SMART Documentation

Release 3.0.0

Rob Chew

Apr 28, 2023

Tutorial

SMART is an open source application designed to help data scientists and research teams efficiently build labeled training datasets for supervised machine learning tasks.

CHAPTER 1

Feature Highlights

- **Active Learning** algorithms for selecting the next batch of data to label.
- **Label Suggestions** to suggest likely labels for each data item, using embeddings.
- **Inter-rater reliability** metrics to help determine a human-level baseline and the understand the test validity of your labeling task.
- **Admin dashboard** and other project management tools to help oversee the labeling process and coder progress.
- **Multi-user coding**, for parallel annotation efforts within a project.
- **Self-hosted installation**, to keep sensitive data secure within your organization's firewall.

CHAPTER 2

Quick Start

```
$ git clone https://github.com/RTIInternational/SMART.git
$ cd smart/envs/dev/
$ docker-compose build
$ docker volume create --name=vol_smart_pgdata
$ docker volume create --name=vol_smart_data
$ docker-compose run --rm backend ./migrate.sh
$ docker-compose up -d
```

Open your browser to http://localhost:8000

2.1 Part 1: Installation

Note: Additional installation instructions and developer notes are available at the SMART Github code repository.

To begin installing SMART, first clone the code repository to the local directory of your choice:

```
$ git clone https://github.com/RTIInternational/SMART.git
```

SMART uses Docker in development to aid in dependency management. First, install Docker and Docker Compose. Then navigate to `envs/dev` and run `docker-compose build` to build all the images:

```
$ cd smart/envs/dev
$ docker-compose build
```

Next, create the docker volumes where persistent data will be stored: `docker volume create --name=vol_smart_pgdata` and `docker volume create --name=vol_smart_data`.

```
$ docker volume create --name=vol_smart_pgdata
$ docker volume create --name=vol_smart_data
```

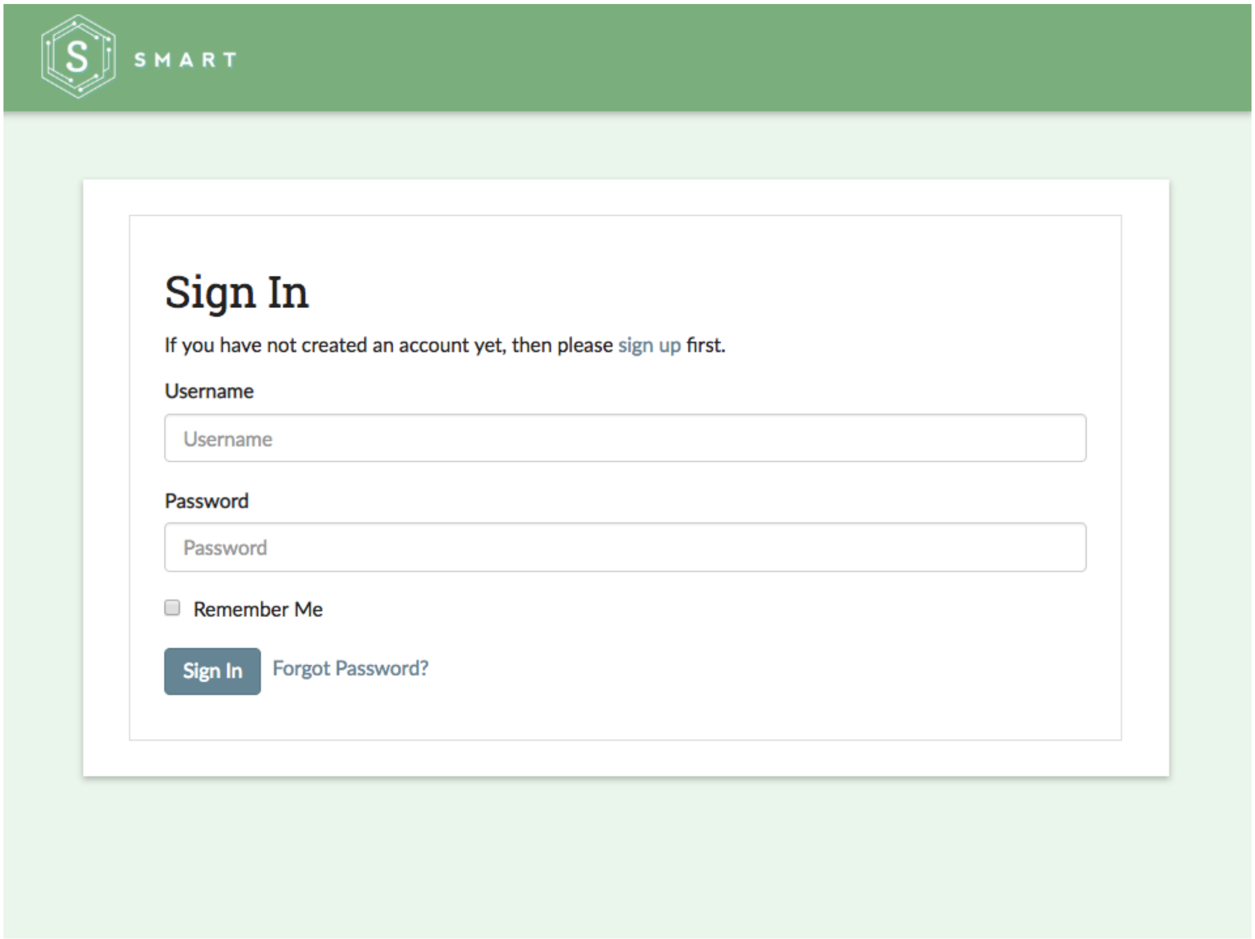
Then, migrate the database to ensure the schema is prepared for the application.

```
$ docker-compose run --rm backend ./migrate.sh
```

Lastly, run `docker-compose up` to start all docker containers. This will start up the containers in the foreground so you can see the logs. If you prefer to run the containers in the background use `docker-compose up -d`. When switching between branches there is no need to run any additional commands (except build if there is dependency change).

```
$ docker compose up -d
```

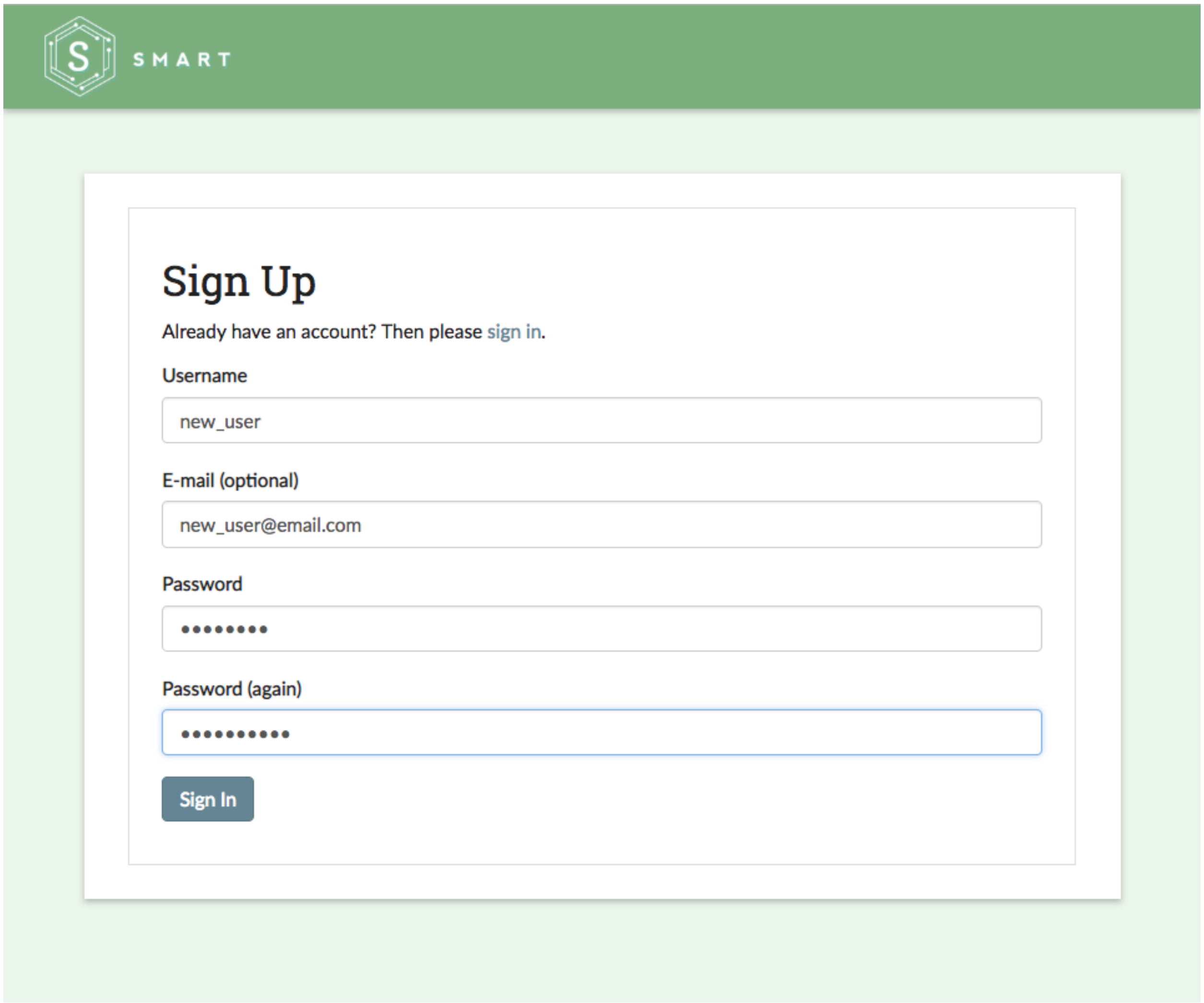
To see SMART in action, navigate to http://localhost:8000 in your web browser of choice. You should be welcomed by the SMART login screen:



If you have not created an account yet, then please sign up first.

Note: By default, SMART will use port 8000 for the front-end and port 5432 for the back-end processes. See the SMART code repository README for instructions on how to change the default ports.

Finally, create a profile to start your own new labelling projects or to be added to an existing one:



Sign Up

Already have an account? Then please sign in.

Username

new_user

E-mail (optional)

new_user@email.com

Password

••••••••

Password (again)

••••••••••

Production Installation

SMART provides both development and production docker builds, depening on user need. The production setup is fairly close to the development one, with some changes that make it easier for the application to run in production. One noteable difference is that celery is currently not supported for the production build due to some complications it can introduce when running the application in production on a managed server. See *envs/prod/README.md* for more information, including instructions on setting up regular backups of the production database.

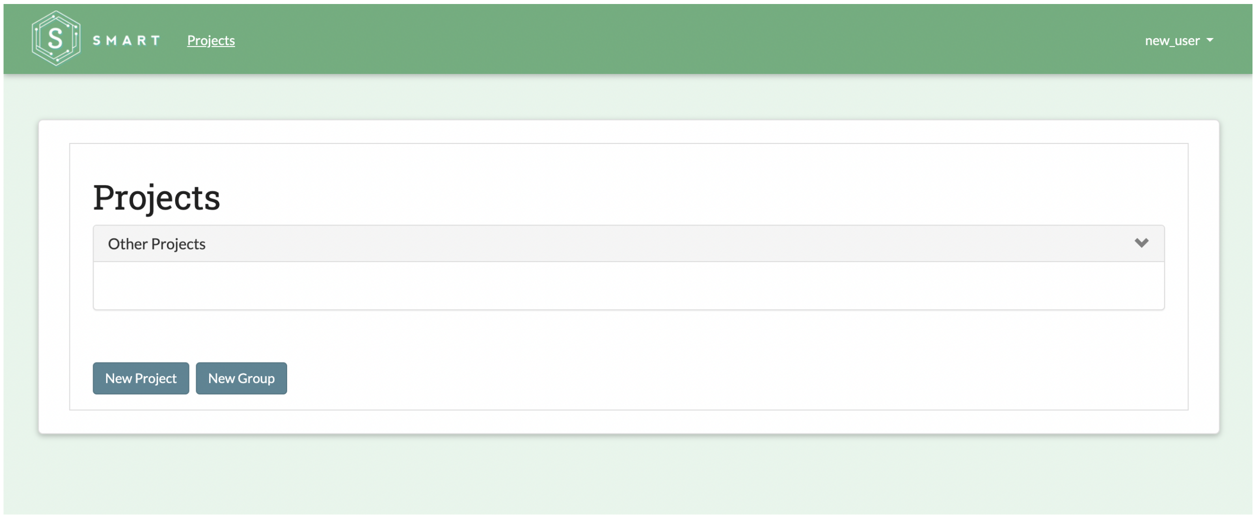
Note: If you intend to run SMART on a server where other processes may be running, you may run into problems with ports already being in use. This can be fixed by changing the default SMART ports in the docker compose files.

Warning: Please note that there may be delays in updating SMART's dependencies between releases. If your project includes sensitive data, we recommend running SMART within a closed, secure network environment.

2.2 Part 2: Creating a New Project

Starting a new labelling project in SMART is as easy as pressing the "New Project" button on the SMART landing page. All users have the ability to start their own coding projects, though they may be restricted to modifying or deleting existing projects depending on their user roles.

