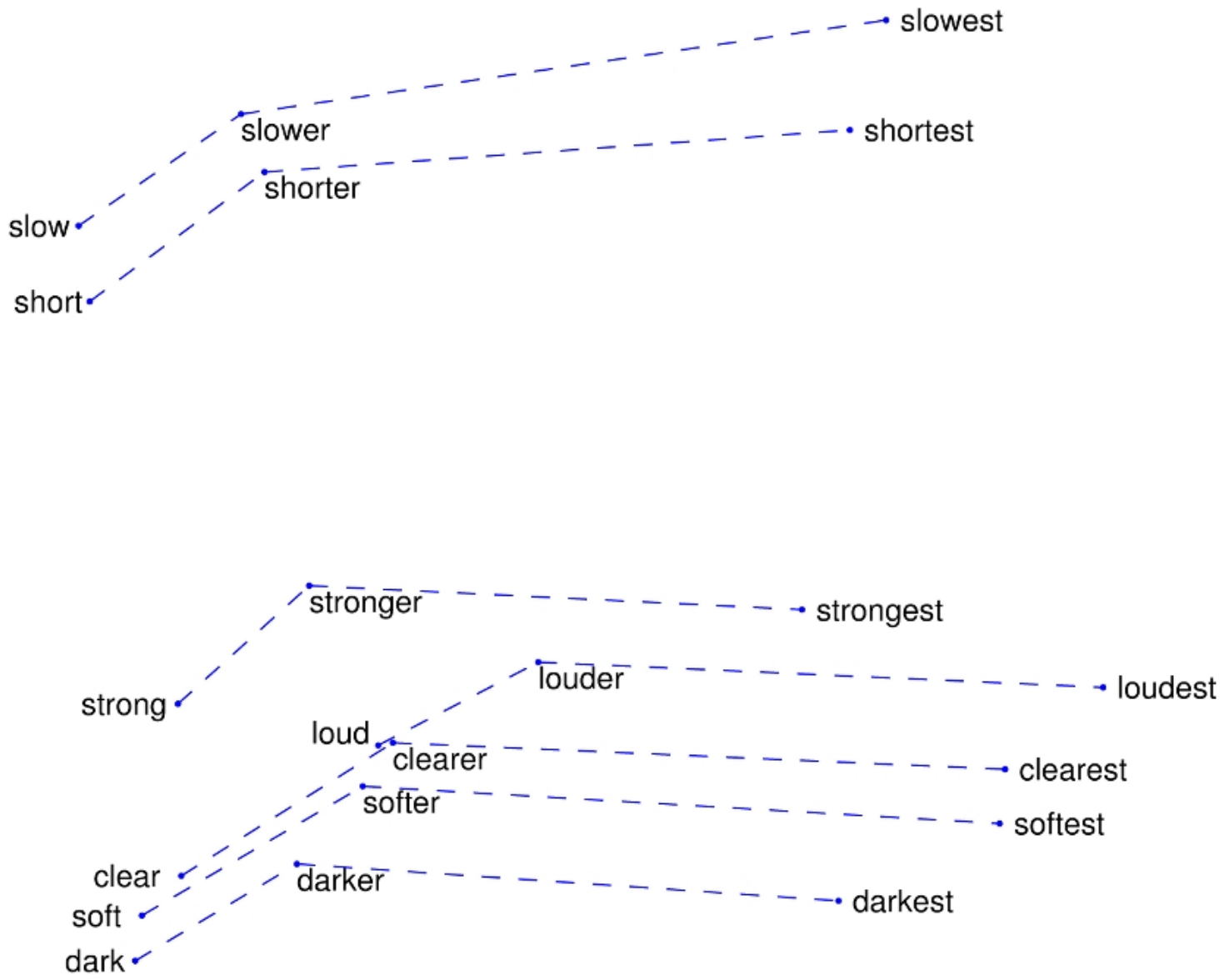
Word vector representations:

king - queen = man - woman

king - man + woman = queen

(produces dimensions of meaning)


niece
aunt
sister
nephew
uncle
brother
woman
man
heiress
madam
heir
sir
countess
duchess
empress
earl
queen
duke
emperor
king


slowest
slower
shortest
shorter
slow
short
stronger
strongest
louder
loudest
strong
loud
clearer
clearest
softer
softest
clear
darker
darkest
soft
dark

