MARMARA UNIVERSITY

FACULTY OF ENGINEERING

DEPARTMENT OF

COMPUTER ENGINEERING

Data Labelling System Requirement Analysis

CSE3063 Object Oriented Software Design

Team
Group Number: 2
150118013 - Gökberk Çelikmasat
150118065 - Şeref Kutay Akgün
150117005 - Kadir Acun
150119825 - Merve Rana Kızıl
150119824 - Zeynep Ferah Akkurt
150117905 - Erhan Yalnız

1. Introduction

Vision: The objective of the project is the implementation of a labeling system simulation with Java based object-oriented software design approach. The project is designed and developed by a team of 6 members.

Scope: Data Labeling System Simulation

Special Requirements

- Java
- Visual Studio
- Eclipse
- Git
- IntelliJ
- PlantUML
- Json-simple.jar
- Visual Paradigm
- Log4j-api.jar
- Log4j-core.jar

Stakeholders

- Murat Can Ganiz (Customer)
- Lokman Altın (Customer)
- Gökberk Çelikmasat
- Şeref Kutay Akgün
- Kadir Acun
- Merve Rana Kızıl
- Zeynep Ferah Akkurt
- Erhan Yalnız

2. Requirement Specification

2.1 Functional Requirements

2.1.1 Must Have Requirements

- A configuration file must be created and copied into the source folder of the project.
- External libraries like json-simple.jar, log4j-api.jar and log4j-core.jar.
- Labeling operations will be done with command prompt.
- Labeling operations will be done randomly.
- It is a multi-user system.
- A user may label a particular instance more than once.
- Labeling operation will continue until all the instances are labeled by all users.
- Metrics will be printed to a txt file. Metrics that are obtained in the previous runs and the current metrics will be combined.
- System will print its actions to the command line and a log file one by one.
- Metrics and log file will be updated after each label assignment.
- Simulation can be stopped at any time and the reports can be accessed.

2.1.2 Could Have Requirements

- Improved artificial intelligence of the simulated labeling operations.
- There can be different types of labeling operations for different types of user types.

2.2 Non-Functional Requirements

- Performance
- Portability
- Availability
- Security
- Safety
- Maintainability

- Reusability

2.3 Simulation Rules/Specifications

This is an insight into the way the group aims to achieve the implementation of the labeling system. This system provides information about different outcomes which about the labels may occur at the end of the simulation to customer with different types of users. Calculated results give an idea about quality of the labeling operations and the quality of the users. This system may provide an information about how to enhance and upgrade the labeling mechanism. This is not a comprehensive list; it will be updated once the team has prepared the detailed design, class diagram and the sequence diagrams for the simulation.