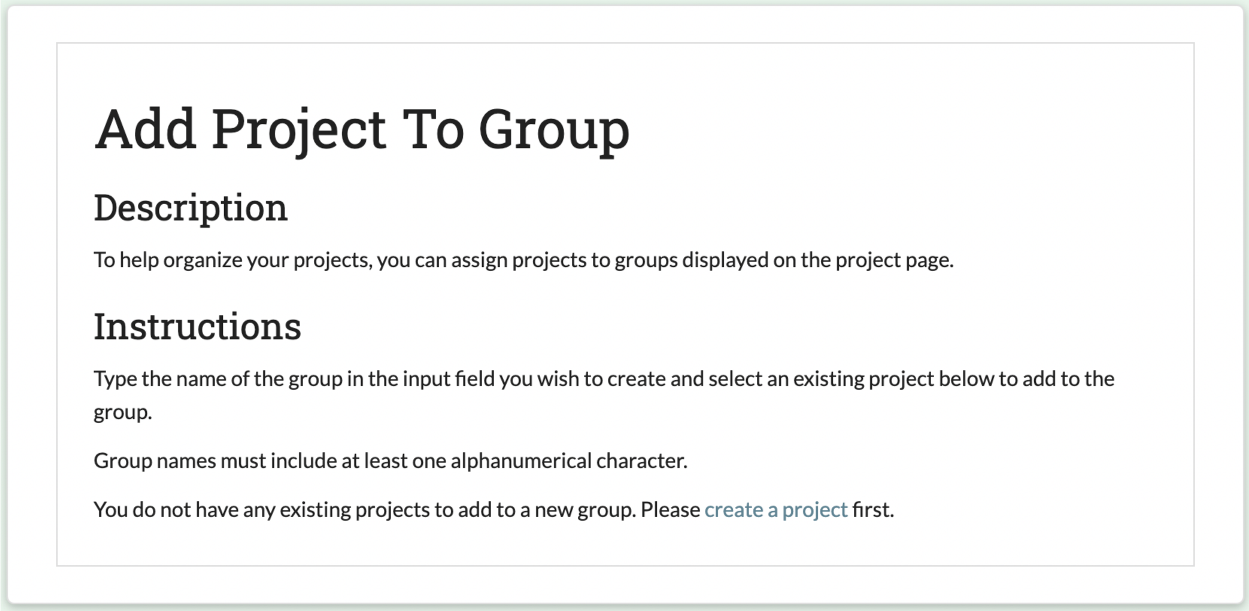
For new users without any existing projects, the SMART landing page should look like this:



For the purpose of this tutorial, we'll create some projects to classify Reddit posts to see if they are about cats. Below we'll give instructions on how to create a classifier project and a general labeling project where we do not want to use a classifier.

2.2.1 Project Groups [NEW]

As of SMART 3.0.0, projects can now be grouped. By default any project not assigned a group will be put in the "Other Projects" group. Project groups can only be created and assigned once at least one project exists. At this point, if the "New Group" button is pressed you should see the following page:



Add Project To Group

Description

To help organize your projects, you can assign projects to groups displayed on the project page.

Instructions

Type the name of the group in the input field you wish to create and select an existing project below to add to the group.

Group names must include at least one alphanumerical character.

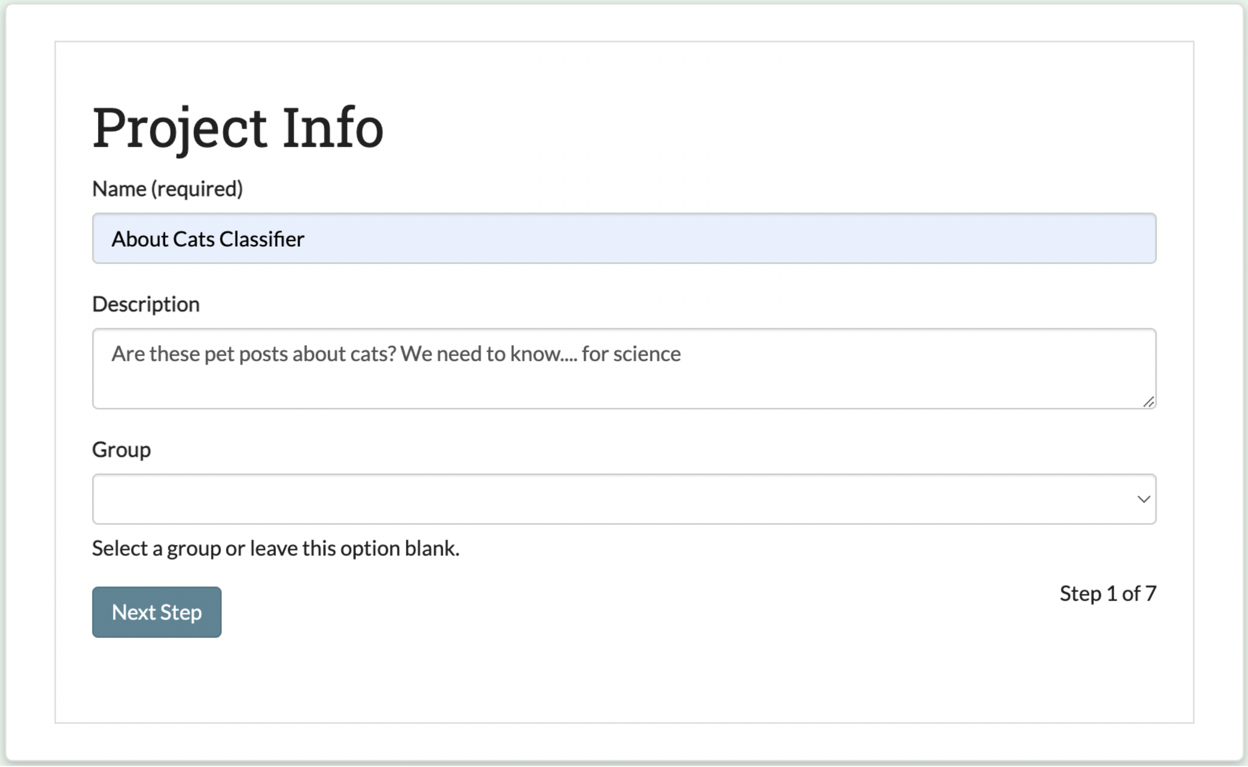
You do not have any existing projects to add to a new group. Please create a project first.

For instructions on how to use project groups see *Project Groups*.

2.2.2 Project Description

The first step in creating your project is to provide a project name and description. The name will be the internal reference for the project which users will see on their landing pages and the description will be available for users on the project Details page. Below, we fill out the name and description for two projects:

- An "About Cats Classifier" project, which will use the modeling features to classify posts in one of three ways, and,
- An "About Cats Labeling" project which will classify posts in many ways, though not use a classifier



Project Info

Name (required)

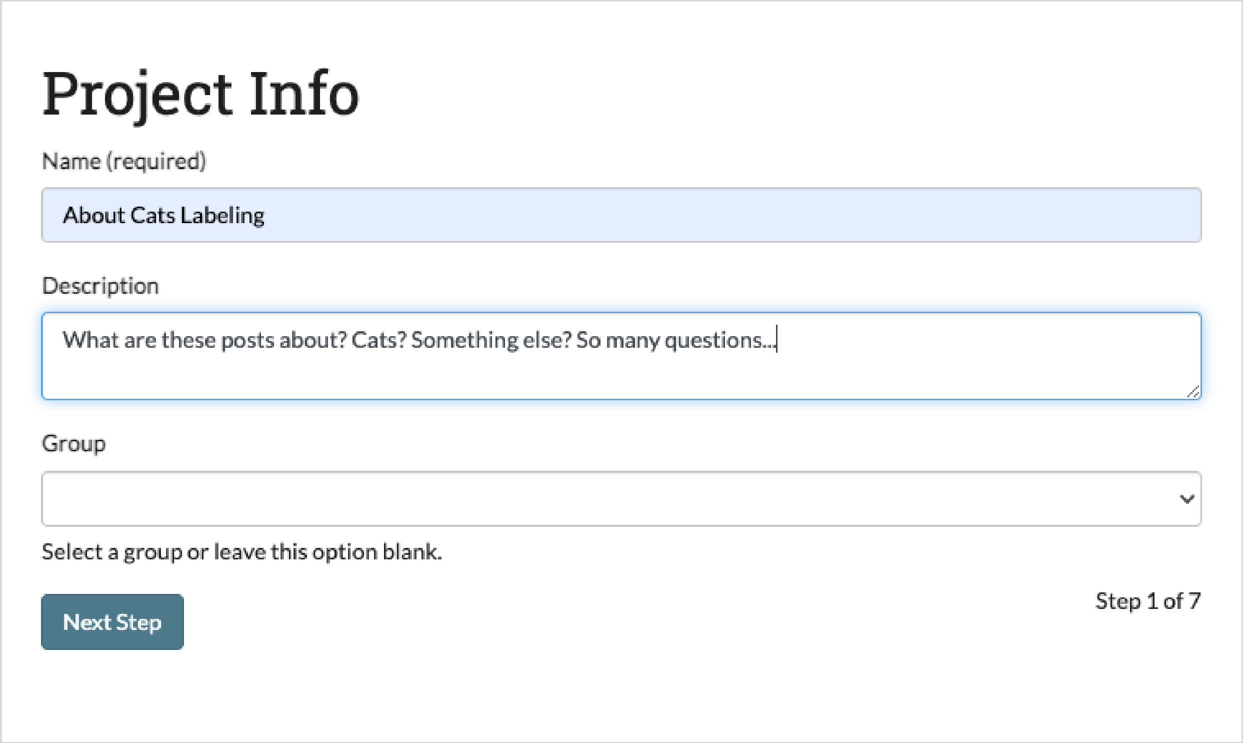
About Cats Classifier

Description

Are these pet posts about cats? We need to know.... for science

Group

Select a group or leave this option blank.



Project Info

Name (required)

About Cats Labeling

Description

What are these posts about? Cats? Something else? So many questions...

Group

Select a group or leave this option blank.

2.2.3 Creating Label Definitions

In the Labels section, we will create categories for labeling. These labeled observations can then be used to train a classification model that predicts into what category a new observation belongs. To add new categories, just fill-in the names of the categories you're interested in predicting into the input boxes. If you have more than two labels, use the "add label" link to add more rows to the form. If you decide that you want to remove a label after adding it, use the "remove label" link to remove the label name.

You can also upload a file with labels and descriptions by clicking the "Choose File" button.

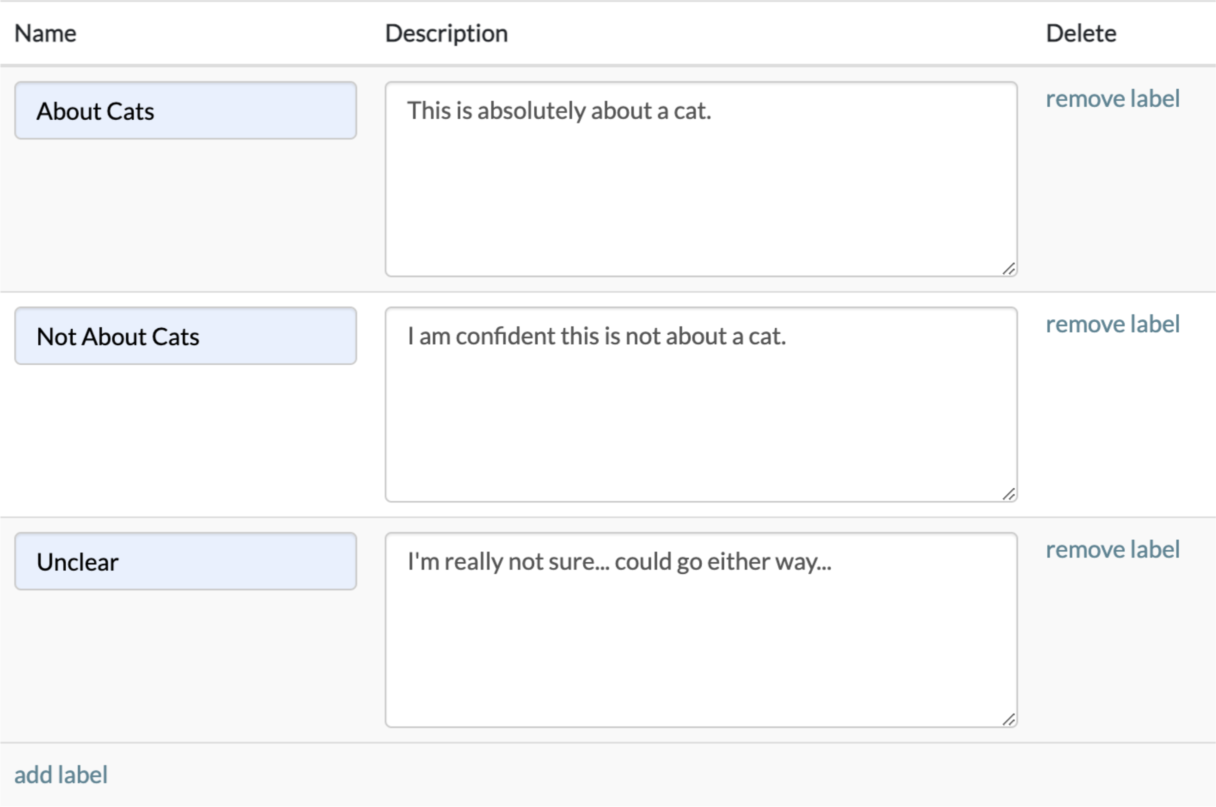
Note:

- SMART requires at least two category labels and the labels must be unique.
- If you plan on uploading a data file that contains labels, the label categories in the file must match those provided on this page.
- You may add up to 10,000 labels to each project.
- Label files must contain `Label` and `Description` columns.
- The label file must be a CSV (comma separated values) file.

