# Search Engine for Efficient Data Discovery of NYC Open Data

By Data InQuirer


Zahra K.

Second Year MS in Data Science
Candidate, New York University

Aakash Kaku

First Year MS in Data Science
Candidate, New York University

Chaitra Hegde

First Year MS in Data Science
Candidate, New York University

# Introduction and Problem Formulation

- With the advent of massive datasets available on open data portal like NYC Open Data, Data.Gov, efficient search of repository in real time plays a very crucial role
- As a project, we are determined to build a search engine for NYC Open data which contains almost 1500 datasets ~ 650 GB
- The aim of the project is to build an interactive search engine that helps users to find relevant datasets from a corpus of dataset that satisfies certain keyword conditions specified by the users

# Related work and Background

- Bellman browser : Make use of data summaries that helps in doing quick and crude exploratory analysis of the data
- Many approximate matching algorithms are developed that helps to performs joins based on string distances
- For this project, we make extensive use of the data summaries as they are found to be more efficient during the search time. Users can interactively play with such a browser and mine the data that is more relevant for them.

# Method

- As mentioned above, method of data summary is adopted to build the search engine. The data summary should contain informative features that efficiently and effectively describes each dataset in the repository.
- **Trade-off between speed and accuracy**: Data Summary will lead to loss of information. Hence, there is a trade-off between speed and accuracy
- The trade off between speed and accuracy can be well answered if the features that summarizes the data are very informative and well suited to the task

# Architecture and Design (1/3)

**Stages of Search Engine Building**:

1. **Feature Engineering**: The most important stage of the dataset searching process. During this stage, relevant features that can informatively characterize a dataset are chosen or built.
2. **Extracting Feature and Building Data Summary**
3. **Building query templates to perform diverse and informative search**

# Architecture and Design (2/3)

**Extracting Feature and Building Data Summary:** Pyspark, HPC and parallel computing were used to derive features from ~650 GB of data. The set of features extracted are mentioned below:

| File name | Category | Column Name | Description | Download Count | View Count |
|---|---|---|---|---|---|
| Tags | Word | Column Type | Word Count | Number of Rows | Number of Columns |
| Maximum | Minimum | Nulls | Non-Nulls | Attribution Link | Field Name |

# Architecture and Design (3/3)

**Building query templates to perform diverse and informative search:**

| Sr. | Query |
|---|---|
| 1 | Dataset with certain number of rows and columns |
| 2 | Search for multiple keywords in OR fashion |
| 3 | Search for multiple keywords in AND fashion |
| 4 | Search for files with certain columns in it |
| 5 | Search for files with a certain column and value |
| 6 | Search keywords in meta data and data |

| Sr. | Query |
|---|---|
| 7 | Search for keyword in metadata |
| 8 | Search for a keyword with threshold on number of times it occurs |
| 9 | Search for a keyword with threshold on number of rows in the dataset |
| 10 | Dataset belonging to certain category |
| 11 | See the Schema of Dataset |

# Technical Depth and Innovation

- The data summary is prepared by extracting most useful information from the data and metadata of the files. To compute and extract such features, a spark implementation using SQL and RDD were used.
- Increase Relevance of Search: Similar words to the given keyword are determined using cosine similarity applied over word2vec representation of all the natural occurring words
- To reduce running time, in every possible place, threading is made use to parallelize the execution of code.

