DATA LABELING SYSTEM
REQUIREMENTS ANALYSIS DOCUMENT

VISION

Our vision for Data Labeling System is assigning predetermined labels to a dataset. Our system will be developed as a simulation that takes arguments of user configuration file and input file. System has used by the bots, so our mechanism should be like randomization factor. System will consider statistics for users, compares them then calculates metrics for instances. It will give a report about user, instance, and dataset performance metrics. Simulation runs until interrupted by user or random labeling mechanism is done.

PROBLEM STATEMENT

Main objective is developing solution for some classification problems related to labeling mechanisms in datasets such as sentiment classification problem. After first iteration, the problem evolves into collecting statistics for users and finding quality of data labeling mechanism and users.

SCOPE

Our Data Labeling System will be simulation for the second iteration again. This system will be getting datasets from a JSON file. Configuration JSON file has id, name and file path for adding datasets to the system and these attributes will be provided by user.

There are some labeling mechanism for processing. Labeling mechanism will be random label mechanism which randomly chose one of the labels from the set of labels and assigns it to the instance then calculates frequency of this labels to find final value of instances.

After processing all instances, the system will create output Json files for metrics, states, logs.

SYSTEM CONSTRAINTS

Our system will run on any Java IDE as a console application with customized Json library original library was Json-Simple.

We must simulate at least 3 users.

Datasets can be added by providing id, name, file path.

The current dataset that will be processed is provided by using the JSON file.

STAKEHOLDERS

Murat Can Ganiz (Customer)

Lokman Altın (Customer)

Edanur Öztürk

Feyza Nur Bulgurcu

Sueda Bilen

Ömer Erkan

Zehra Kuru

Farouk Tijjani Mohammed Deribe

GLOSSARY OF TERMS (ALPHABETICALLY LISTED)

Consistency: The quality of always performing in a similar way.

Dataset: The structure that keeps labels and instances.

Log History: Object that holds history of all labels of instances.

Instance: Content that need to be labeled in dataset and it must be labeled one time.

Json File: The file which includes meta data.

Label: Content that has used for categorizing the instances.

Labeling Mechanism: Method for labeling the instances.

Metric: Collections of statistical data about results of the system.

Negative: Indicates instance content as negative.

Notr: Indicates instance content as notr.

Simulation: Is an approximate imitation of the operation of a process or system

Positive: Indicates instance content as positive.

Probability: The percentage of revision when it is labelled again.

Random Labeling Mechanism: The mechanism that label instances randomly.

User: who runs the system.

USE CASES

1) User executes the Data Labeling System.

2) Simulation starts.

3) Dataset will be uploaded to system.

4) For each user, system decides labeling mechanism based on user’s type

5) For each user, system randomly links some instances to users.

6) Random Labeling Mechanism will choose labels from the dataset randomly and assigns it to an instance.

7) System creates a log for the labeling processes.

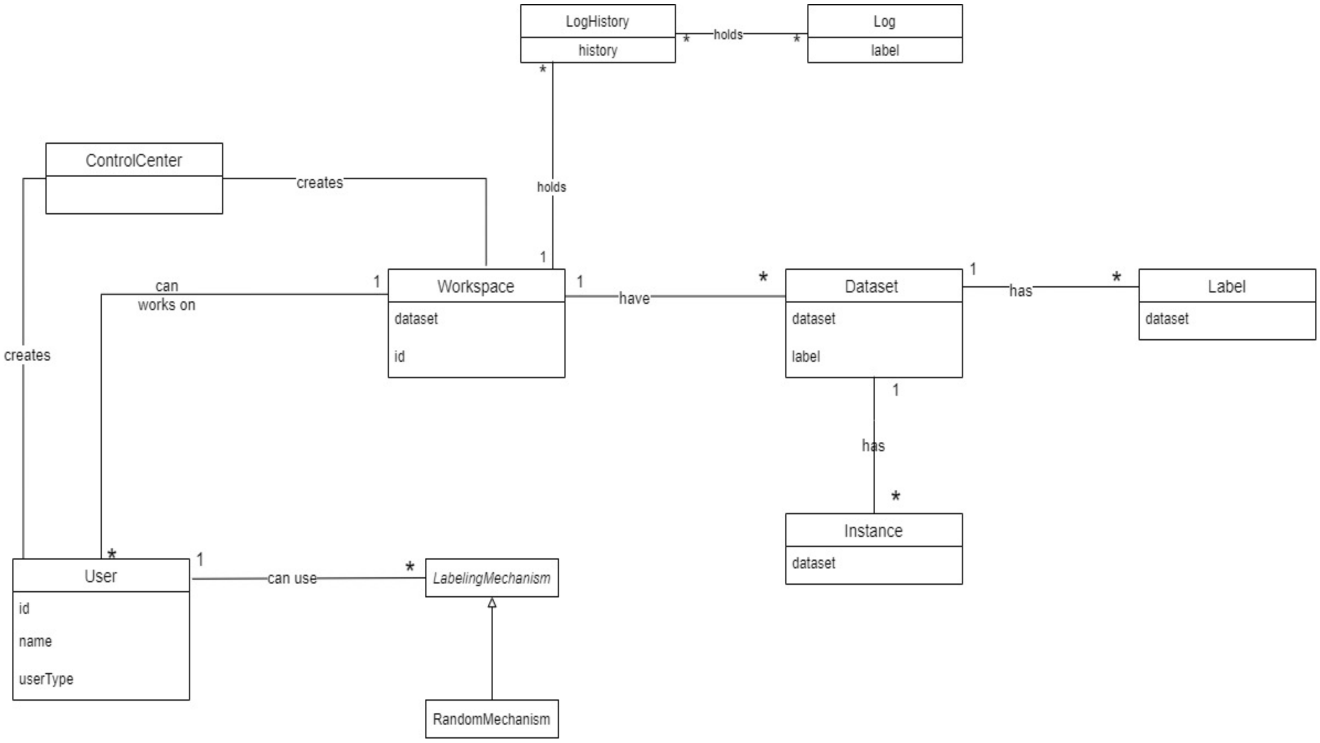
8) System assigns the final value of instances by using the users’ label assignment history.

9)System calculates the metrics after each labelling process and writes on json file.

10)When there is no remaining possibilities system stop the simulation or can be interrupted by user.

11) System prints output files based on assignments.

DOMAIN CLASS DIAGRAM


LogHistory
history
*
holds
*
Log
label
*
holds
1
ControlCenter
creates
creates
can
works on
1
Workspace
dataset
id
1
have
*
Dataset
dataset
label
1
has
*
Label
dataset
1
has
*
Instance
dataset
*
1
User
id
name
userType
can use
*
LabelingMechanism
RandomMechanism