Warning:

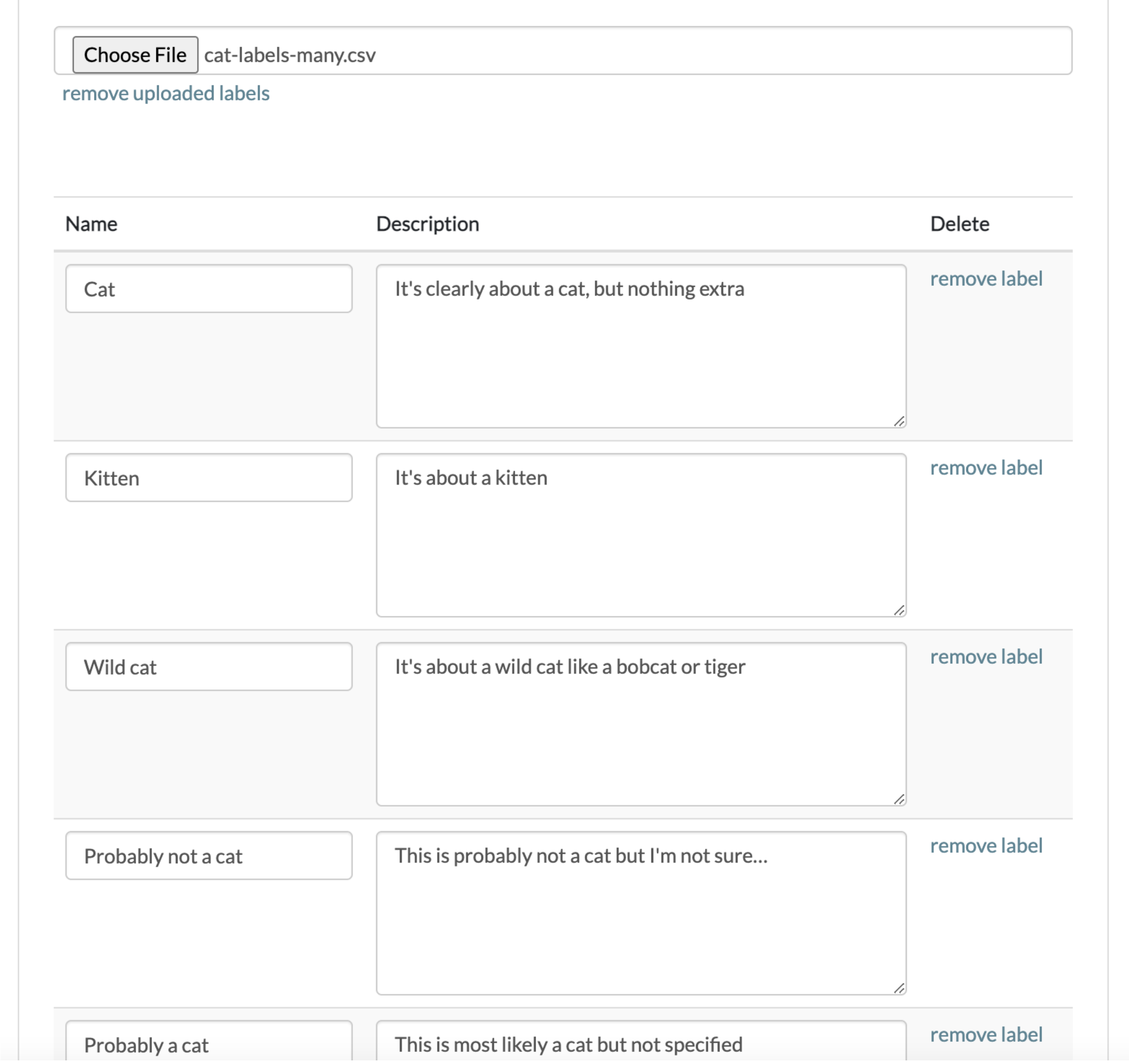
- You cannot add, remove, or update any label categories for a project after the project is created. However, you can always export your labeled data and upload into a new project to add or remove new label categories as needed.

For our first project “About Cats Classifier”, we only have three labels which we want SMART to predict, so we fill them in below:



Name	Description	Delete
About Cats	This is absolutely about a cat.	remove label
Not About Cats	I am confident this is not about a cat.	remove label
Unclear	I'm really not sure... could go either way...	remove label

For our second project “About Cats Labeling”, we have many more labels, which we will fill in using a file.



cat-labels-many.csv

Name	Description	Delete
Cat	It's clearly about a cat, but nothing extra	remove label
Kitten	It's about a kitten	remove label
Wild cat	It's about a wild cat like a bobcat or tiger	remove label
Probably not a cat	This is probably not a cat but I'm not sure...	remove label
Probably a cat	This is most likely a cat but not specified	remove label

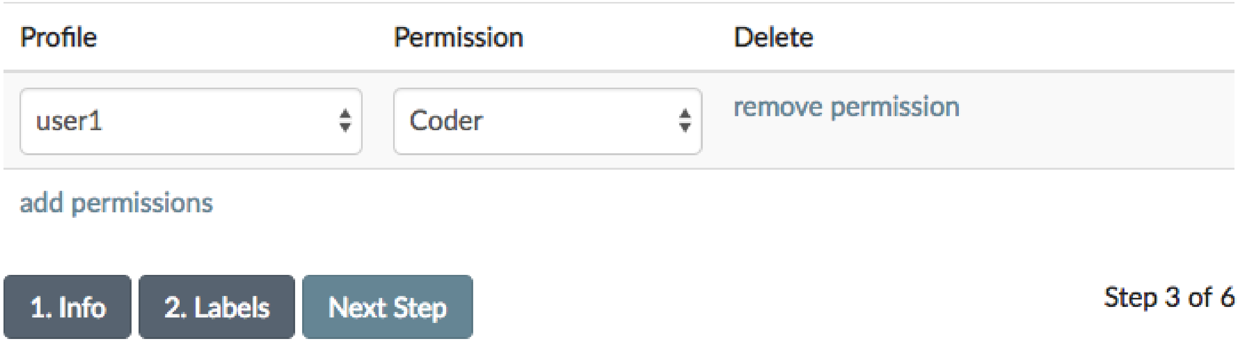
2.2.4 Project Permissions

To help organize your labeling projects, you can assign special permissions to other project members. Project members can be assigned one of two user-roles:

- **Admins** are able to update the project description, upload additional data, control project permissions, and annotate data.
- **Coders** are able to view project details and annotate data.

In this panel, you can select project members and assign their role types. Clicking the “add permissions” link adds more rows to the form. If you decide that you want to remove a permission after adding it, click the “remove permission” link next to the inputs to remove the permission. If an intended project member is not listed below, please check to see if they have created an account.

In the development environment, SMART includes three user profiles for testing purposes (`root`, `user1`, and `test_user`). Inviting additional users to a project is optional. For the purposes of this tutorial, we will add `user1` as a coder for both projects:



Profile	Permission	Delete
user1	Coder	remove permission

add permissions

Step 3 of 6

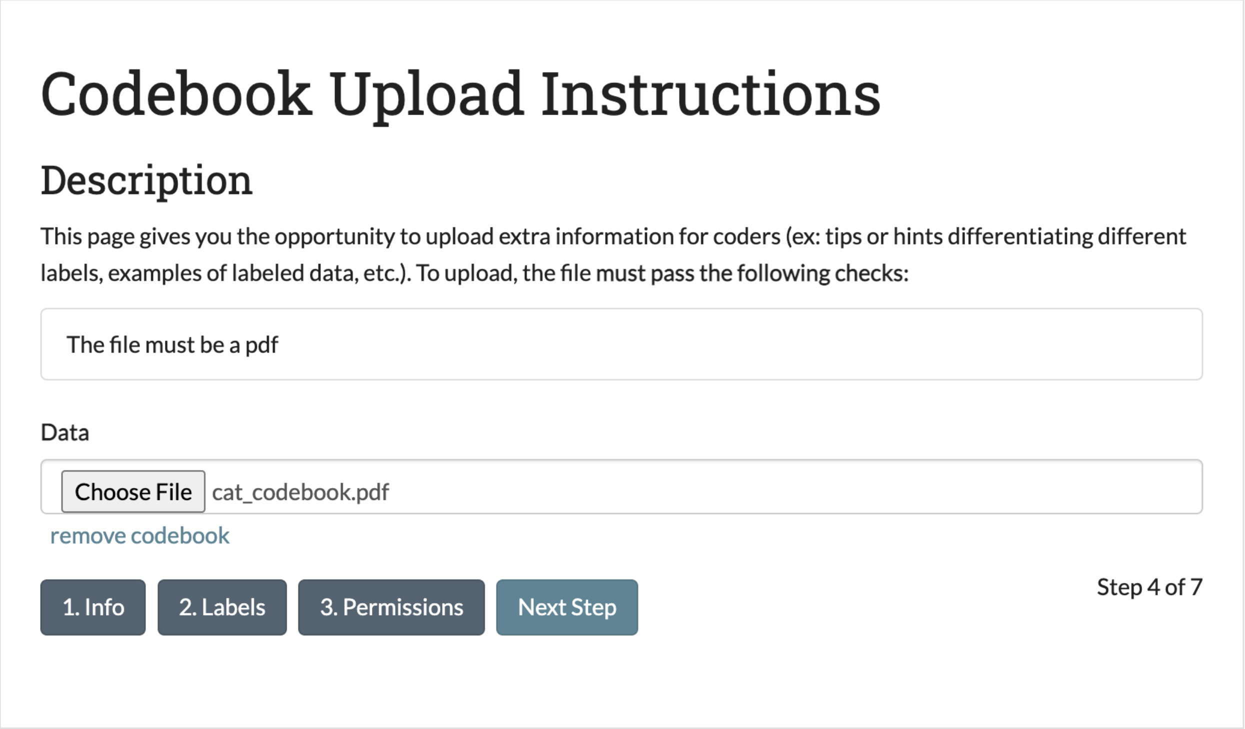
Note:

- The project creator is always assigned Admin privileges.
- Each user profile can only be assigned one permission type.
- Each row must be completely filled in with both a profile and permission.
- You can update permissions after creating a project.

2.2.5 Adding Codebook

This page gives you the opportunity to upload extra information for coders that maybe helpful for clarifying the labelling task (e.g., tips for differentiating categories, examples of labeled data, etc.). This is particularly useful if the categories you’re interested in labelling are numerous or nuanced.

A demo codebook for the tutorial can be found in the `smart/demo/` directory, which we will be using for both projects. To upload the codebook, click the “Choose File” button and select `cat-codebook.pdf`:



Codebook Upload Instructions

Description

This page gives you the opportunity to upload extra information for coders (ex: tips or hints differentiating different labels, examples of labeled data, etc.). To upload, the file must pass the following checks:

The file must be a pdf

Data

Choose File cat_codebook.pdf

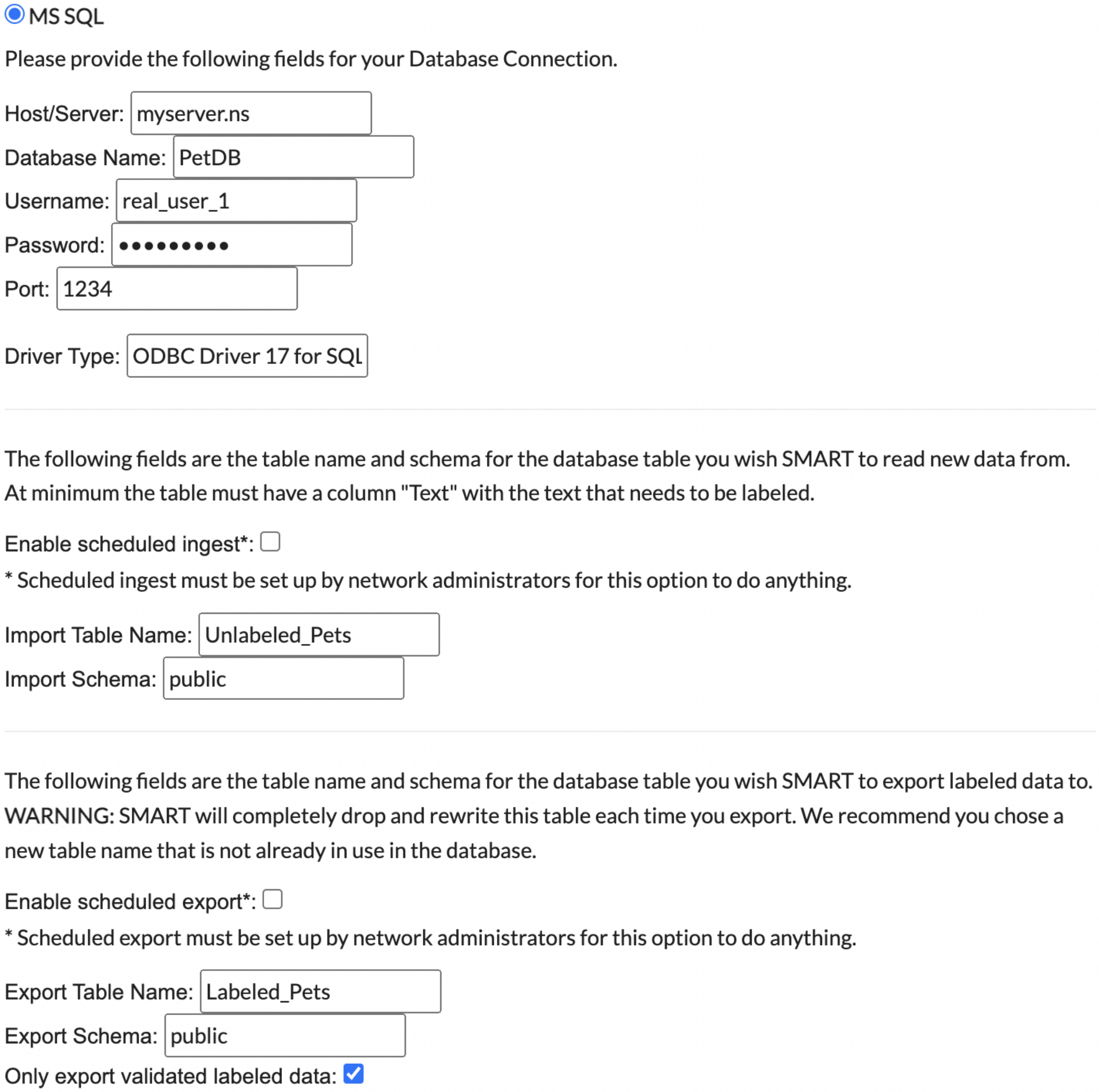
Note: The codebook file must be a PDF.