**Demo Time!**


OpenData

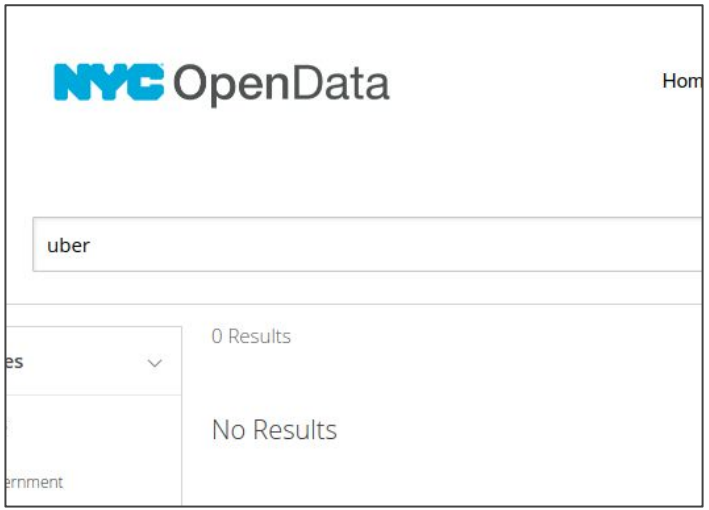
# Results and Evaluation

- 11 different type of advance search capabilities are present in the search engine which are not present in the search engine of NYC Open Data Website
- It is found to give better results than the website's search engine
- Users can tailor the search as per their needs by querying the summary tables in an efficient and easy manner.
- Data summaries have many other applications like finding similar datasets, finding similar column in a table, etc. Data summary tables can easily be leveraged for performing such analysis as well.

# Future Work

- Another possible search methodology could be to find appropriate clusters of dataset to search over and perform search only on those dataset that fall in the cluster
- For above methodology, some kind of clustering algorithm would be needed
- Ranking the search results using some relevance metric