GloVe: Global Vectors for Word Representation

Trained on non-zero entries of a global word-word co-occurrence matrix

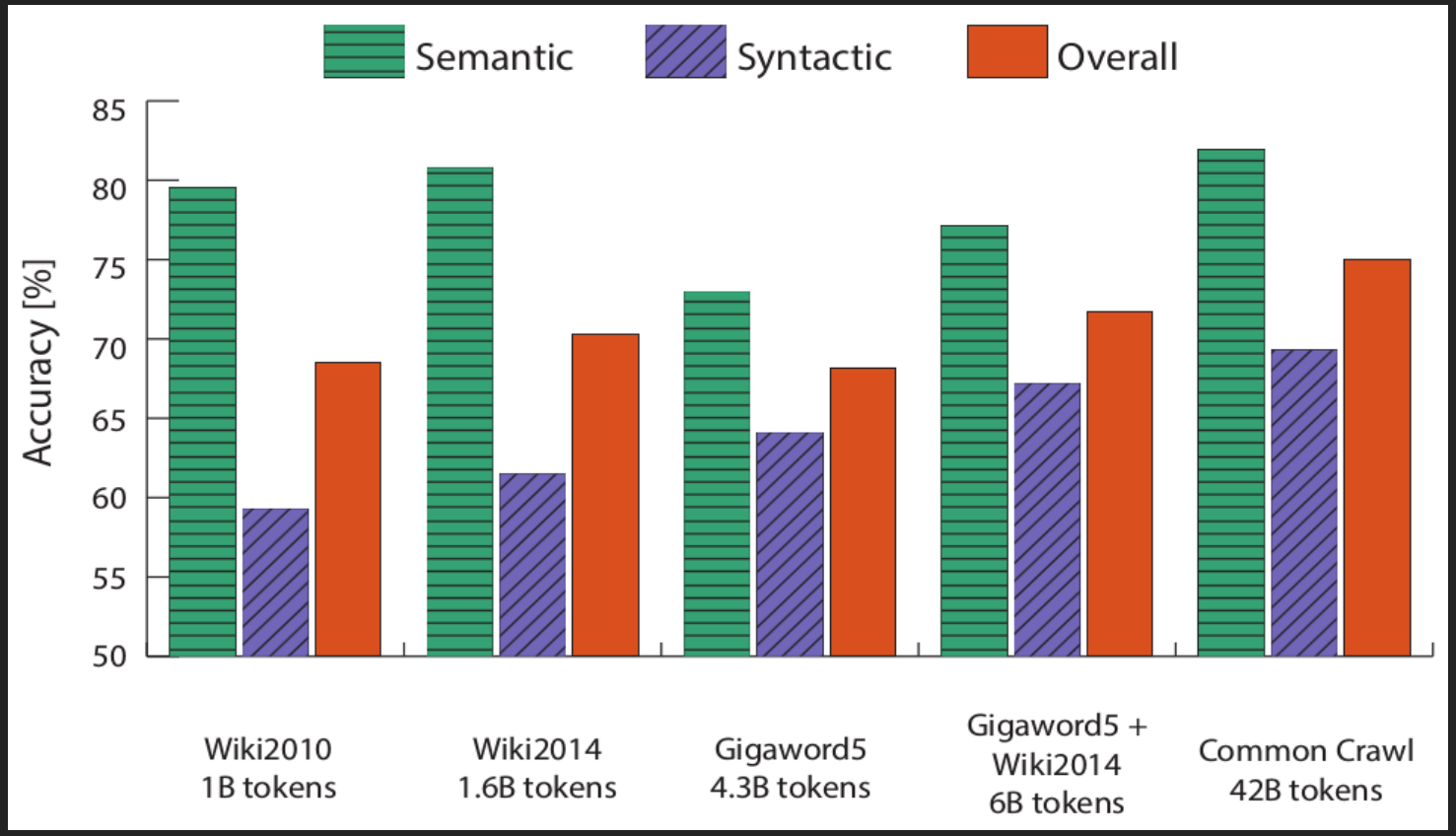
Populating matrix requires a single pass
Subsequent training is far faster

