DataButler

A python-based tool for automatic data profiling and cataloging

Keerthi Priya Pullela, Rahul Madhu, Rukmini Sunil Nair, Sagar Kurada, Xema Pathak, Matthew A. Lanham

Purdue University, Krannert School of Management

[kpullel, rmadhu, nair69, skurada, xpathak, lanhamm] @purdue.edu

Abstract

DataButler is an open-source python-based tool for data profiling and cataloging. With increasing bytes of data, data scientists and analysts need intelligence to automatically discover key information about the data.

DataButler also allows analysts to be more efficient without them having to perform EDA on every dataset that they work on. Apart from computing basis relational statistics

DataButler also focuses on the automatic categorization of a dataset and discovers metadata information.

Background

Data profiling is the process of verifying users' structured data, semi-structured data, and unstructured data, gathering data structure, data pattern, statistical information, distribution messages, and reviewing data attributes for data governance, data management, data migration, and data quality control.

Use cases of data profiling

Data Exploration | Data Cleaning | Data Integration | Query Optimization | Scientific Data Management

Functionalities

Existing studies on data profiling focus on identifying the datatype of the column and sanitizing data based on statistical indicators such as mean and median. However, our DataButler is expected to help perform the following tasks:

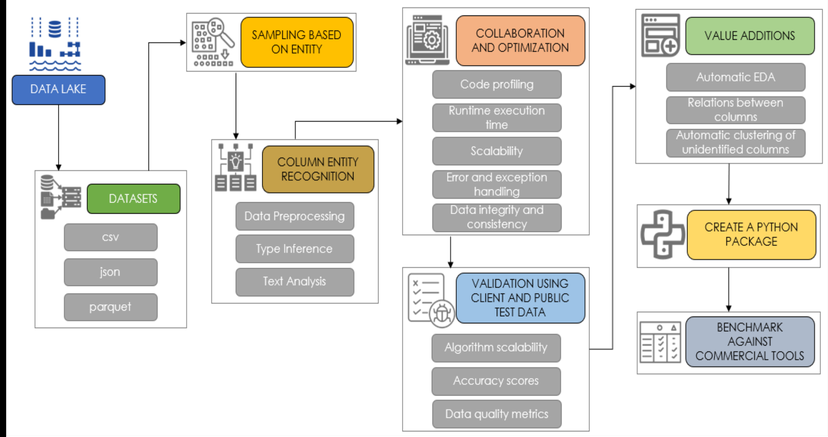
- Identify crucial metadata of the given dataset
- Identify invalid values for the given column in a dataset
- Identify sensitive information such as credit card details, SSN, et al. that can be hashed to protect user information and ensure data privacy
- Perform initial exploratory data analysis to give a bird's eye view of the unknown dataset

Methodology

The following is the condensed methodology workflow we follow -

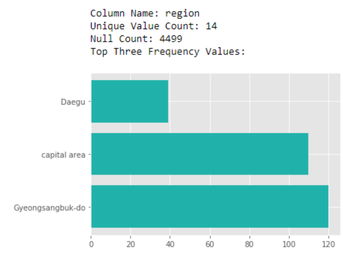
COLUMN ENTITY RECOGNITION	The data type, as well as contents of each column is identified. Ex: animals, currency, credit card number etc.
FUNCTIONS AND LOGIC FOR EACH ENTITY	Various algorithms and python libraries are employed for each entity; these are all brought together and optimized.
COLLABORATION AND OPTIMIZATION	The data type, as well as contents of each column is identified. Ex: animals, currency, credit card number etc.
VALIDATION USING CLIENT AND PUBLIC TEST DATA	We measure how well our test data fits our framework.
VALUE ADDITION	Automatic EDA, interdependencies among columns, automatic clustering of unidentified columns etc.
CREATE A PYTHON PACKAGE	Convert the framework to a Python package in order for it to be open source and easy to use.
BENCHMARK AGAINST COMMERCIAL TOOLS	Compare, contrast and measure the performance of our framework against commercial tools such as Informatica.

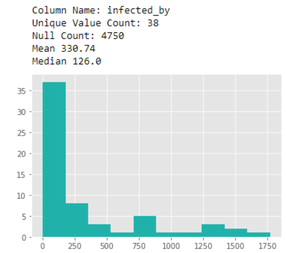
Here is a detailed process flow -



Results

Some sample results – Frequency distributions and EDA for each entity, along with confidence scores.






```
Possible Primary Key(s): ['id']
Columns with missing data: ['sex', 'birth_year', 'region', 'group', 'infection_reason', 'infection_order', 'infected_by', 'co
ntact_number', 'released_date', 'deceased_date']

                CONFIDENCE SCORES:

+----------------+--------------+------------------+
|  Column Name   | Tested Entity | Confidence Score |
+----------------+--------------+------------------+
|      sex       |    gender    |      100.0       |
|    country     |    country   |      100.0       |
|     region     |     city     |       8.57       |
|     group      |     name     |       7.14       |
| confirmed_date |     date     |      100.0       |
| released_date  |     date     |      100.0       |
| deceased_date  |     date     |      100.0       |
+----------------+--------------+------------------+
```

Benchmarking

Attributes	DataButler	Trifacta	Informatica	IBM InfoSphere	SAP Information Steward	Pandas Profiling
Distribution: Quantitative Data	✓	✓	✓	✓	✓	✓
Distribution: Qualitative Data	✓	✓	✓	✓	✓	✓
EDA: Statistics	✓	✓		✓	✓	✓
EDA: Missing Values	✓	✓	✓	✓	✓	✓
EDA PDF/HTML Output	✓	✓		✓		✓
Primary Key Identification	✓		✓	✓		
Primary Key based Joining		✓	✓	✓	✓	
Entity Recognition	✓				✓	
Confidence Scores	✓				✓	
Similar Entity Recognition			✓			
Open Source	✓					✓

Conclusions and Future Scope

- Data Butler is a python-based tool for categorizing data into 19 entities', containing numeric, string and datetime data for csv, text and json files.
- In the future, the scope can be extended to more entity recognitions, merging multiple datasets on primary keys with a better user interactive interface.
- Github Repository: https://github.com/DataButler .