# Appendix



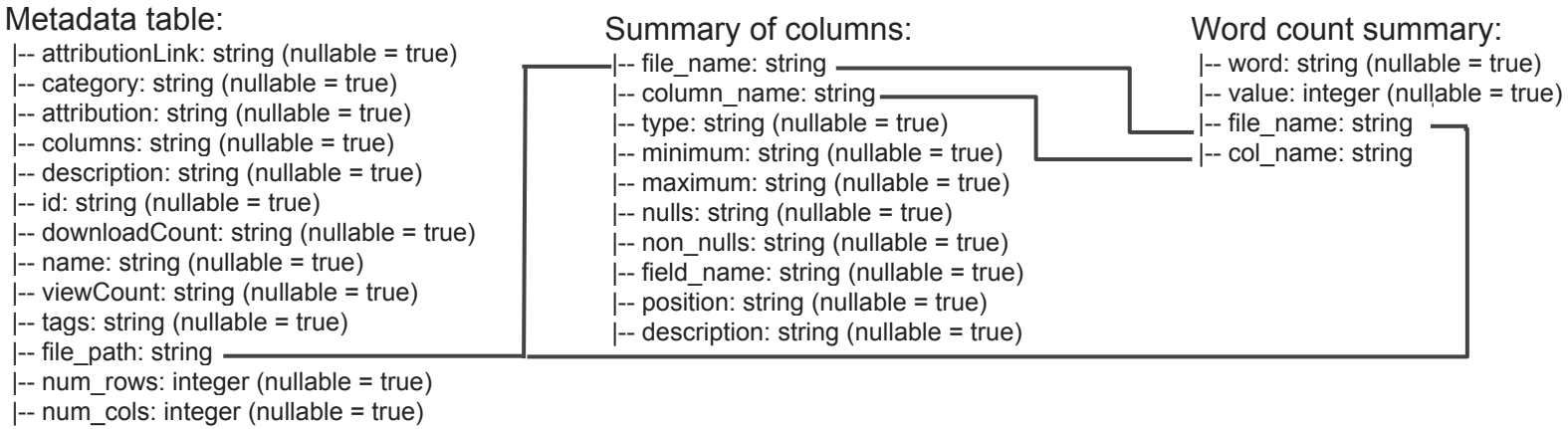

uber

| name | file_path |
|---|---|
| FHV Base Aggregate Report | /user/bigdata/nyc_open_data/2v9c-2k7f.json |
| 2013-2017 School Math Results - SWD | /user/bigdata/nyc_open_data/7k5d-rk33.json |
| 2016-17 Physical Education - MTI All- Star Schools | /user/bigdata/nyc_open_data/g9wv-7n2m.json |
| 2013-2017 School ELA Results - All | /user/bigdata/nyc_open_data/kdm9-vp7d.json |
| 2013-2017 School Math Results - All | /user/bigdata/nyc_open_data/kha6-7i9i.json |
| 2013-2017 School ELA Results - Gender | /user/bigdata/nyc_open_data/m8nr-4ivu.json |
| 2013-2017 School ELA Results - EvELL | /user/bigdata/nyc_open_data/mib5-bwqy.json |
| 2013-2017 School Math Results - EvELL | /user/bigdata/nyc_open_data/rqjq-29wc.json |
| 2013-2017 School Math Results - Gender | /user/bigdata/nyc_open_data/x4ai-kstz.json |
| 2013-2017 School Math Results - Ethnic | /user/bigdata/nyc_open_data/xx92-6788.json |
| 2013-2017 School ELA Results - SWD | /user/bigdata/nyc_open_data/ybcb-4665.json |
| 2013-2017 School ELA Results - Ethnic | /user/bigdata/nyc_open_data/ynau-kwze.json |

**Comparison of Search for word ‘uber’**

# Appendix

## Schema of summary tables:



# Appendix

Query 9: Find the datasets that contains '**income**' in the column name:

| path | name |
|---|---|
| /user/bigdata/nyc_open_data/gffu-ps8j.json | Income By Type Of Income And AGI Range |
| /user/bigdata/nyc_open_data/nwet-nc6h.json | Tax Credits By Agi Range |
| /user/bigdata/nyc_open_data/3vvi-fwjs.json | Tax Liability By AGI Range |
| /user/bigdata/nyc_open_data/ipc3-2nbm.json | Personal Income By AGI Range |
| /user/bigdata/nyc_open_data/9ay9-xkek.json | Local Law 44 - Unit Income Rent |

NYC OpenData

# Appendix

Query 10: Find datasets with a column that contain a keyword **‘address’** with keyword ‘**building**’ in it **:**

| path | name |
|---|---|
| /user/bigdata/nyc_open_data/9a87-6m4x.json | SBS ICAP Contract Opportunities |
| /user/bigdata/nyc_open_data/u35m-9t32.json | DCLA Cultural Organizations |
| /user/bigdata/nyc_open_data/m3fi-rt3k.json | New York City Council Discretionary Funding (2009-2013) |
| /user/bigdata/nyc_open_data/un8d-rbed.json | SBS ICAP Contract Opportunities - Historical |
| /user/bigdata/nyc_open_data/ye3c-m4ga.json | Civil List |

NYC OpenData

# Appendix

Query 2: Find datasets that contains word **‘classroom’** at least 50 times:

| file_name | name |
|---|---|
| /user/bigdata/nyc_open_data/8gpu-s594.json | Application for State Aid |
| /user/bigdata/nyc_open_data/tm6d-hbzd.json | Incidents Responded to by Fire Companies |
| /user/bigdata/nyc_open_data/gshi-yqza.json | Transportable Classroom Units- Buildings & Schools |

Query 7: Find the datasets that contains ‘**alcohol**’ in it:

```
['alcohol']
['drinking', 'marijuana', 'alcohol']
```

| file_name | name |
|---|---|
| /user/bigdata/nyc_open_data/tm6d-hbzd.json | Incidents Responded to by Fire Companies |
| /user/bigdata/nyc_open_data/8sdw-8vja.json | Asset Management Parks System (AMPS) - Work Orders |
| /user/bigdata/nyc_open_data/m3fi-rt3k.json | New York City Council Discretionary Funding (2009-2013) |
| /user/bigdata/nyc_open_data/6r4h-c2y6.json | Sustainability Indicators (2012) |
| /user/bigdata/nyc_open_data/8nqg-ia7v.json | Mental Health Service Finder Data |