

# Open Data Link

A dataset search engine for open data

Paul Ouellette and Justin Fagnoli

# Open Data Link

- ▶ Dataset search engine for open data.
- ▶ Search methods:
  - ▶ Semantic keyword search
  - ▶ Joinable table search
  - ▶ Unionable table search

# Motivation

- ▶ Governments and other organizations publish a lot of open data, but discovery is still difficult.
- ▶ Data scientists can identify ways to integrate datasets.
- ▶ Data publishers can see the wider context of their data.

# Demo

# Outline

System overview

Joinable table search

Unionable table search

Semantic Keyword Search

Future work

# Outline

System overview

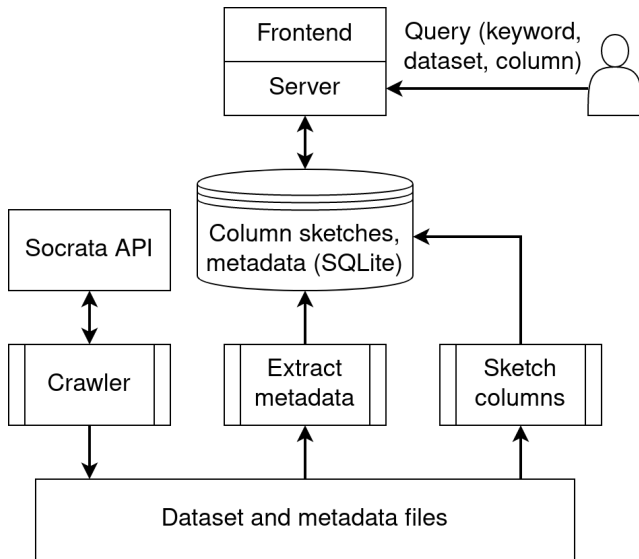
Joinable table search

Unionable table search

Semantic Keyword Search

Future work

# System overview



# Dataset crawl

- ▶ 10k of 42k datasets on Socrata.
- ▶ 172k columns.
- ▶ Most datasets are small.
- ▶ Largest datasets have over 100 million rows.



# Outline

System overview

Joinable table search

Unionable table search

Semantic Keyword Search

Future work

## Minhash<sup>2</sup>

- ▶ Data sketch for estimating Jaccard similarity of sets.

$$J(S, T) = \frac{|S \cap T|}{|S \cup T|}$$

- ▶ A minhash signature is composed of the results of a number of minhashes.
- ▶ The probability that the minhashes for two sets are the same equals the Jaccard similarity of the sets<sup>1</sup>.
- ▶ Minhash LSH hashes similar signatures to the same bucket.

---

<sup>1</sup>Mining of Massive Datasets, Chapter 3.

<sup>2</sup>A. Broder, "On the Resemblance and Containment of Documents", Compression and Complexity of Sequences 1997.

## LSH Ensemble<sup>3</sup>

- ▶ Set containment is a better measure for computing joinability.

$$C(Q, X) = \frac{|Q \cap X|}{|Q|}$$

- ▶ We can convert Jaccard similarity to containment, given the sizes of the domains.
- ▶ The size of the indexed domain is not constant, so domains are partitioned by cardinality.
- ▶ A minhash LSH index is constructed for each partition.

---

<sup>3</sup>Erkang Zhu, Fatemeh Nargesian, Ken Q. Pu, Renée J. Miller, “LSH Ensemble: Internet-Scale Domain Search”, VLDB 2016.

# Outline

System overview

Joinable table search

**Unionable table search**

Semantic Keyword Search

Future work

# Unionable table search

- ▶ The LSH Ensemble index is queried for each column of the query table.
- ▶ Candidate tables are those that appear in  $\geq 40\%$  of the joinability queries.
- ▶ Candidates are ranked by alignment: the fraction of candidate columns that are unionable with a query column.

# Outline

System overview

Joinable table search

Unionable table search

Semantic Keyword Search

Future work

# Semantic Keyword Search

- ▶ Problem: Given a list of keywords, return datasets which are more similar than threshold  $t$ .
  - ▶  $0 \leq t \leq 1$
- ▶ Motivation: Data scientists want a simple way to find new and insightful datasets

# Our Approach

- ▶ Search on the metadata, not on the data in the dataset
  - ▶ Data in dataset is too noisy
- ▶ Metadata that we have:
  - ▶ Dataset description
  - ▶ Column description
  - ▶ Datasets tags