2.2.6 Setup Database Connection (optional) [NEW]

New in SMART 3.0.0, this page lets you connect SMART to an MSSQL database. This lets you provide a table in the database that you want SMART to pull from when adding data to the project. You may also name a new table which SMART will export labeled data to. SMART will error if the export table already exists.

Note: A database connection can be set up or removed any time after the project has been created by going to the *Updating a Project* page.

Below is an example of setting up a database connection. For our Cat projects, we will not be setting up a connection.



● MS SQL

Please provide the following fields for your Database Connection.

Host/Server: myserver.ns

Database Name: PetDB

Username: real_user_1

Password: •••••••••

Port: 1234

Driver Type: ODBC Driver 17 for SQL

The following fields are the table name and schema for the database table you wish SMART to read new data from. At minimum the table must have a column "Text" with the text that needs to be labeled.

Enable scheduled ingest*: □

* Scheduled ingest must be set up by network administrators for this option to do anything.

Import Table Name: Unlabeled_Pets

Import Schema: public

The following fields are the table name and schema for the database table you wish SMART to export labeled data to. WARNING: SMART will completely drop and rewrite this table each time you export. We recommend you chose a new table name that is not already in use in the database.

Enable scheduled export*: □

* Scheduled export must be set up by network administrators for this option to do anything.

Export Table Name: Labeled_Pets

Export Schema: public

Only export validated labeled data: ☑

The fields for a database connection are:

- **Host/Sever**: Place the database is hosted.
- **Database Name**: Name of the specific database to connect to.
- **Username and Password**: Credentials for a user that is authorized to connect to the database. (NOTE: for security reasons this information is not saved to the SMART internal database).
- **Port**: The port to connect through.
- **Driver Type**: The driver needed to connect to the database. Currently only MS SQL is supported.

Ingest

This section of the form is needed to set up which table and schema data should be pulled into SMART. The layout of this data uses the same rules as a regular data file upload.

Scheduled Ingest: If this button is checked, then SMART will add this project to the list of projects that will pull in new data from the indicated tables whenever the management command *ingest_database_data* is used. This will allow a server admin to set up re-occuring pulls through a service like cron. See the README in the *envs/prod* folder on the repository for more information.

Export

This section of the form is needed to set up the table and schema where labeled data should be exported. These exports contain the same fields which show up when someone downloads the labeled data on the Details or Project list page.

Scheduled Export: Just like scheduled ingest, SMART allows projects to set up scheduled export through the *export_database_data* management command.

Exporting only validated data: This checkbox affects if un-validated labeled data will be included in database exports. Labeled data can be validated in the *Annotate Data Page* through the history table or the "IRR | Requires Adjudication" tab. Resolved IRR data is automatically considered validated. By default, database exports will include all labeled data.

2.2.7 Upload Data

Time to upload your data!

SMART now provides two options for uploading data. If you have set-up a database connection in the previous step, you can select "Connect to Database and Import Table" to import your data from the ingest table you indicated in the Database Connection page. Otherwise, select "File Upload" to upload a data file to SMART.

To upload, the data must pass the following checks:

- If choosing file upload, the file needs to have either a .csv, .tsv, or .xlsx file extension.
- The file or datatable requires the data to have one column named `Text`. It can also contain a unique id column named `ID` and a label column named `Label`.
- The largest file size supported is 500MBs.
- The (optional) `ID` column should contain a unique identifier for your data. The identifiers should be no more than 128 characters.
- You may add a dataset that already contains labelled observations. However, all labels present in the upload file must be in list of categories assigned in the *Creating Label Definitions* step.

The `Text` column should contain the text you wish users to label. For our "About Cats" projects, the `Text` column will contain the post text. The `Label` column should contain any pre-existing labels for the corresponding text. If none of your data contains existing labels, then this column can be left blank or removed. Extending our example, if a lead coder has already annotated some posts with their cat outcomes, this column would contain those labeled records.

2.2.8 Project Metadata [NEW]

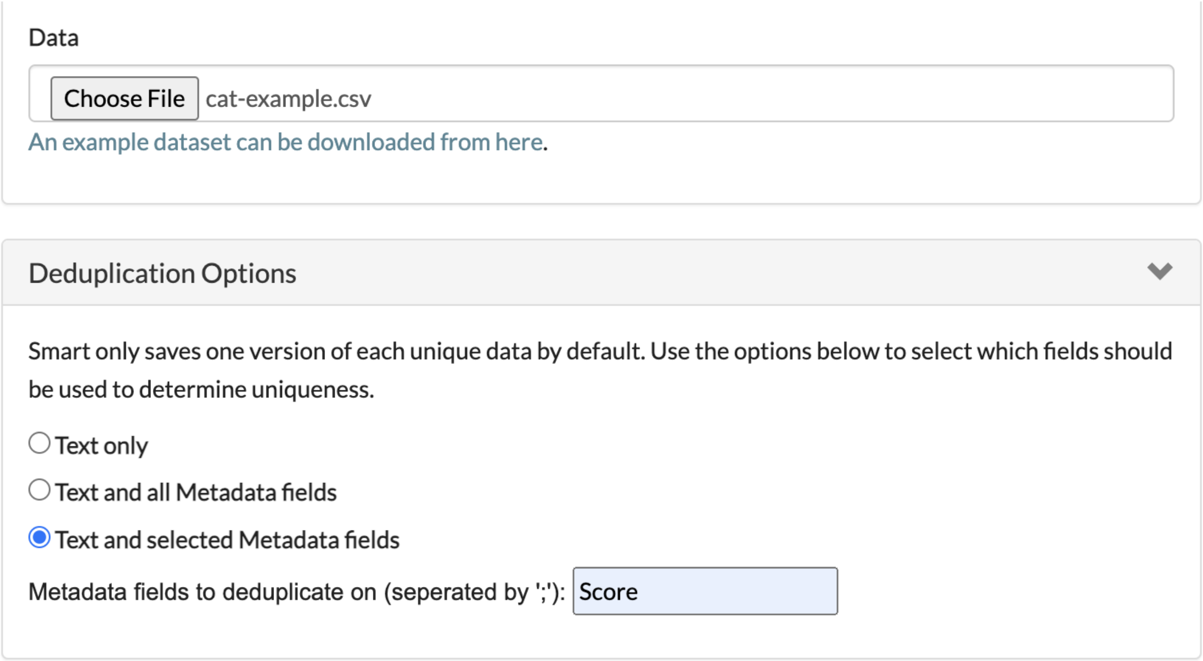
Any fields outside of `Text`, `Label`, or `ID` will be classified as "metadata," also called "respondent data." This is additional data which you would like to be presented along with the text to aid in labeling.

Some details about metadata:

Note:

- Metadata fields can have nulls.
- If you upload a file or pull in a table with metadata fields in project creation, SMART will expect all future data uploads to have those fields.

- SMART will disregard metadata fields in files uploaded after project creation if they did not exist in the first project creation upload.
- Metadata fields can be used in SMART for deduplication. In the example below, the "Score" metadata field has been selected for deduplication. This means that if two posts have the same text but different scores, they will be considered distinct entities for coding by SMART.



Data

Choose File cat-example.csv

An example dataset can be downloaded from here.

Deduplication Options

Smart only saves one version of each unique data by default. Use the options below to select which fields should be used to determine uniqueness.

○ Text only

○ Text and all Metadata fields

● Text and selected Metadata fields

Metadata fields to deduplicate on (seperated by ';'): Score

Tip:

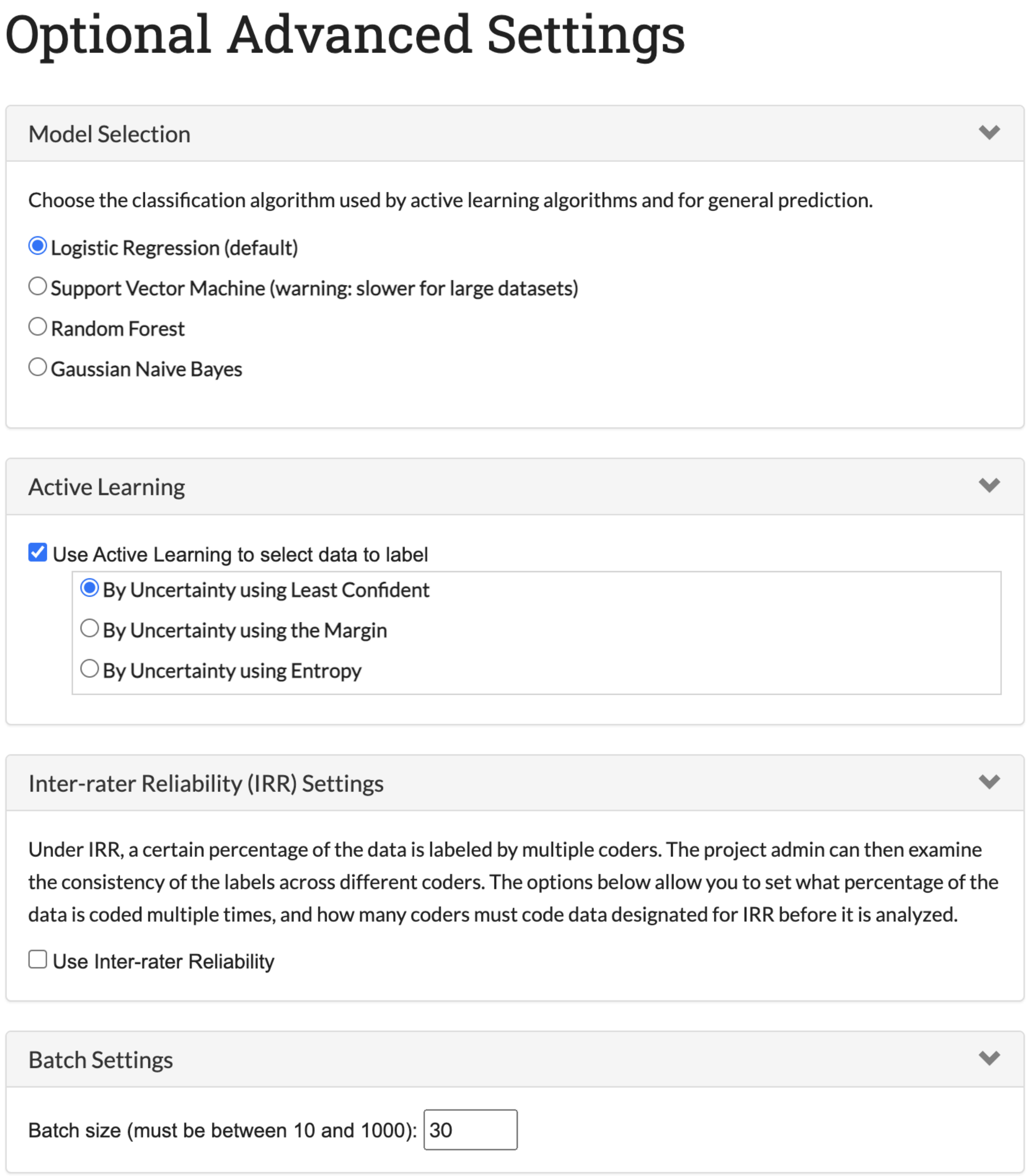
- SMART will keep up to two million unique records per data set.
- If there are multiple rows with the same text and deduplicating metadata (see above), only one of the records will be saved, and the first label, if given, will be saved.

2.2.9 Advanced Settings

The Advanced Settings page allows you to customize your labelling experience and utilize advanced features such as *Active Learning* or *Inter-rater Reliability (IRR)*.

Please reference the *Advanced Feature Details* section of the documentation to learn more about these and other options.

For our first project "About Cats Classifier" we will turn on the model selection and active learning, but leave out Inter-Rater Reliability.



Optional Advanced Settings

Model Selection

Choose the classification algorithm used by active learning algorithms and for general prediction.