$$\text{GloVe} = O(|C|^{0.8})$$

$$\text{On-line window-based (i.e. word2vec)} = O(|C|)$$

GloVe On Various Corpora

Semantic: "Athens is to Greece as Berlin is to _?"
Syntactic: "Dance is to dancing as fly is to_?"



GloVe over Big Data

GloVe using 42 billion tokens from Common Crawl outperformed word2vec w/ 100 billion tokens (Google News)

Largest GloVe model to prove scalability uses **840 billion tokens** from Common Crawl

Source code and pre-trained models at http://www-nlp.stanford.edu/projects/glove/

Mix and Match: Word Vectors

- More data, less fine tuning needed
- Best model: mix of all excl. word2vec

Input embeddings	DEV-NO-F	DEV-F	CV-NO-F	CV-F	TEST-NO-F	TEST-F
CW-50	65.83	78.39	63.59	74.71	66.62	76.03
GloVe-300	74.17	77.81	72.89	76.57	76.05	75.87
HPCA-200	61.45	77.14	70.58	76.66	64.56	76.00
word2vec-300	71.46	73.54	69.07	71.93	71.38	71.59
random embeddings	-	74.17	-	64.54	-	71.43
CW-50+GloVe-300+HPCA-200	**79.01**	**79.48**	**76.20**	**77.70**	**78.14**	**77.12**

Automatic Noun Compound Interpretation using Deep Neural Networks and Word Embeddings (Dima and Hinrichs, 2015)

Examples of Previous Work

Unsupervised Algorithms

+ N-gram & language models
+ GloVe: Global Vectors for Word Representation

Filtering

+ Web tables for gazetteers
+ Dirt Cheap Web-Scale Parallel Text
+ Extracting US phone numbers

Gazetteers for NER

+ Want the widest variety of topics possible
+ Aim to keep them modern / up to date
+ Capture relationships between similar words (disambiguation)

Google Sets

Web tables as a source of gazetteers + relations

Querying ["**cat**"],
returns ["**dog**", "**bird**", "**horse**", "**rabbit**", ...]

Querying ["**cat**", "**ls**"],
returns ["**cd**", "**head**", "**cut**", "**vim**", ...]

