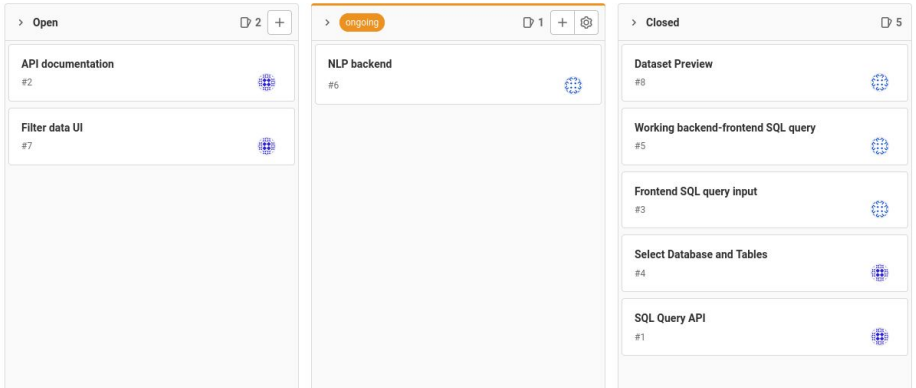
End-user Search Queries on CSV datasets

Team-12
Tathagata Raha - 2018114017
Dipanwita Guhathakurta - 2018112004
Pawan Baswani - 2021701035

Overview



5 issues closed, 1 in progress
40 merged commits

Recent progress

- Software requirements & Design Specifications
- API documentation (ongoing)
- Working pipeline for table & dataset selection
- Working pipeline for SQL query
- Working pipeline for Table Preview
- Frontend for Natural Language query

Progress - UI Design

Dataset & Table selection

Implemented Now UI-based frontend for selection of datasets and single/multiple tables for queries.

Data preview

Embed description and preview for each selected table through dropdowns.

SQL query

Accept SQL query input in an in-browser IDE and embed output preview as tables

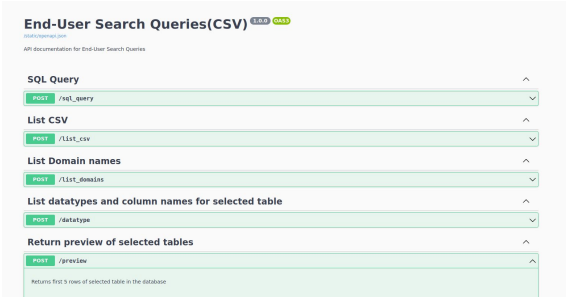
Natural Language Input

Accept text input from the user for natural language processing

Progress - Backend

01	/list_domains	• CSV tables are organized into domains. Returns a list of available domains in the database
02	/list_csv	• Returns a list of datasets (tables) given a domain. User can select one/multiple tables within the chosen domain
03	/preview	• For each dataset, return a JSON form of the table head() to embed into the frontend
04	/datatype	• For each dataset, return a JSON containing column names and column datatypes.
05	/sql_query	• Accepts a list of user-selected tables and SQL query, and executes the query on the dataset. Returns JSON form of output for user preview and saves the output temporarily for download.

Detailed Swagger API docs with demo at /docs



Individual contributions

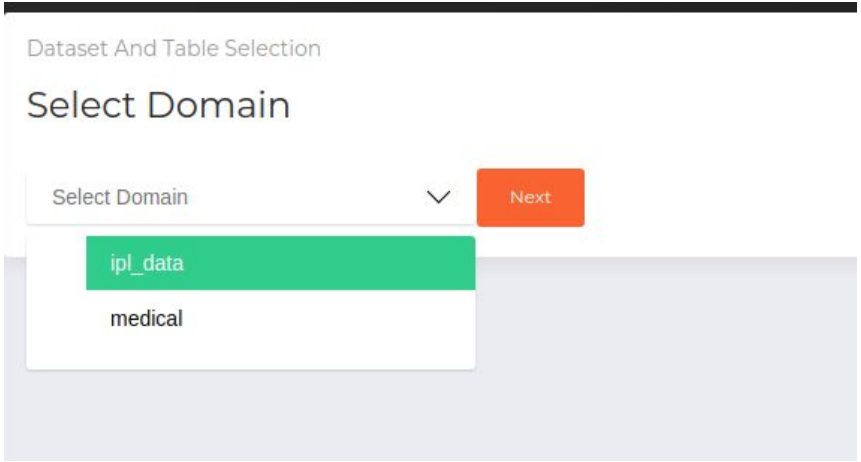
Issue	Tathagata	Dipanwita	Pawan
Wizard (Frontend)			
Code Editor (Frontend)			
Query Output (Frontend)			
Table and Data Preview (Frontend)			
NLP query (Frontend)			