● Logistic Regression (default)

○ Support Vector Machine (warning: slower for large datasets)

○ Random Forest

○ Gaussian Naive Bayes

Active Learning

☑ Use Active Learning to select data to label

● By Uncertainty using Least Confident

○ By Uncertainty using the Margin

○ By Uncertainty using Entropy

Inter-rater Reliability (IRR) Settings

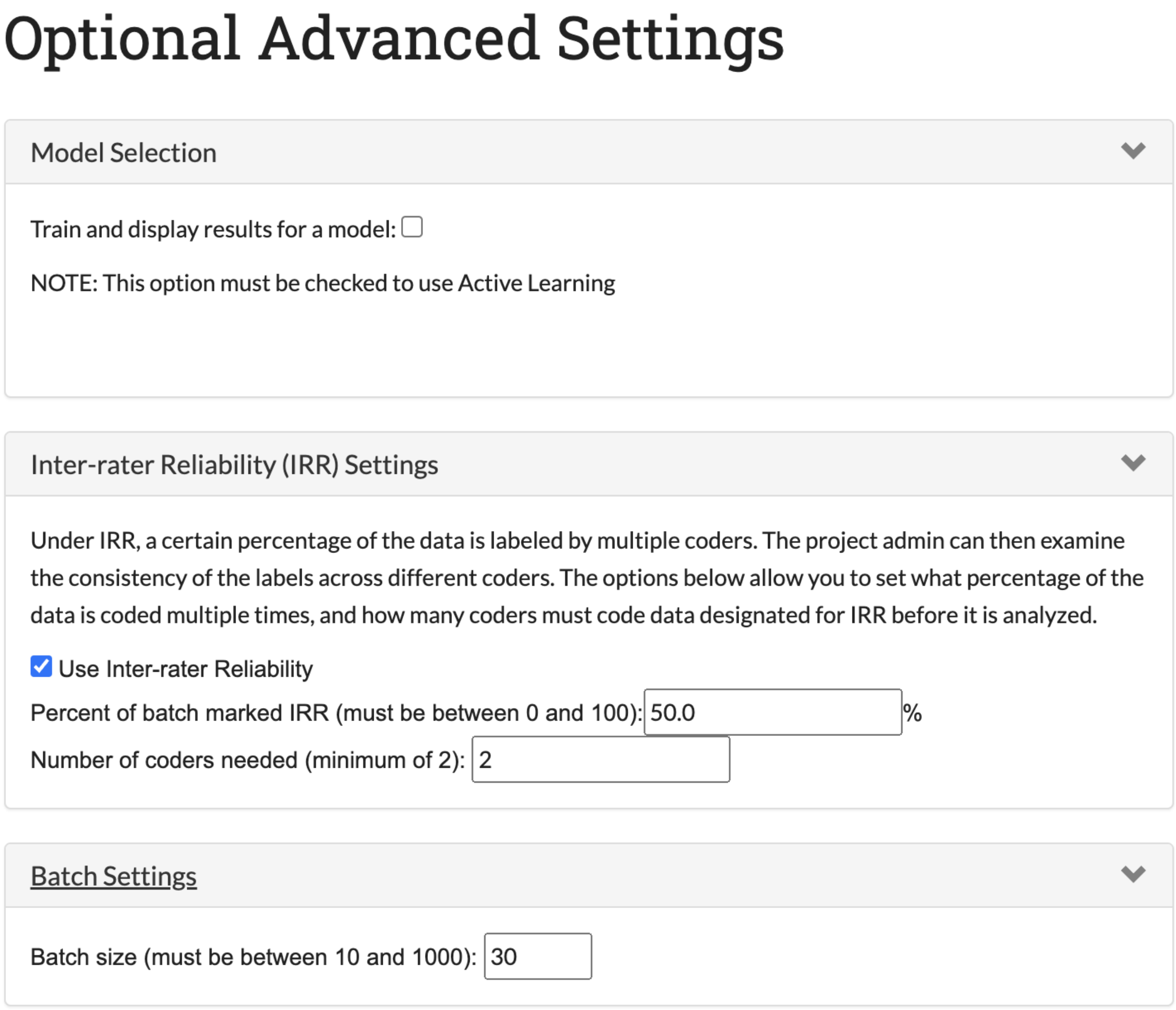
Under IRR, a certain percentage of the data is labeled by multiple coders. The project admin can then examine the consistency of the labels across different coders. The options below allow you to set what percentage of the data is coded multiple times, and how many coders must code data designated for IRR before it is analyzed.

□ Use Inter-rater Reliability

Batch Settings

Batch size (must be between 10 and 1000): 30

For our second project “About Cats Labeling” we will want to go with the default settings for Model Selection and Active learning. We will turn on Inter-Rater-Reliability and set it to 50% (so half of our posts will be double coded).



Optional Advanced Settings

Model Selection

Train and display results for a model: □

NOTE: This option must be checked to use Active Learning

Inter-rater Reliability (IRR) Settings

Under IRR, a certain percentage of the data is labeled by multiple coders. The project admin can then examine the consistency of the labels across different coders. The options below allow you to set what percentage of the data is coded multiple times, and how many coders must code data designated for IRR before it is analyzed.

☑ Use Inter-rater Reliability

Percent of batch marked IRR (must be between 0 and 100): 50.0 %

Number of coders needed (minimum of 2): 2

Batch Settings

Batch size (must be between 10 and 1000): 30

Note:

- You may be wondering… "Can I make a project with both a classifier and a large number of labels?" The answer is yes, SMART will let you do that. However, the active learning and model components of SMART only turn on when it deems you have **at least one labeled observation per category**. Even then, the model will likely be inaccurate until there are a sufficient number of labeled observations per category. For this reason, we do not recommend enabling the classifier component for projects with more than 5 label categories.

Tip: The data used in this tutorial is shipped with SMART and can be found in the `smart/demo/` directory. To upload this file, click the "Choose File" button and select `cat-example.csv`.

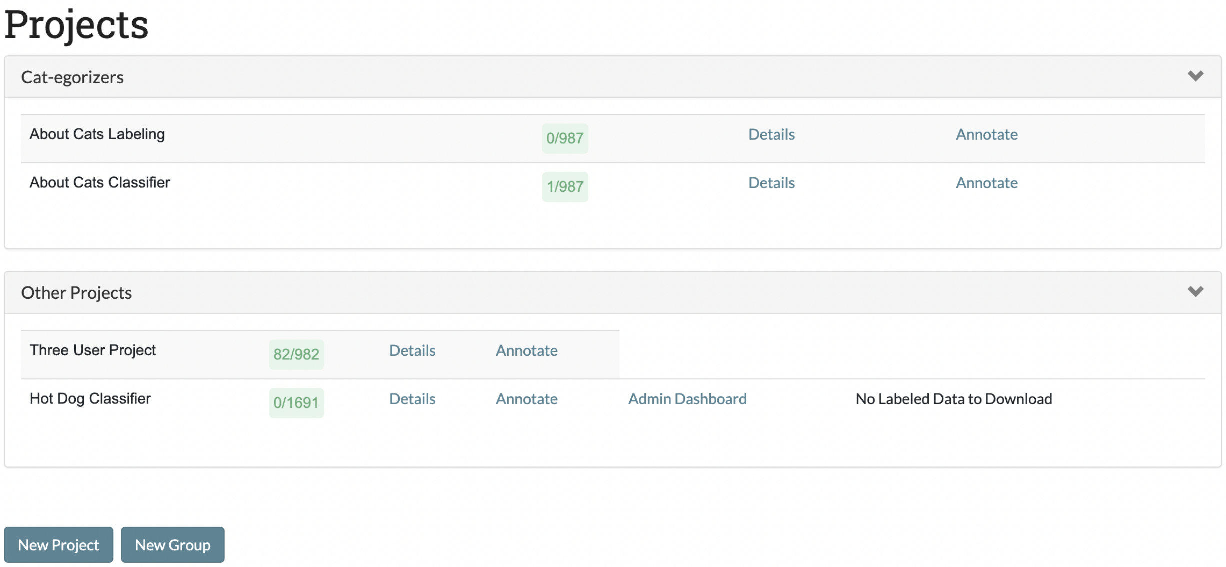
2.3 Part 3: Reviewing Projects & Editing Project Settings

2.3.1 Projects Page

The projects page serves as the central page for a SMART user. The page provides a list of all projects the user is on, and provides links to major parts of each project. Users with admin privileges will be able to see links to a project's respective *Admin Dashboard* as well as the *Download Labeled Data and/or Model* button. Coders will only see the *Details Page* and *Annotate Data Page* links. This is also the page where you go to *Create a New Project*.

The projects will also give a high level count of the portion of data in the project which has been fully labeled. Note that items in the *Recycle Bin Page* will not be counted in the denominator of the fraction.

The image below is the project page for User1. We see that the user is a coder for all projects except for the `Hot Dog Classifier` project, where it is an Admin.



Projects

Cat-egorizers

About Cats Labeling	0/987	Details	Annotate
About Cats Classifier	1/987	Details	Annotate

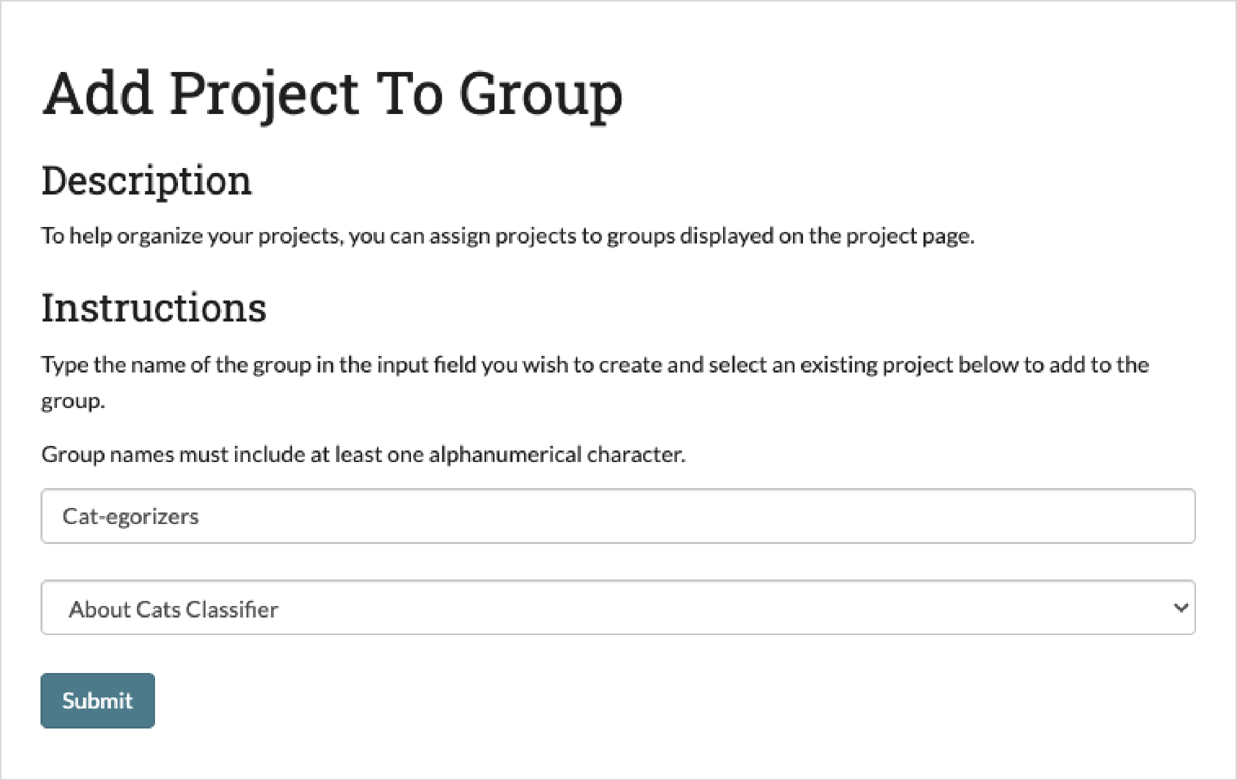
Other Projects

Three User Project	82/982	Details	Annotate		
Hot Dog Classifier	0/1691	Details	Annotate	Admin Dashboard	No Labeled Data to Download

2.3.2 Adding a Project to a Group

To add a project to a group, we must first create the group and add one project to it by clicking the "New Group" button, writing the name of the group, and adding a project.

Below we add our "About Cats Classifier" project to a new group.



Add Project To Group

Description

To help organize your projects, you can assign projects to groups displayed on the project page.

Instructions

Type the name of the group in the input field you wish to create and select an existing project below to add to the group.

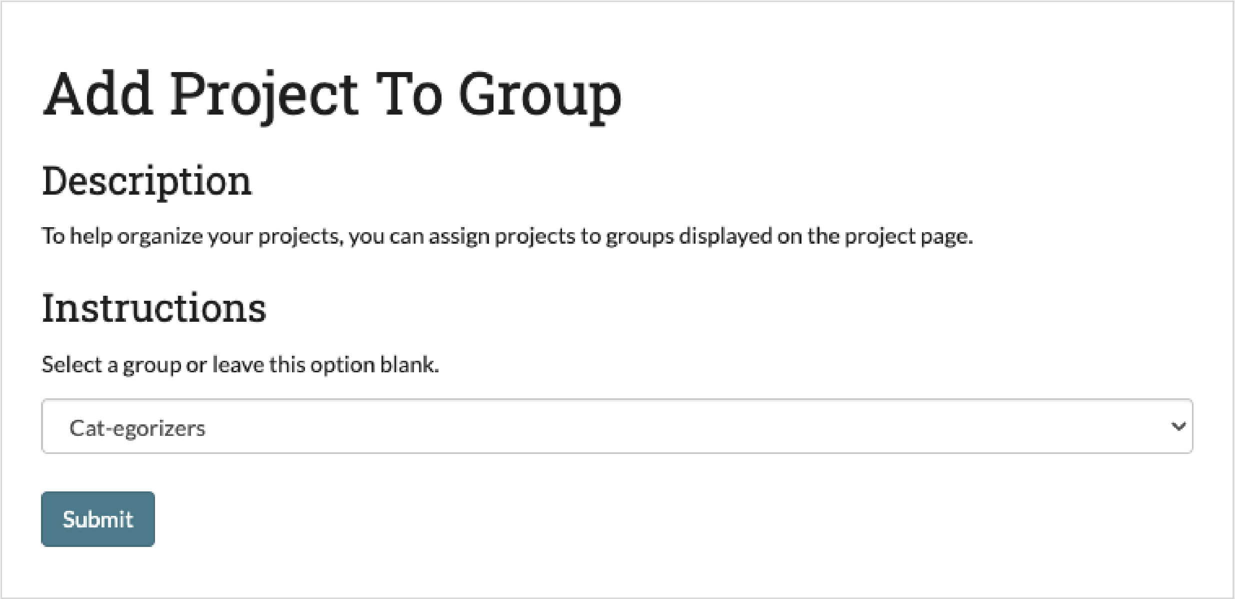
Group names must include at least one alphanumerical character.

Cat-egorizers

About Cats Classifier

You can add more projects to that group by clicking "New Group" and typing the same group name, but you can also add projects to the list of existing groups through selecting that project in the project page and then going to `Details`-> `Update Project` -> `Add Project to Group`.

Below we add our "About Cats Labeling" project to the Group we made for the "About Cats Classifier" project.



Add Project To Group

Description

To help organize your projects, you can assign projects to groups displayed on the project page.

Instructions

Select a group or leave this option blank.

