ML Data Classification System: Vision and Scope

DataPlumbers
Computer Science Department
California Polytechnic State University
San Luis Obispo, CA USA

October 4, 2018

Contents

Credits

Name	Date	Role	Version
Zachary Richardson	October 4, 2018	Data Plumber	1.0
Tony Chen	October 4, 2018	Data Plumber	1.0
Bennette Robinson	October 4, 2018	Data Plumber	1.0
Kyler Ramsey	October 4, 2018	Data Plumber	1.0
Samuel Nayerman	October 4, 2018	Data Plumber	1.0

Revision History

Name	Date	Reason for Changes	Version
———	October 8, 2018	Initial draft	1.0

1 Business Requirements

1.1 Background

Data handling is becoming increasingly important in variety of nontechnical fields. Transparency, reproducibility, and verifiability are key requirements to establishing veracity in published articles, research, and legal findings. Simultaneously, the continuously expanding collection of available data-sets makes it rather difficult to find, analyze, and share results. Subsequently, the privacy of sensitive information must be respected during the research and authoring process, and also in the final results.

1.2 Business Opportunity

Our business opportunity with MarkLogic is to help them leverage their software solutions for their customers by having a working product of the Data Discovery Workbench. By providing a software solution to data classification, users will be able to more effectively classify, manage, and distribute findings from a given data set. The customers will realize the value that the Data Discovery Workbench can bring to their work enterprise or research by saving them the time and cost in classifying large collections of untrained data sets while being able to discover other relevant data sets of interest.

1.3 Business Objectives and Success Criteria

BO	**Info**
BO-1	Develop the Data Classifier
BO-2	Develop the Data Catalog
BO-3	Develop the Data Anonymizer

SC	**Info**
SC-1	Trained machine learning model
SC-2	Effective classification of data sets
SC-3	Working Data Anonymizer

1.4 Customer or Market Needs

- Users must be able to find data that is relevant to their topic area from amongst vast amounts of data.
- Users must carefully track the lineage of source data and the manipulations performed on that data. This meta-data must be available, alongside the final published content, to journalists, academics, and other information consumers.
- Users must methodically redact or anonymize data to maintain privacy.
- Users must be able to update their results if the input data changes or the analysis methods change. This includes the fact that some source data is temporal in its accuracy.

1.5 Business Risks

Mislabeling of classified or sensitive data.

2 User Description

2.1 User/Market Demographics

Data scientists, journalists, academics, and other information consumers.

2.2 User Personas

A data scientist John working for a health care company contracts MarkLogic to overhaul the company's database. To better organize the existing data, John would like to classify different data types from his company's various databases. Different branches of the company have collections of unclassified and classified customer, enterprise, and public information. John is able to feed multiple data sets into MarkLogic's data classifier, customize categories in the data catalog, and make informed decisions to help his company manage their new database software.

2.3 User Environment

The environment for the users needs to be a clear and concise page, where users can view data sets at different points in the training model, analytics behind the data, and other all-encompassing metrics. The user will also need a computing device capable of installing and running a modern internet browser, while having access to an internet connection.

2.4 Key User Needs

User should be able to easily operate and process data discovery between categorized data sets from different data formats.

3 Vision of the Solution

3.1 Vision Statement

The Data Discovery Workbench is a collection of tools and technologies that will be built over time and eventually implement several key components: Data Classification, Data Cataloging, and Data Anonymizer.

3.2 Solution Overview

Our solution is to implement a Data Classifier, Data Catalog, and Data Anonymizer to help facilitate data consumers in research fields, migrate complex layers of data, and increase efficiency in data discovery.

The **Data Classifier** will have predefined and learned categories of data elements that will be recognized by our machine-learning feature. Users will be able to manipulate the data elements of a data category to fit their needs of other untrained data sets. This will enable users to seamlessly classify data into existing trained data sets.

The **Data Catalog** will assist in the data visualization aspect of the data classification, allowing users to query for the types of data or data sets theyre looking for. It will also allows users to search for data sets that have been found to be relevant for similar research. This is made possible by searching what others have checked out.

The **Data Anonymizer** will process data sets so that no users will be allowed to see sensitive information. It is the job of the Anonymizer to respect

metadata from the Catalog and invoke anonymizationtools asneededbased on the data and user that is asking for the data.