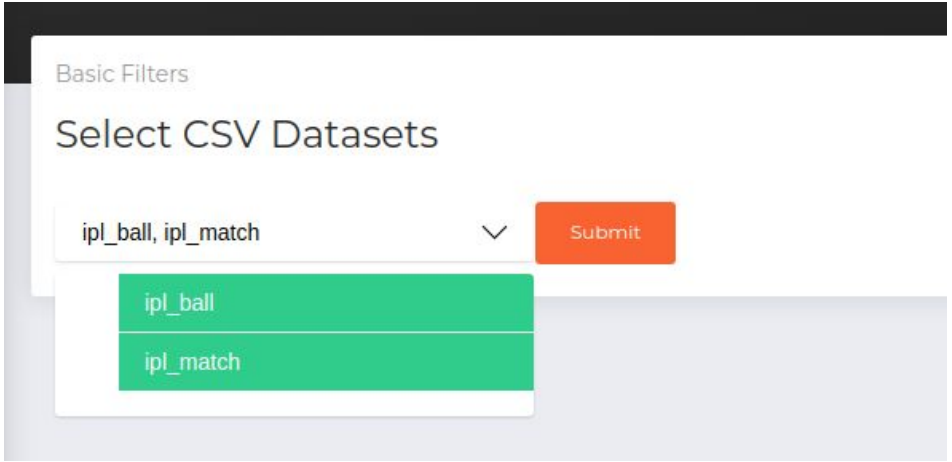
Individual contributions

Issue	Tathagata	Dipanwita	Pawan
Wizard (Backend)			
Query SQL (Backend)			
Query Output (Backend)			
Table and Data Preview (Backend)			
NLP query (Backend)			
API Documentation (Backend)			