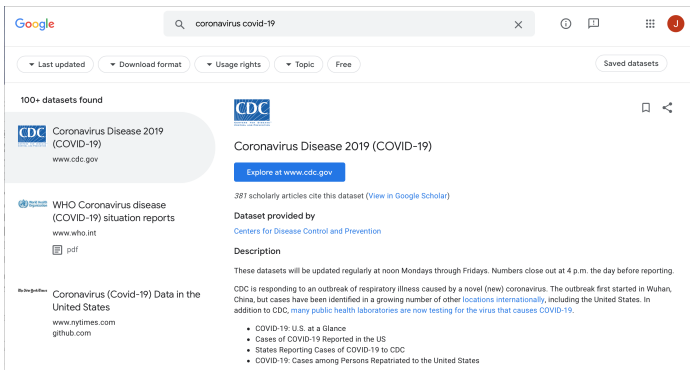


## Our Approach (Cont.)

- ▶ Use semantic NOT syntactic similarity
  - ▶ Example: Fish & Seafood
  - ▶ Example: Coronavirus & Respiratory System

# Others Approach

## ► Google Dataset Search



The screenshot shows the Google Dataset Search interface. At the top, the search bar contains 'coronavirus covid-19'. Below the search bar, there are filters for 'Last updated', 'Download format', 'Usage rights', 'Topic', and 'Free'. A 'Saved datasets' button is on the right. The results section shows '100+ datasets found'. The first result is 'Coronavirus Disease 2019 (COVID-19)' from the CDC, with a link to 'www.cdc.gov'. The second result is 'WHO Coronavirus disease (COVID-19) situation reports' from 'www.who.int', with a PDF icon. The third result is 'Coronavirus (Covid-19) Data in the United States' from 'www.nytimes.com' and 'github.com'. The main content area displays details for the CDC dataset, including a 'Description' section that states: 'These datasets will be updated regularly at noon Mondays through Fridays. Numbers close out at 4 p.m. the day before reporting. CDC is responding to an outbreak of respiratory illness caused by a novel (new) coronavirus. The outbreak first started in Wuhan, China, but cases have been identified in a growing number of other locations internationally, including the United States. In addition to CDC, many public health laboratories are now testing for the virus that causes COVID-19.' A list of bullet points follows: 'COVID-19: U.S. at a Glance', 'Cases of COVID-19 Reported in the US', 'States Reporting Cases of COVID-19 to CDC', and 'COVID-19: Cases among Persons Repatriated to the United States'.

Google coronavirus covid-19

▼ Last updated ▼ Download format ▼ Usage rights ▼ Topic Free Saved datasets

100+ datasets found

**CDC** Coronavirus Disease 2019 (COVID-19)  
www.cdc.gov

**WHO** Coronavirus disease (COVID-19) situation reports  
www.who.int  
pdf

**nytimes.com** Coronavirus (Covid-19) Data in the United States  
www.nytimes.com  
github.com

**CDC**

Coronavirus Disease 2019 (COVID-19)

Explore at [www.cdc.gov](https://www.cdc.gov)

387 scholarly articles cite this dataset ([View in Google Scholar](#))

**Dataset provided by**  
[Centers for Disease Control and Prevention](#)

**Description**

These datasets will be updated regularly at noon Mondays through Fridays. Numbers close out at 4 p.m. the day before reporting.

CDC is responding to an outbreak of respiratory illness caused by a novel (new) coronavirus. The outbreak first started in Wuhan, China, but cases have been identified in a growing number of other locations internationally, including the United States. In addition to CDC, many public health laboratories are now testing for the virus that causes COVID-19.

- COVID-19: U.S. at a Glance
- Cases of COVID-19 Reported in the US
- States Reporting Cases of COVID-19 to CDC
- COVID-19: Cases among Persons Repatriated to the United States

# System Overview

- ▶ FastText: words  $\rightarrow$  vectors
- ▶ SimHash: vectors  $\rightarrow$  bit vectors
- ▶ LSH: similarity search on bit vectors

- ▶ Vectors represent the semantics of words
- ▶ Closer a pair of vectors, closer the semantics of the two words
- ▶ closeness or similarity of vectors  $:=$  Cosine-Similarity

# Simhash

- ▶ Vector of floats  $\rightarrow$  Vector of bits

hash := an array of length  $H$  For vector with dimension  $d$ : Compute whether it is above or below  $d$  hyperplanes  $H$  times

# SimHash LSH

- ▶ L hash tables of bit vectors
- ▶ Query each L hash table for M candidates
- ▶ Compute cosine similarity of unhashed vectors to return top-M results

# LSH Forest

- ▶ Prefix Tree of bit vectors
- ▶ Variable length hash in tree solves tunability problem
- ▶ Query L Prefix Trees (the LSH Forest) for M candidates
- ▶ Compute cosine similarity of unhashed vectors to return top-M results

# Outline

System overview

Joinable table search

Unionable table search

Semantic Keyword Search

Future work



## Future work

- ▶ Organizing datasets into a directory structure for navigation.
- ▶ Use semantic similarity of attribute names in unionable table search.
- ▶ Similar dataset search based on metadata similarity.
- ▶ Keyword search over data values.