Cat-egorizers

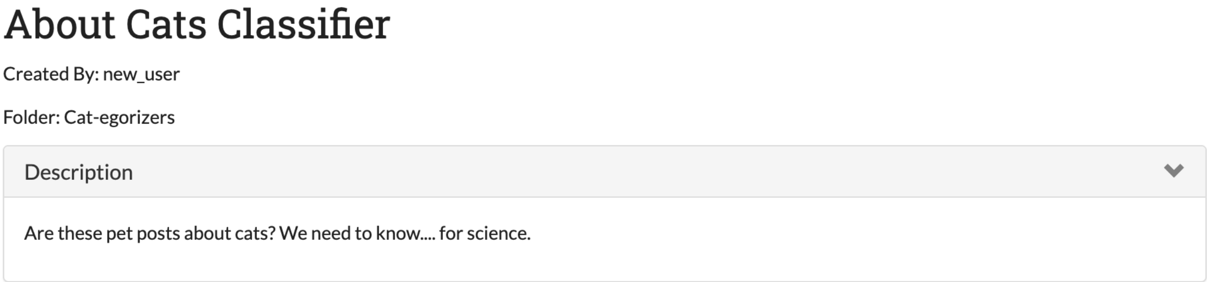
2.3.3 Details Page

The Details page provides an overview of the information and settings for your project. Each project has its own Details page, which is created when you start a new project. You can navigate to any project Details page from the *Projects Page* or to a specific Details page by pressing the "Details" link on the top navigation bar when on a project *Annotate Data Page* page or *Admin Dashboard* page.

(Examples in this section are from the "About Cats Classifier" project)

The Details page lets you review:

- The project Description.



About Cats Classifier

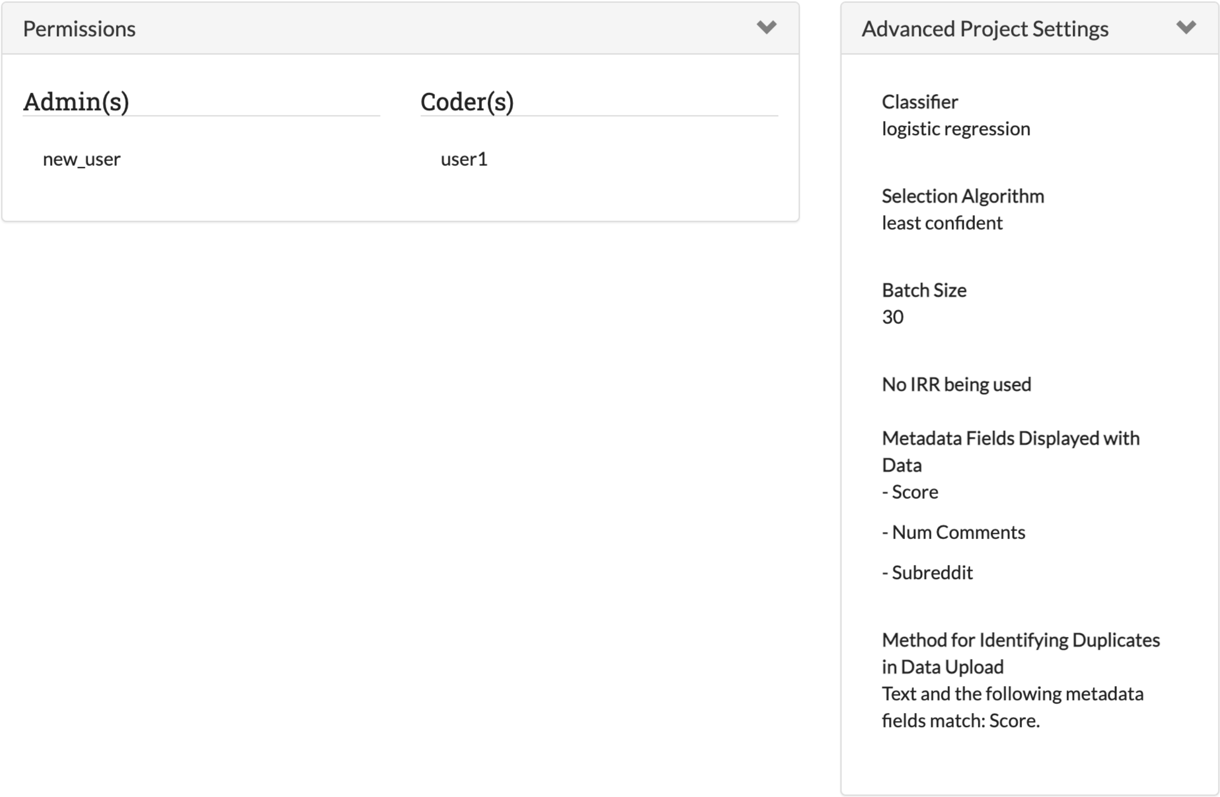
Created By: new_user

Folder: Cat-egorizers

Description

Are these pet posts about cats? We need to know.... for science.

- What permissions have been assigned to what users.
- The advanced settings (i.e. *Active Learning*, *Inter-rater Reliability (IRR)*, classifier, batch size, *Project Metadata [NEW]*, deduplication settings).



Permissions

Admin(s)	Coder(s)
new_user	user1

Advanced Project Settings

Classifier
logistic regression

Selection Algorithm
least confident

Batch Size
30

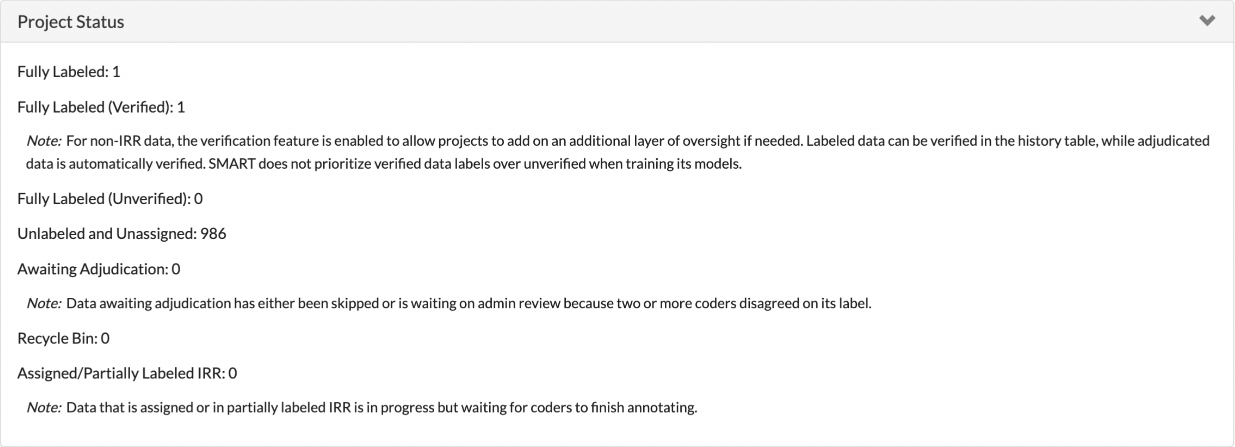
No IRR being used

Metadata Fields Displayed with Data
- Score
- Num Comments
- Subreddit

Method for Identifying Duplicates in Data Upload
Text and the following metadata fields match: Score.

- **The status of the data currently loaded in your project. This splits into:**

  - *Fully Labeled*: data which has been either labeled by one user (if not IRR data) or has received the required number of labels and they either agreed or the disagreement was resolved (if IRR data).
  - *Unlabeled and Unassigned*: This data has not been touched since it was loaded into SMART.
  - *Awaiting Adjudication*: This data is sitting in the *Admin Annotation Page* table waiting review by an administrator. It was either IRR data where the coders did not agree, or the data was sent to an administrator by a coder.
  - *Recycle Bin*: This data was dropped from the project for some reason and is not included in the data totals.
  - *Assigned/Partially Labeled IRR*: This data is in progress. Either it has been passed out to someone for coding (see *Unassign Coders [NEW]* on how to un-assign data from coders), or it is IRR data which has received some labels but not enough to be either adjudicated or resolved.



Project Status

Fully Labeled: 1

Fully Labeled (Verified): 1

Note: For non-IRR data, the verification feature is enabled to allow projects to add on an additional layer of oversight if needed. Labeled data can be verified in the history table, while adjudicated data is automatically verified. SMART does not prioritize verified data labels over unverified when training its models.

Fully Labeled (Unverified): 0

Unlabeled and Unassigned: 986

Awaiting Adjudication: 0

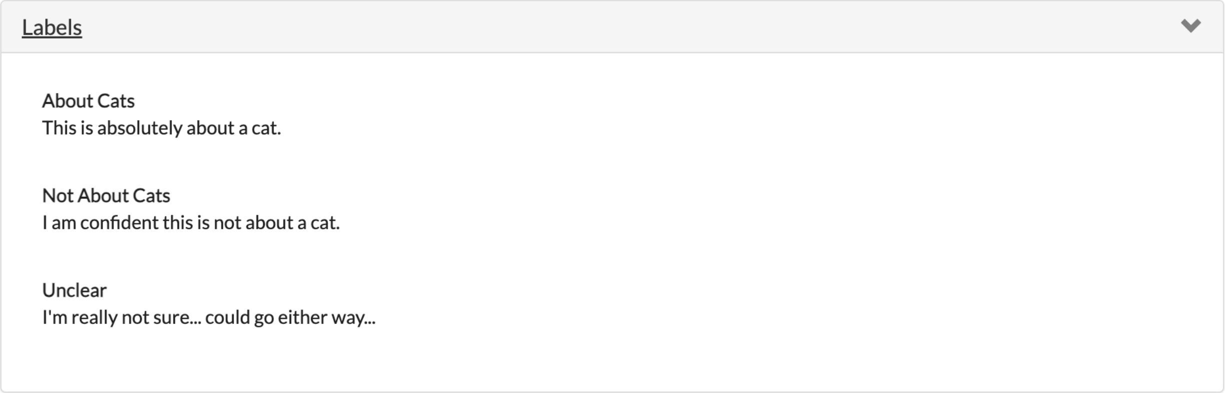
Note: Data awaiting adjudication has either been skipped or is waiting on admin review because two or more coders disagreed on its label.

Recycle Bin: 0

Assigned/Partially Labeled IRR: 0

Note: Data that is assigned or in partially labeled IRR is in progress but waiting for coders to finish annotating.

- The labels being used and their descriptions (if applicable).



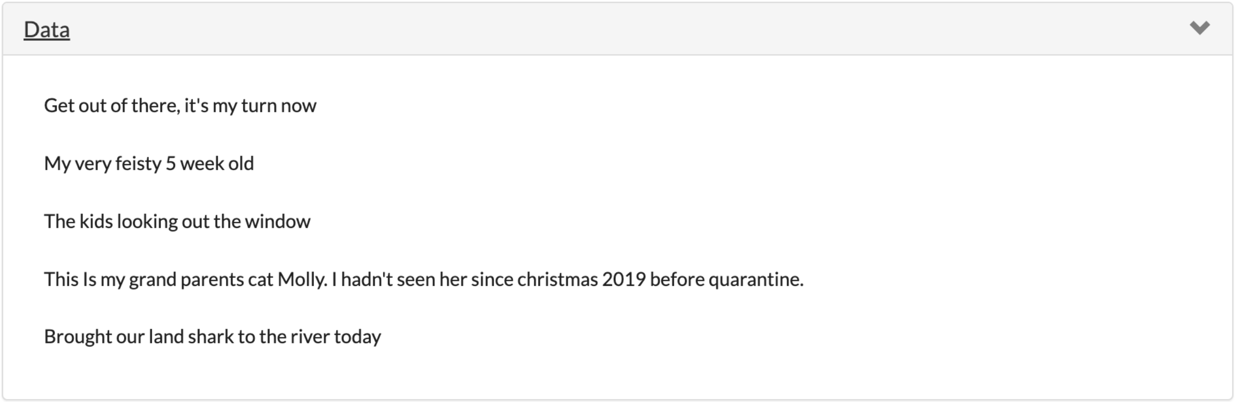
Labels

About Cats
This is absolutely about a cat.

Not About Cats
I am confident this is not about a cat.

Unclear
I'm really not sure... could go either way...

- A sample of your data.



Data

Get out of there, it's my turn now

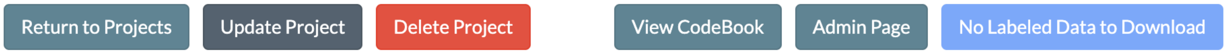
My very feisty 5 week old

The kids looking out the window

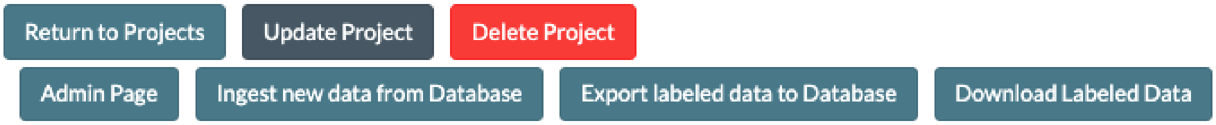
This Is my grand parents cat Molly. I hadn't seen her since christmas 2019 before quarantine.

Brought our land shark to the river today

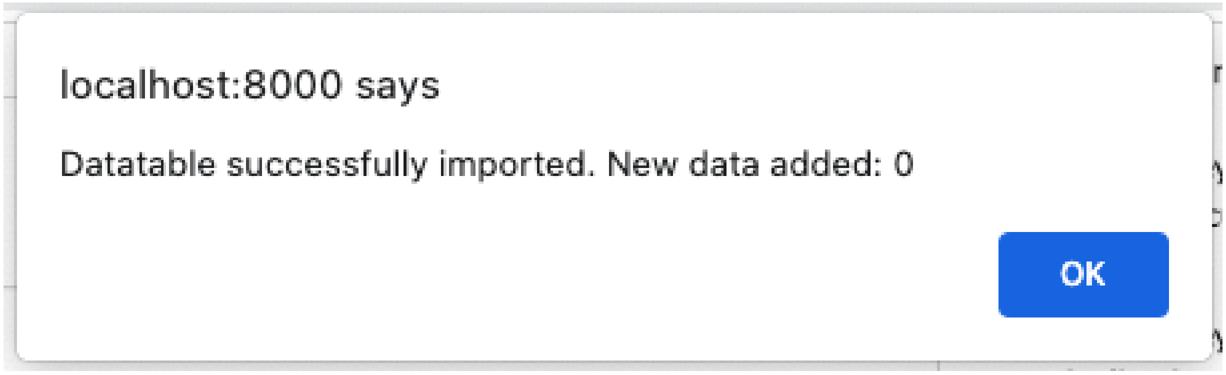
At the bottom of the Details page, there are buttons to delete the project, edit the project settings, or download the labeled data and (if applicable) trained model. These buttons are only visible to users with admin privileges for the project.