3.3 Major Features

There are two web-based user interfaces and a backend consisting of machine learning steps and processing of data generated by those steps. Additionally, there are a number of stretch goals which may be added to enhance the functionality. For this project, the Data Classifier will be the main focus.

Feature	**Description**
Element Category GUI	Allow users to see predefined and learned categories of data elements that will be recognized by the Data Classifier (e.g. names, social security numbers, phone numbers, or other domain specific information). They will be presented with both the name of the category and a description (if present). Allow users to label data element categories as sensitive.
Machine Learning - Create Models	Process training sets to create models to be used to identify categories of data elements (e.g. names, phone numbers) from input data sets.
Machine Learning - Apply Models	Apply models to input data sets to do the actual categorization of the data elements.

Process Learned Data	Generate a machine-readable description of an input data set based on the categorization process. The description will provide as much detail as was learned about each data element including whether it is sensitive or tagged in some other way (see stretch goals).This machine-readable data will ultimately go into the Data Catalog, but for this project it is acceptable for it to exist as XML or JSON files.
Data Classifier GUI	Allow users to browse, edit, and add to the classifications that were performed automatically. A user may remove an erroneous classification, add a new classification for a data element that was not automatically classified, or add/remove tags (such as sensitive).

Stretch Goals

Feature	**Description**
Element Category GUI	If new categories of data are learned, the GUI will provide the user a mechanism for naming and describing these categories. In addition to allowing users to specify which data element categories are sensitive, allow users to create and apply other arbitrary tags for data element. In this way users will be able to not only tag a data element category as sensitive, they will beable to tag it with any other term that is relevant to them.

Machine Learning - Create Models	Apply unsupervised learning techniques to discover new categories of data elements from unlabeled input data sets. This should be done through the application of existing techniques.
Data Classifier GUI	Capture changes that are made by the user to auto-generated classifications for feedback into the Machine Learning Training phase. For example, notice when a user manually applies a classification. That is a newlabel. Make those new labels available as training data.

3.4 Assumptions and Dependencies

The Data Catalog component would depend on the completion of the Data Classifier. The main goal is to deliver the Data Classifier component first and then continue to delivering both Data Catalog and Data Anonymizer if time permits.

The users of the product:

- have a working internet connection
- have access to their own data sets to use as input to the software

4 Scope and Limitations

4.1 Scope of Initial and Subsequent Releases

Initial Release targets the end of CSC 405. One or two subsequent releases will occur in CSC 406.

4.2 Limitations and Exclusions

Data sets used in production will be unknown until the tool is deployed to users.

5 Business Context

5.1 Stakeholder Profiles

- **MarkLogic**: MarkLogic is our main stakeholder for this project. As the client, they are hoping to acquire a product from our team that can help meet the needs of their own customers.
- **Prof. da Silva**: The professor for our class will be acting as the facilitating advisor for our project. The success of this project will depend on meeting his project criteria.
- **Data Plumbers**: Our development team will be responsible for delivering value to our customer, MarkLogic, by committing to continuous software releases until the product requirements are met.

5.2 Project Priorities

The priority is to have a finished product of the Data Classifier component, along with its 2 major GUIs. The Data Catalog and Data Anonymizer components will be stretch goals for this project but we will do our best to release these components if possible.

5.2.1 Release 1

A rough prototype of the Data Classifier by the end of CSC 402.

5.3 Operating Environment

To access the product:

- Users will need to have a modern internet browser, such as Google Chrome, Mozilla FireFox, Microsoft Edge, and Safari, installed on their computer in order to properly operate the software components.

6 Competitive Analysis

6.1 Overview

MarkLogic is primarily an enterprise NoSQL database company. While MarkLogic has been in business for over a decade, many of it's competitors are newer companies. Competition includes companies such as MongoDB, DynamoDB, and CouchDB. Currently, we are unaware of a NoSQL competitor that provides a data classification system.

6.2 MongoDB

MongoDB is a free and open-source cross platform NoSQL database program. It is document oriented and uses a JSON like structure with schemata. Unlike MarkLogic, MongoDB is open source. MongoDB provides no means of data classification.

6.3 Amazon DynamoDB

DynamoDB is a NoSQL database service that is built upon Amazon's cloud computing infrastructure. It is known for automatically configuring and scaling of your database cluster.