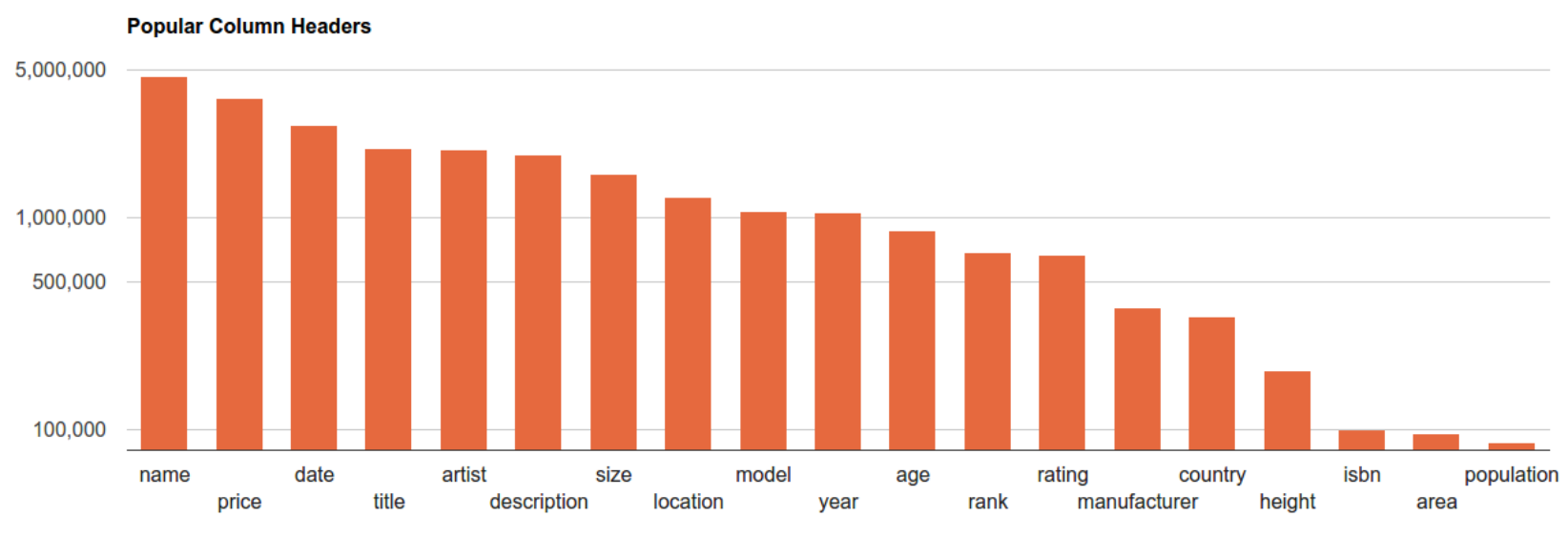
Web Data Commons Web Tables

Extracted 11.2 billion tables from WARC files,
filtered to keep relational tables via trained classifier

Only 1.3% of the original data was kept,
yet it still remains hugely valuable

Resulting dataset:
11.2 billion tables ⇒ 147 million relational web tables

Web Data Commons Web Tables



Popular column headers: name, title, artist, location, model, manufacturer, country …

Released at webdatacommons.org/webtables/

Web Data Commons Web Tables

Movies		Camera Models		Music Albums		Footballers	
Name	#Tables	Name	#Tables	Name	#Tables	Name	#Tables
avatar	11080	nikon d 200	1390	thriller	4268	robin van persie	7439
inception	8121	canon eos 20 d	480	aftermath	2466	david beckham	3041
taxi	6292	canon eos 40 d	355	twist shout	2017	cristiano ronaldo	2927
titanic	4270	nikon d 5000	351	true blue	1737	lionel messi	1748
fantastic four	2113	canon eos 30 d	346	like prayer	1616	ronaldo	1716
moulin rouge	1616	nikon d 80	339	like virgin	1414	gareth bale	1708
black knight	1298	canon eos 50 d	304	yellow submarine	1405	fernando torres	1641
deception	1286	nikon d 90	274	dark side moon	1201	frank lampard	1461
minority report	1201	canon eos 10 d	248	abbey road	971	thierry henry	1332
ice age	1201	nikon d 60	233	something new	919	ronaldinho	1195
unfaithful	1179	nikon d 100	191	please please me	886	roberto carlos	817
glitter	943	canon eos d 30	172	shine light	833	xabi alonso	735

Web-Scale Parallel Text

Dirt Cheap Web-Scale Parallel Text from the Common Crawl (Smith et al.)