Note: If you have set up a database connection, there will also be buttons for ingesting new data from the ingest table, and exporting labeled data to the export table. Note that the export will **completely drop and rewrite the export table every time.**



When you click on the "Ingest new data from Database" button, SMART will import the ingest database table and compare it with the data already in SMART. If there are any new items in the database which do not already exist in SMART, these will be added. SMART will return the number of new items added in a window, or if an error was thrown, the error:



localhost:8000 says

Datatable successfully imported. New data added: 0

2.3.4 Updating a Project

The Update Project page is accessible from the *Details Page* of a project. This page can be used for the following operations:

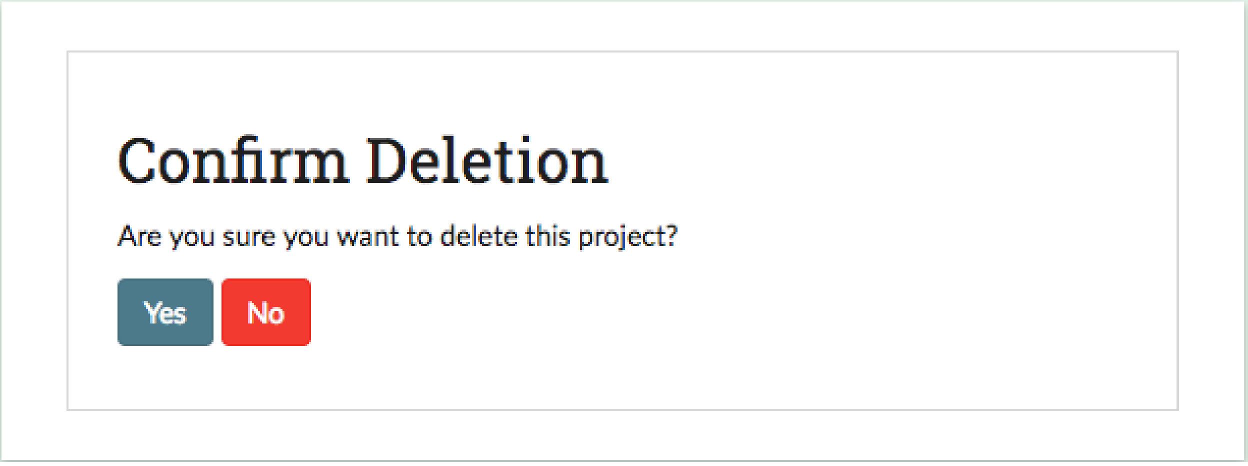
- Edit the project name and description.
- Add or remove an MSSQL database connection, or change settings (NOTE: you will be required to re-enter database credentials to make changes).
- Add additional data to label.
- Add or change the codebook file.
- Add, remove, or change project permissions.
- Edit label descriptions.
- Add the project to an existing group.

Tip:

- SMART allows up to two million records total per project. This includes additional data added later.
- New data is checked against existing project data for duplication.

2.3.5 Deleting a Project

The button to delete a project can be found on the *Details Page* page of a project. To delete a project, click this button and then select "Yes" at the prompt.



Confirm Deletion

Are you sure you want to delete this project?

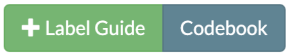
2.4 Part 4: Annotating Data

Once your project has been created, you are ready to start labeling! To begin, you can navigate to any project Annotation page from the *Projects Page* or to a specific Annotation page by pressing the "Annotate" link on the top navigation bar when on a project *Details Page* or *Admin Dashboard* page.

The Annotate page consists of either two or five tabs, depending on your User permissions. The sections below are marked by either ADMIN (available only to those with admin privileges) or ALL (available to everyone with at least coder privileges).

Note: If a user with admin privileges is on the annotation page, then other admin will be unable to access admin-only tabs until the first admin has left the page. This is to prevent multiple admin from labeling the same data simultaneously. However, if the first Admin has had at least **15 minutes** of inactivity, the second Admin will be given the page and the first Admin will be locked out. This is to prevent the page from remaining locked due to someone leaving a tab open. Coders and Admin can always access the Annotate Data and History pages. See *user-roles* for a chart of user permissions.

In addition, each tab has access to the project's *Label Guide (feature)* and *Codebook (feature)* using the buttons shown below:



2.4.1 Annotate Data Page

Overview

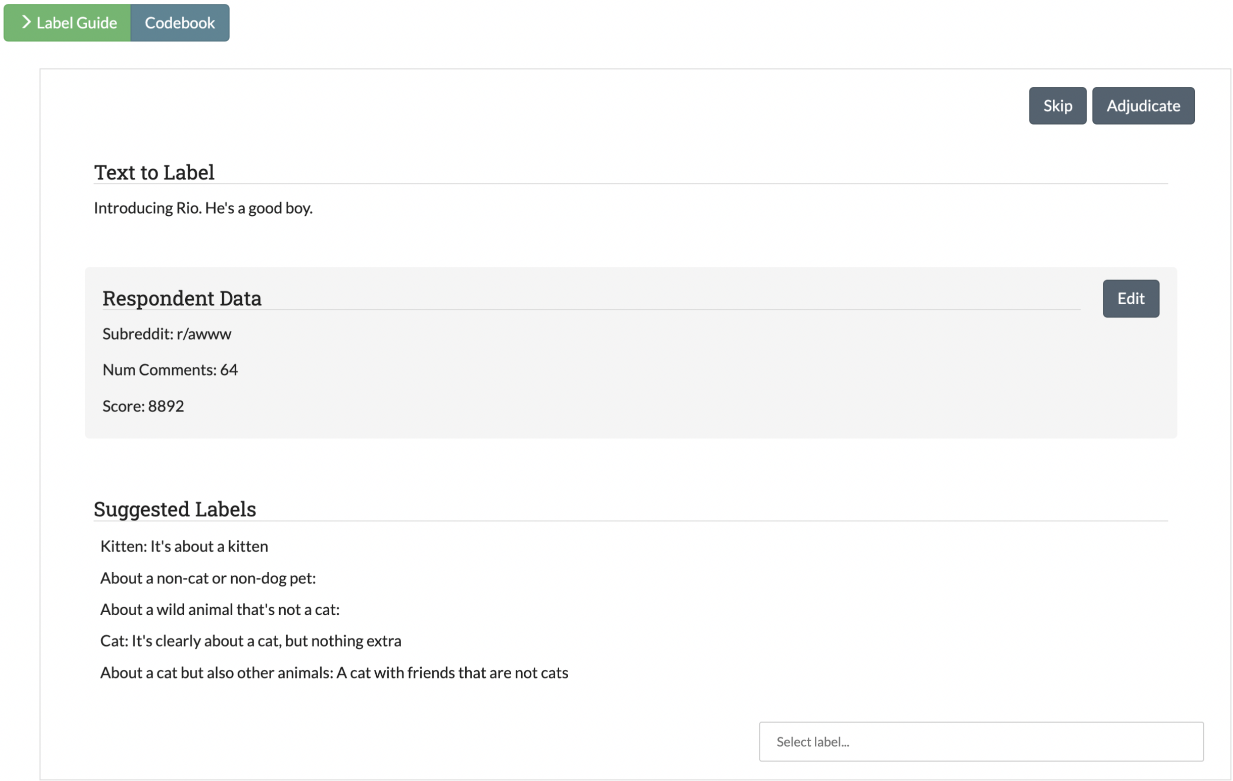
User: ALL

The Annotate Data tab is where most users will spend a majority of their time. When you enter this page, SMART will pass you a portion of the current batch as a deck of "cards", to be presented to you one at a time. You can then choose one of two actions:

- **Label:**
    - *For projects with at most five labels*: Assign a label to the observation by clicking on the button corresponding to the desired label.
    - *For projects with more than five labels*: Assign a label to the observation by either clicking on one of the suggested labels (see *Label Embeddings* for more information on label suggestions), or by searching for the correct label in the dropdown.
- **Skip**: This option is used when you want to skip an observation for now. This will send it back into the pool of data to label.
- **Adjudicate**: This option is used when you are unsure of what label to choose or you want to send an observation to the project administrators for review. When selecting this option, you will be required to provide a reason why you are sending the item to the administrators. Data that is not IRR is sent to the *Admin Annotation Page* to be reviewed by any user with admin privileges. Data that is IRR will still need to wait for the required number of coders to weigh in (adjudicating counts as your label).
- **[NEW] Edit Metadata**:Clicking the "Edit" button allows you to edit the values for the metadata fields for that card. This option is also available in the History and Skew tables. This allows users to record additional information about their data in metadata fields, or fix incorrect values.

If the data is not being used for *Inter-rater Reliability (IRR)* and is labeled, then this data will be marked as labeled and removed from the pool of unlabeled data. If data is IRR, then it may still be presented to additional coders on the project, but will not be presented to you again.

Below we have an annotate page for the "About Cats Labeling" project.



Text to Label

Introducing Rio. He's a good boy.

Respondent Data

Subreddit: r/awww

Num Comments: 64

Score: 8892

Suggested Labels

Kitten: It's about a kitten

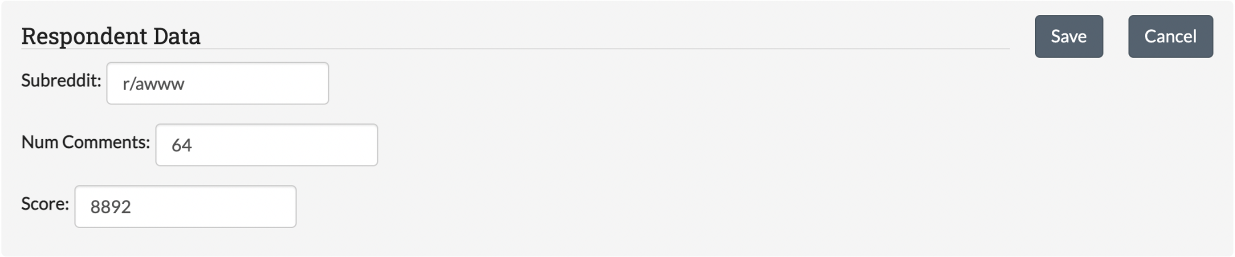
About a non-cat or non-dog pet:

About a wild animal that's not a cat:

Cat: It's clearly about a cat, but nothing extra

About a cat but also other animals: A cat with friends that are not cats

Below, we see an example of the card above after the "edit" button has been pressed



Respondent Data

Subreddit: r/awww

Num Comments: 64

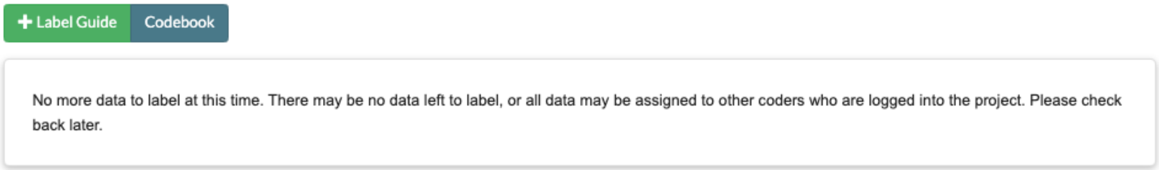
Score: 8892

Refilling the Batch

A user's card deck will continue to refill itself from the batch until it is empty. Once a batch has been coded or skipped, a new batch of unlabeled data will be requested from SMART. This batch will be selected using the chosen active learning algorithm, or randomly, depending on if Active Learning was enabled in *Advanced Settings*. The batch may also be selected randomly in projects with active learning enabled for three other reasons:

- It is the first batch.
- Each possible label has not been used at least once.
- There has not been a full batch worth of data marked as labeled (possibly some was skipped or is IRR and waiting for additional labels).

If a model is currently running, then the new batch will be delayed until the model has finished running, and you will be presented with the message in the image below. Note that this does not apply to projects that have disabled having a model. Projects that have disabled Active Learning but have a model will still have to wait for the model to run, but it will be done faster as predictions will not have to be generated for the unlabeled data (see *Part 5: Administrator Dashboard* for more details).



No more data to label at this time. There may be no data left to label, or all data may be assigned to other coders who are logged into the project. Please check back later.

Tip: If you are seeing the message above, try refreshing the page. The batch might have become available after the application was last queried. If the message is still there, then wait a few minutes for the model to finish and refresh again.

Note: You will also see the “no more data” message if all available data in the project is some combination of labeled, awaiting adjudication, IRR which you’ve already labeled, or assigned to someone else. See *Unassign Coders [NEW]* for how to free up data assigned to coders who do not plan to label it.