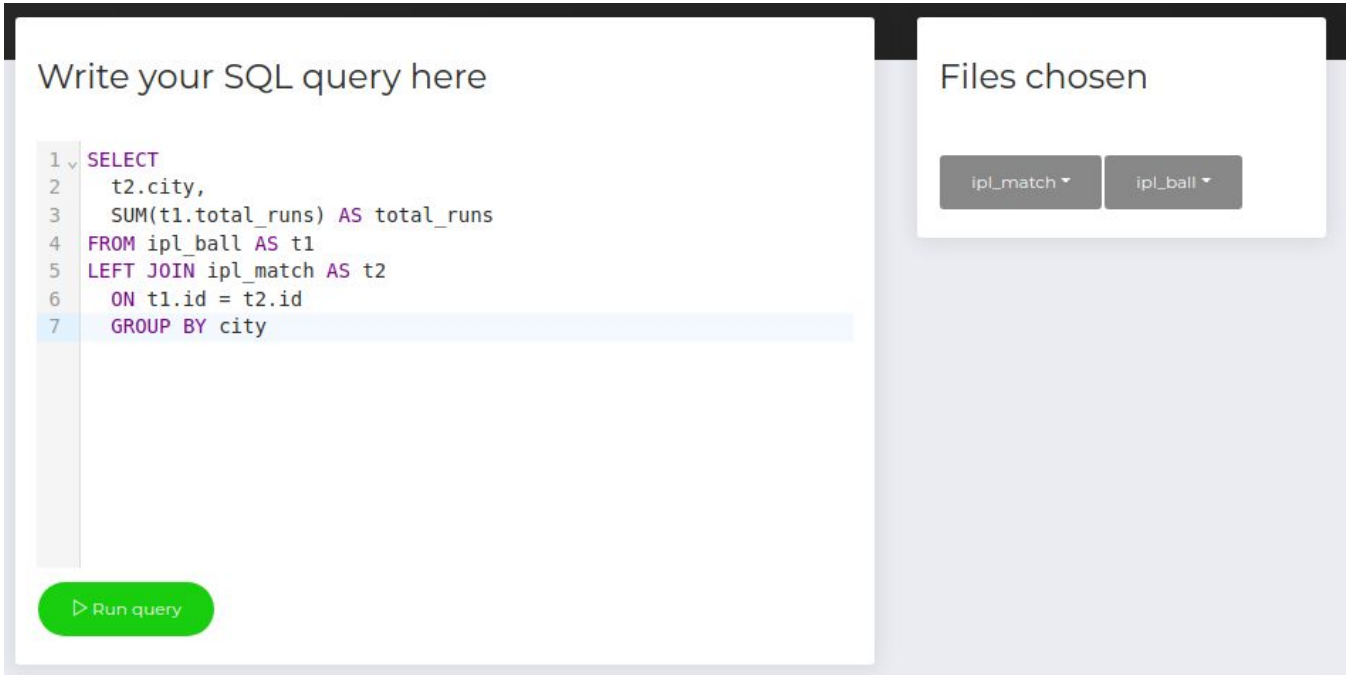
Screenshots

Wizard


Dataset And Table Selection
Select Domain
Select Domain
Next
ipl_data
medical


Basic Filters
Select CSV Datasets
ipl_ball, ipl_match
Submit
ipl_ball
ipl_match

SQL Editor

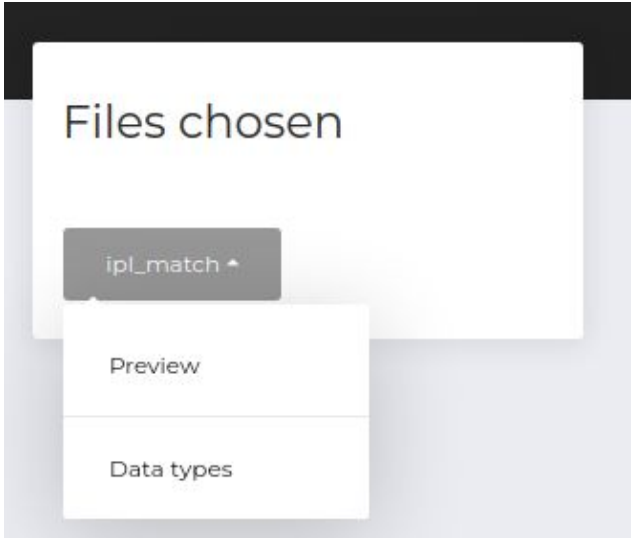

Write your SQL query here
1 SELECT
2 t2.city,
3 SUM(t1.total_runs) AS total_runs
4 FROM ipl_ball AS t1
5 LEFT JOIN ipl_match AS t2
6 ON t1.id = t2.id
7 GROUP BY city
Run query
Files chosen
ipl_match
ipl_ball

Query output

Output

city	total_runs
	3893
Abu Dhabi	8830
Ahmedabad	3746
Bangalore	20237
Bengaluru	5127

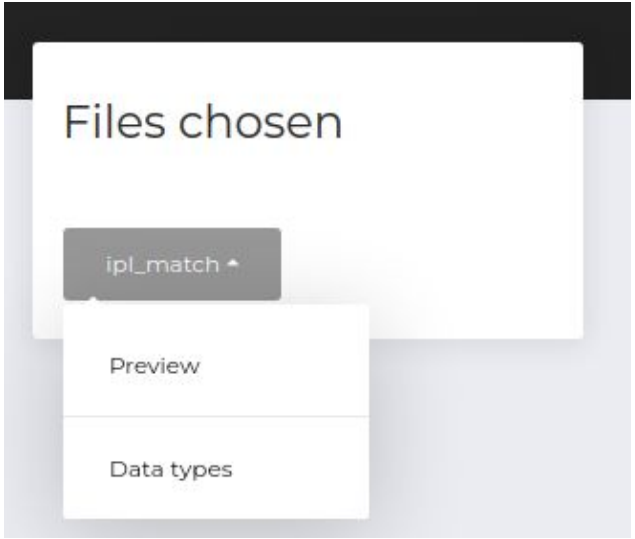
Table preview



Preview of ipl_match

id	city	date	player_of_match	venue	neutral_venue	team1	team2	to
335982	Bangalore	2008-04-18	BB McCullum	M Chinnaswamy Stadium	0	Royal Challengers Bangalore	Kolkata Knight Riders	Roy Cha Bar
335983	Chandigarh	2008-04-19	MEK Hussey	Punjab Cricket Association Stadium, Mohali	0	Kings XI Punjab	Chennai Super Kings	Che Kin
335984	Delhi	2008-04-19	MF Maharoof	Feroz Shah Kotla	0	Delhi Daredevils	Rajasthan Royals	Raj Roy
335985	Mumbai	2008-04-20	MV Boucher	Wankhede Stadium	0	Mumbai Indians	Royal Challengers Bangalore	Mu
335986	Kolkata	2008-04-20	DJ Hussey	Eden Gardens	0	Kolkata Knight Riders	Deccan Chargers	Dec Cha

Table data types



Data types of ipl_match

Column	Data type
id	int64
city	object
date	object
player_of_match	object
venue	object
neutral_venue	int64
team1	object

Natural language queries

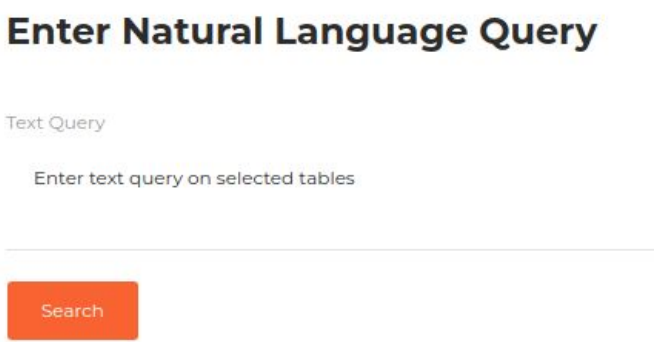
```
agent.get_query("how many deaths of age below 40 had stomach cancer?")
```

```
'SELECT COUNT(death_count) FROM dataframe WHERE age  < 40 AND cancer_site = "Stomach" '
```

```
>>> agent.get_query("Sum of total profit where units sold is more than 100")
'SELECT SUM(total_profit) FROM dataframe WHERE units_sold  > 100 '
```

```
agent.get_query("What are the types of cancers found in people in 2013?")
```

```
'SELECT cancer_site FROM dataframe WHERE year = "2013" '
```



API documentation