Processed all text from URLs of the style:
website.com/[langcode]/

"...nothing more than a set of common two-letter language codes ... [we] mined 32 terabytes ... in just under a day"

Web-Scale Parallel Text

	French	German	Spanish	Russian	Japanese	Chinese
Segments	10.2M	7.50M	5.67M	3.58M	1.70M	1.42M
Source Tokens	128M	79.9M	71.5M	34.7M	9.91M	8.14M
Target Tokens	118M	87.5M	67.6M	36.7M	19.1M	14.8M
	Arabic	**Bulgarian**	**Czech**	**Korean**	**Tamil**	**Urdu**
Segments	1.21M	909K	848K	756K	116K	52.1K
Source Tokens	13.1M	8.48M	7.42M	6.56M	1.01M	734K
Target Tokens	13.5M	8.61M	8.20M	7.58M	996K	685K
	Bengali	**Farsi**	**Telugu**	**Somali**	**Kannada**	**Pashto**
Segments	59.9K	44.2K	50.6K	52.6K	34.5K	28.0K
Source Tokens	573K	477K	336K	318K	305K	208K
Target Tokens	537K	459K	358K	325K	297K	218K

(source = foreign language, target = English)

Web-Scale Parallel Text

Both EuroParl & United Nations are large and well curated parallel texts,

but both have *very* specific domains & genres.

Europarl	CommonCrawl	Most Likely Tokens
9	2975	hair body skin products water massage treatment natural oil weight acid plant
2	4383	river mountain tour park tours de day chile valley ski argentina national peru la
8	10377	ford mercury dealer lincoln amsterdam site call responsible affiliates displayed
7048	675	market services european competition small public companies sector internal
9159	1359	time president people fact make case problem clear good put made years situation
13053	849	commission council european parliament member president states mr agreement
1660	5611	international rights human amnesty government death police court number torture
1617	4577	education training people cultural school students culture young information

Web-Scale Parallel Text

"...resulting in improvements of up to 1.5 BLEU on standard test sets, and 5 BLEU on test sets outside of the news domain."

Minimal cleaning & filtering still resulted in a substantial improvement in SMT performance

Manual inspection across three languages:
80% of the data contained good translations

Extracting US Phone Numbers

"Let's use Common Crawl to help match businesses from Yelp's database to the possible web pages for those businesses on the Internet."

Yelp extracted ~748 million US phone numbers from the Common Crawl December 2014 dataset

Regular expression over extracted text (WET files)

Extracting US Phone Numbers

Total complexity: 134 lines of Python
Total time: 1 hour (20 × c3.8xlarge)
Total cost: $10.60 (Python using EMR)

Matched against Yelp's database:

- 48% had exact URL matches
- 61% had matching domains

More details (and full code) on Yelp's blog post:
Analyzing the Web For the Price of a Sandwich

WikiReverse

Created by volunteer Ross Fairbanks *for fun*

Task: Find hyperlinks to Wikipedia from the web

Result: Dataset of over 36 million links

Code and data released online at wikireverse.org

Similar work by UMass and Google Research:
Wikilinks: A Large-scale Cross-Document Coreference Corpus Labeled via Links to Wikipedia

Common Crawl's Derived Datasets

Natural language processing:

- Parallel text for machine translation
- N-gram & language models (975 bln tokens)
- WDC "Collect ALL the web tables" (3.5 bln)

Large scale web analysis:

- WDC Hyperlink Graphs (128 bln edges)
- WikiReverse.org - Wikipedia in-links analysis

and a million more use cases!

Why am I so excited..?

Open data is catching on!

Even playing field for academia and industry

- *Baidu* used Common Crawl for Deep Speech
- Google Web 1T ⇒ Buck et al.'s N-grams
- Google's Wikilinks ⇒ WikiReverse
- Google's Sets ⇒ WDC Web Tables

Common Crawl releases their dataset
and brilliant people build on top of it

Challenge: Parser training data

Automatic Acquisition of Training Data for Statistical Parsers (Howlett and Curran, 2008)

Use knowledge base of facts or simple sentences:
"Mozart was born in 1756."

Parse more complex sentences with dep constraints:
"Wolfgang Amadeus Mozart (baptized Johannes Chrysostomus Wolfgangus Theophilus) was born in Salzburg in 1756, the second survivor out of six children."

Common Crawl

Read more at commoncrawl.org

Stephen Merity

stephen@commoncrawl.org

commoncrawl.org