2.3.1 Simulation Guidelines

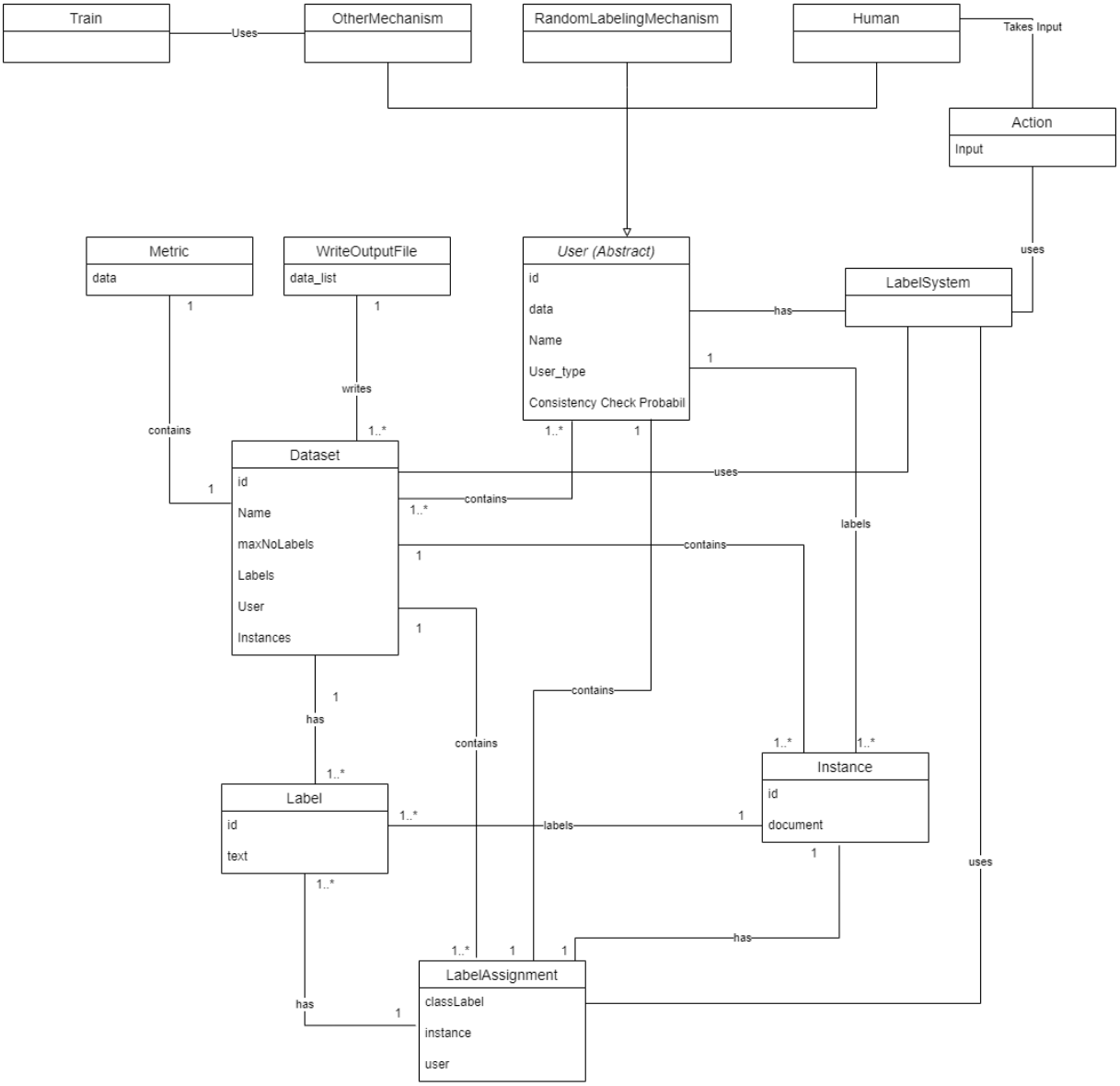
- The system starts the simulation with reading configuration and input files.
- A particular instance can be labeled more than one user and a user can label more than one instance.
- In this system, users can be more than one different type.
- Datasets have an information about maximum number and according to this information, an instance can be labeled as much as this number.
- Number of labels per instance cannot be less than one.
- User randomly chooses one of the labels from the set of labels and chooses an instance which will be tagged.
- Labeling operation of the user is given after each label assignment.
- Labeling operation continues until all the instances are labeled by all users at least once.
- Statistics are calculated after each label assignment.
- Statistics are collected and the results are given at the end of the labeling operation.
- The resulting metrics consists of user, instance and dataset performance metrics. These metrics gives detailed information about how much of the

data is completed while labeling and which users did what during the simulation process.

- After each termination by user, the simulation continues where it last left off. After each simulation the new metrics are overridden onto the previous ones.
- Users can be compared according to collected statistics.

3. Glossary

- Instance: Like samples, examples, records, documents
- Label: A descriptive or identifying word or phrase.
- Dataset: Group of instances and labels.
- LabelingAssignment: Holds required information about instances.
- LabelSystem: A system which holds all labeling operations.
- User: A type of bot that is behave like human who labels
- In addition to random labeling user, a machine learning user can label instances in a given dataset. But needs a trained data in order to do so.


Train
Uses
OtherMechanism
RandomLabelingMechanism
Human
Takes Input
Action
Input
Metric
data
WriteOutputFile
data_list
User (Abstract)
id
data
Name
User_type
Consistency Check Probabil
LabelSystem
has
uses
writes
contains
Dataset
id
Name
maxNoLabels
Labels
User
Instances
labels
Label
id
text
Instance
id
document
LabelAssignment
classLabel
instance
user
1
1..*