CSE 3063 Fall 2020 Java Project 1
Group Number 14
Iteration 2 Requirement Analysis Document

1. Description

This software is designed to label instances in a dataset. Instances may be comments on an e-commerce website and the labels might be based on whether these comments are positive or negative. After the labeling process, the dataset can be used in machine learning models to create intelligent systems.

In our project, different users may use the system to label instances in a dataset. Datasets can be assigned to different users for processing. The software also keeps track of the history of the labeling process to calculate statistics per user, instance and dataset.

These statistics may help us analyze the completeness of a dataset, dominant label among the instances in a dataset. We can also see how much time on average a user spends on an instance to label it. Another useful metric is to see the consistency of a user on the labels given by forcing him/her to label an instance already labeled by him/her on random.

2. Requirements

2.1 The program should support a multi-user system.

2.2 Datasets should be given in JSON files.

2.3 The program should read the dataset from these JSON files.

2.4 After reading the dataset from JSON files, the program should randomly label the data with the corresponding user ID's.

2.5 Since this is a multi labeling system, the user may assign more than one class label ID to a single instance.

2.6 Users can be requested to label an instance that is already labeled by them with a certain probability to track their consistency.

2.7 Every instance in a dataset should track its labels given by different users to determine its final label.

2.8 A dataset should track its total completeness percentage for each user.

2.9 After the labeling is done, the program should create the output as a JSON file.

3. Glossary of Terms

3.1 Dataset: A group of instances to be labeled.

3.2 Class Labels: Predetermined categories to assign the instances.

3.3 User: People who make the labeling of the datasets.

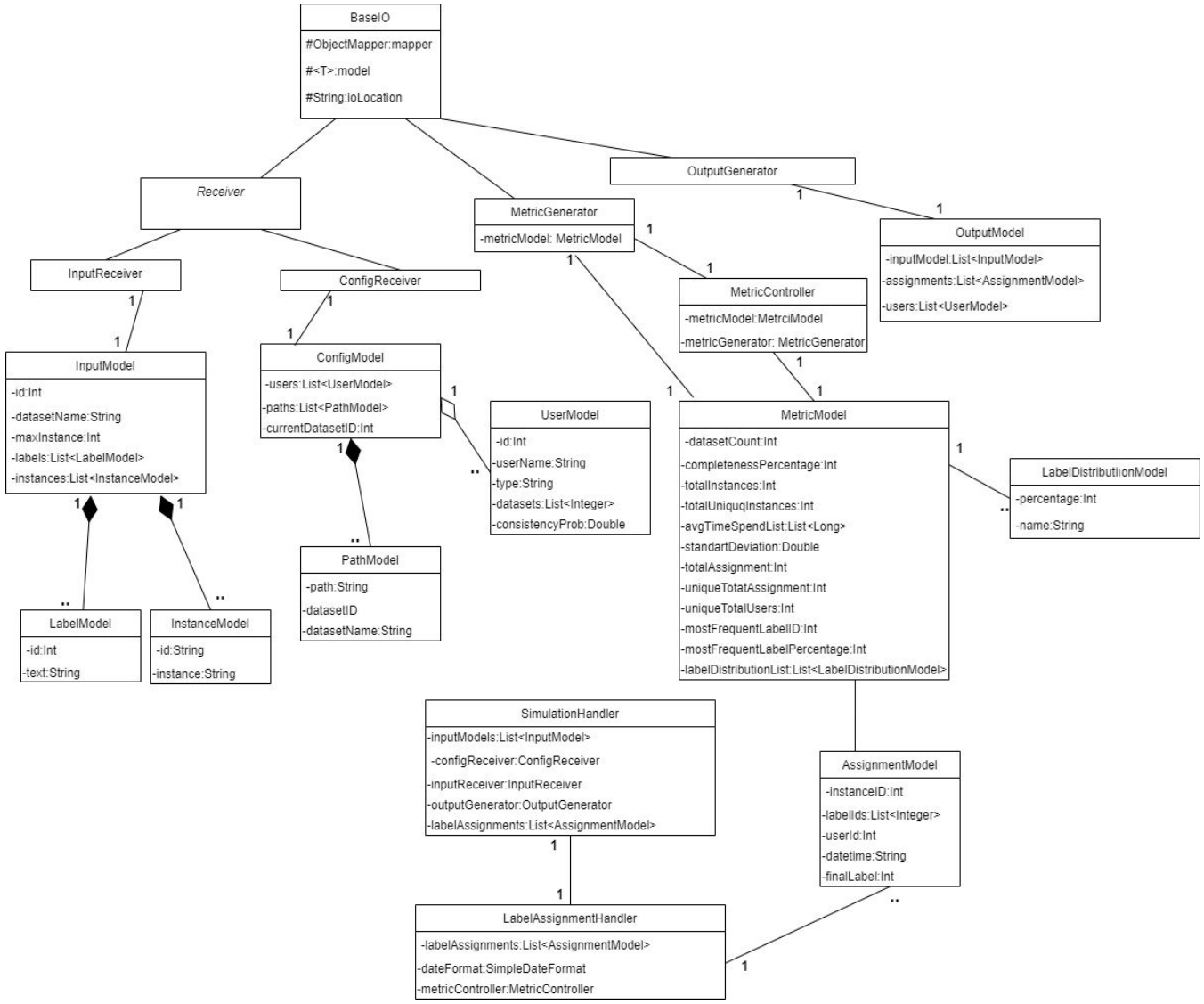
3.4 JSON: JavaScriptObjectNotation files that are used to showcase structured data.

3.5 Instances: An instance is a single comment given to users in the system to be labeled.

4. Use Cases

4.1 Aid researchers while labeling big datasets.
4.2 Help users categorize their own data offline.
4.3 Multiple users can label the same data instance with different Labels.
4.4 Final label of an instance is determined by a consensus of a group of different users. So a user's individual error can’t affect the final label. This will help us minimize the noise in the labeled data.

5. Domain Model



6. System Sequance Diagram(SSD)