2.4.2 History Page

User: ALL

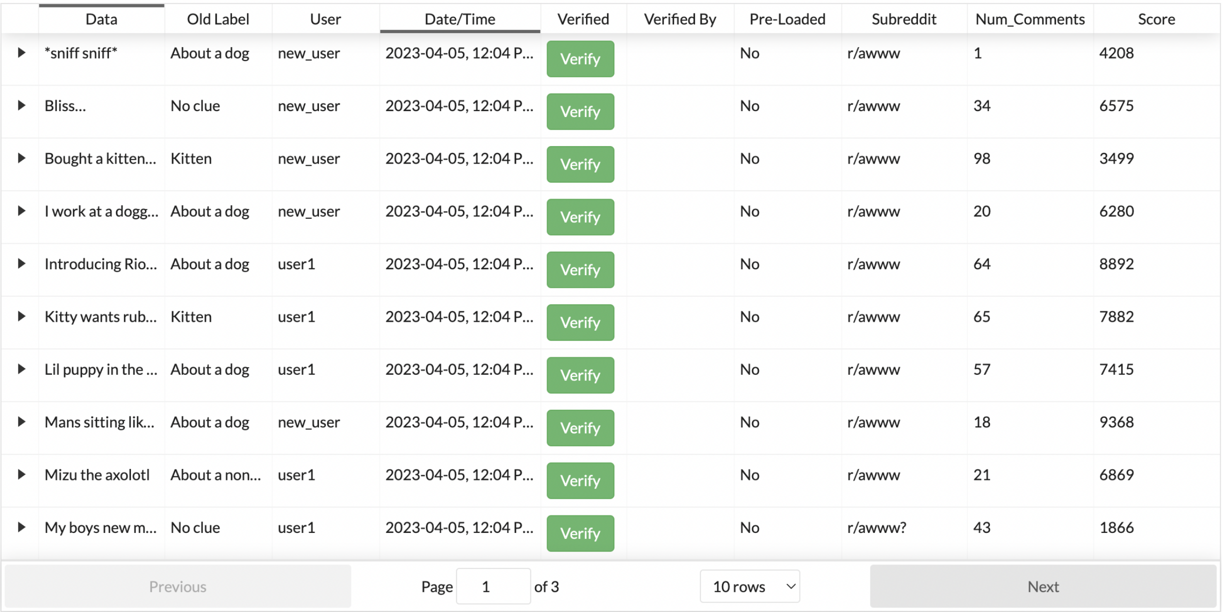
Overview

Perhaps you have been happily coding your data and you accidentally click the wrong label. Now you have data labeled “About a Cat” which is decidedly *not about cats!* Or perhaps you have labeled a number of items when your project leader announces that from this day forth, *Chihuahuas will also be counted as cats!* The history tab exists for scenarios like these ones. In this tab, you are able to view and edit your past labels.

This page includes all data that has been labeled by you personally, and provides the following fields:

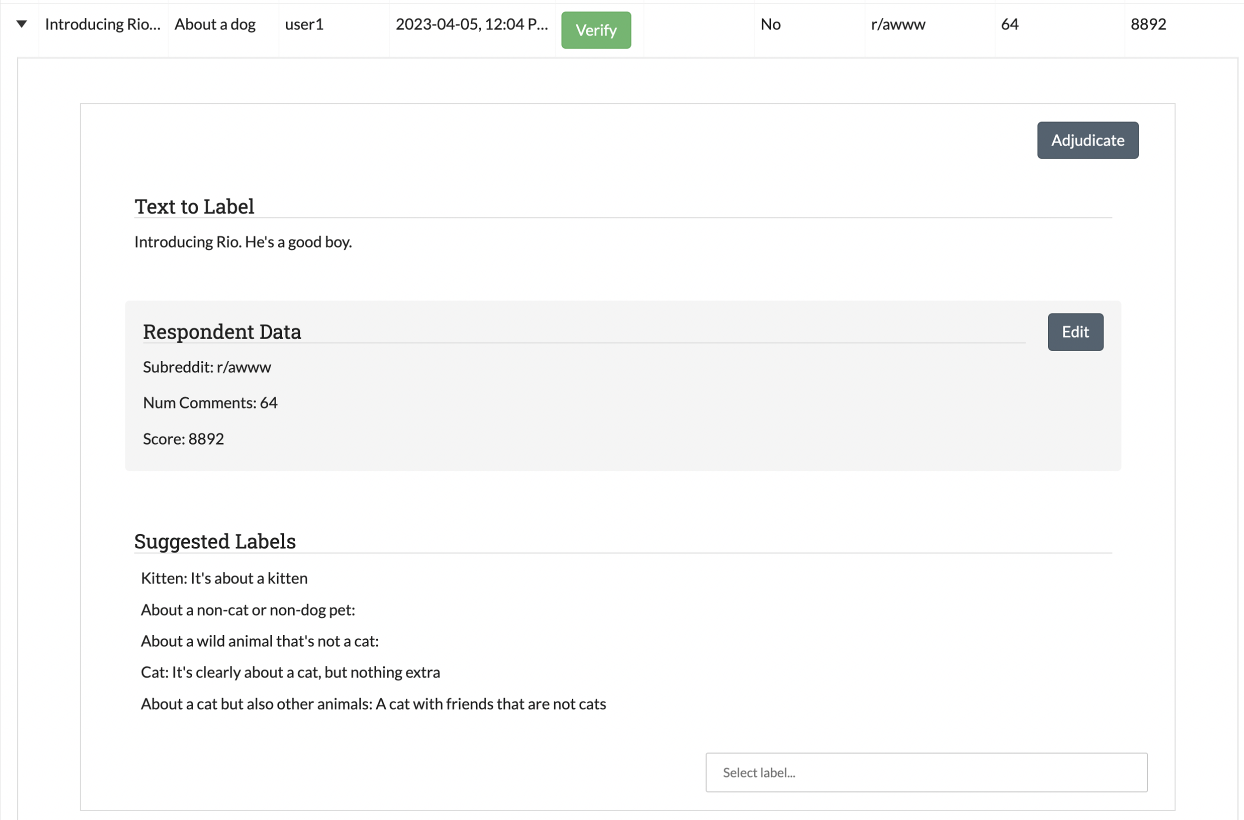
- **Data**: the text being labeled.
- **Old Label**: the current label assigned to the data.
- **User**: The username of the user who labeled the data (for pre-loaded labels this defaults to the project creator).
- **Date/Time**: The date and time where the data was labeled.
- **[NEW] Verified**: This field indicates if the label has been verified. If it has, this field will say “Yes.” If not, it will instead include a button to verify the data. Note that this feature is disabled for IRR data, as IRR data includes it’s own form of verification through either coder agreement or admin adjudication if they don’t agree.
- **[NEW] Verified By**: This is the username of the user who verified the data label.
- **[NEW] Pre-Loaded**: This field indicates if this labeled item was loaded into the system already labeled. Note that if you change the label in the history table, it will no longer be pre-loaded.
- **[NEW] Metadata Fields**: All metadata fields are also listed as columns, and so can be used for sorting or searching within a batch.

Note: Administrative users will be able to see and edit the labeled data for all coders. In the page below, we can see both `new_user`’s and `user1`’s labels.



	Data	Old Label	User	Date/Time	Verified	Verified By	Pre-Loaded	Subreddit	Num_Comments	Score
▶	*sniff sniff*	About a dog	new_user	2023-04-05, 12:04 P...	Verify		No	r/awww	1	4208
▶	Bliss...	No clue	new_user	2023-04-05, 12:04 P...	Verify		No	r/awww	34	6575
▶	Bought a kitten...	Kitten	new_user	2023-04-05, 12:04 P...	Verify		No	r/awww	98	3499
▶	I work at a dogg...	About a dog	new_user	2023-04-05, 12:04 P...	Verify		No	r/awww	20	6280
▶	Introducing Rio...	About a dog	user1	2023-04-05, 12:04 P...	Verify		No	r/awww	64	8892
▶	Kitty wants rub...	Kitten	user1	2023-04-05, 12:04 P...	Verify		No	r/awww	65	7882
▶	Lil puppy in the ...	About a dog	user1	2023-04-05, 12:04 P...	Verify		No	r/awww	57	7415
▶	Mans sitting lik...	About a dog	new_user	2023-04-05, 12:04 P...	Verify		No	r/awww	18	9368
▶	Mizu the axolotl	About a non...	user1	2023-04-05, 12:04 P...	Verify		No	r/awww	21	6869
▶	My boys new m...	No clue	user1	2023-04-05, 12:04 P...	Verify		No	r/awww?	43	1866

To save space, the history table only includes enough text for each data sample to fit the page width. To expand a row for reading and editing, click on the arrow to the left of the text. This will open up a subrow with the entire text and the label/skip options. Note that changing a label to Adjudicate will remove it from the history table as you have effectively given up responsibility for it.



▼	Introducing Rio...	About a dog	user1	2023-04-05, 12:04 P...	Verify	No	r/awww	64	8892

Text to Label

Introducing Rio. He's a good boy.

Respondent Data

Subreddit: r/awww

Num Comments: 64

Score: 8892

Suggested Labels

Kitten: It's about a kitten

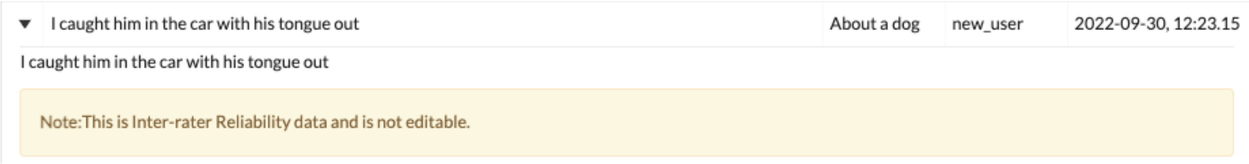
About a non-cat or non-dog pet:

About a wild animal that's not a cat:

Cat: It's clearly about a cat, but nothing extra

About a cat but also other animals: A cat with friends that are not cats

Note: *Inter-rater Reliability (IRR)* data labels can be changed in the history table up until the point where enough people have labeled/skipped it and it is processed. At this point, the data is effectively "labeled by everyone" (either from consensus or from an admin resolving a dispute) and will no longer be editable on anyone's history table. Expanding a resolved IRR datum will simply show a message (see below):



▼ I caught him in the car with his tongue out	About a dog	new_user	2022-09-30, 12:23.15

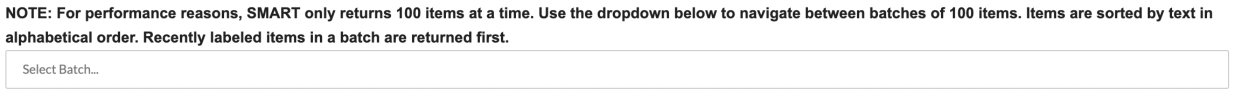
I caught him in the car with his tongue out

Note:This is Inter-rater Reliability data and is not editable.

> **Warning:** *For Active Learning Users:* Active learning algorithms use past labeled data to select future batches. Data labels changed retroactively will appear in the training data for the next batch, but will not affect past batches or the current batch. Excessive label changing may hamper active learning algorithms and make them less effective (see *Active Learning* for more details)

[NEW] Searching, Sorting, and Filtering

Batching: To keep the performance of the history table optimal, SMART sorts the data by alphabetical order and then batches the results into groups of 100 items.

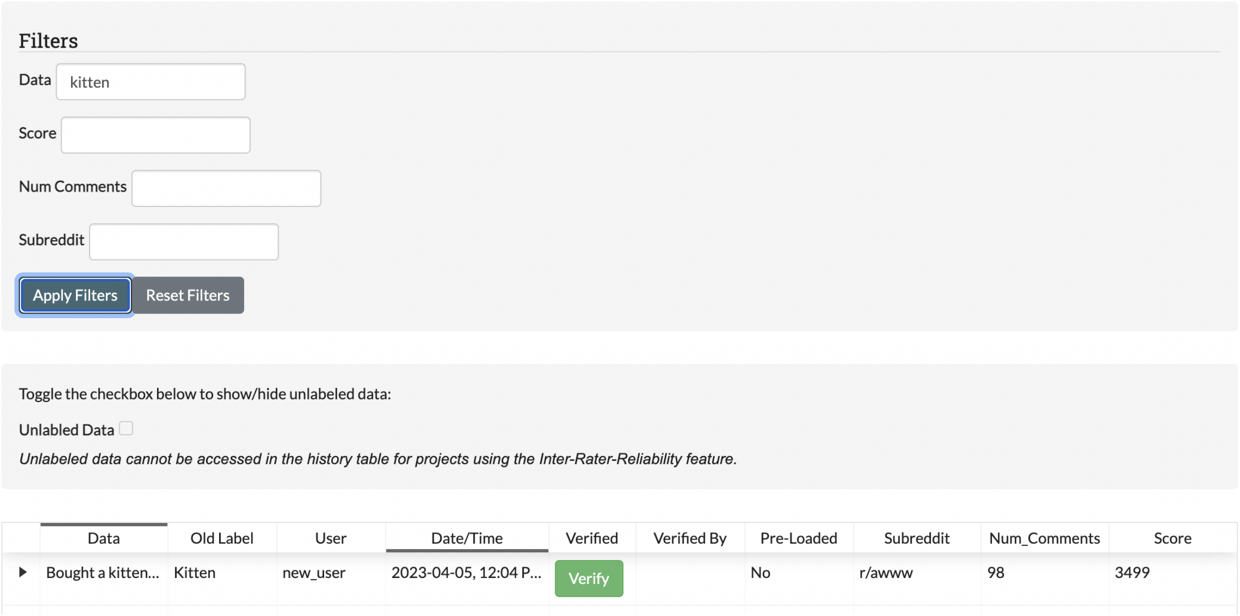


NOTE: For performance reasons, SMART only returns 100 items at a time. Use the dropdown below to navigate between batches of 100 items. Items are sorted by text in alphabetical order. Recently labeled items in a batch are returned first.

Select Batch...

Each batch in the history table is automatically sorted by the date to provide the most recent labels first, and users can sort and filter within the batch inside the table (see *Searching and Sorting Tables*). For items that either don't have a label date or have the same date, they are returned in alphabetical order by text.

Filtering: By default, the history table contains all labeled items. The filter form at the top allows users to filter results to specific text or metadata values. The "Reset Filters" button resets the form and returns the History table back to its original state.



Filters

Data: kitten

Score:

Num Comments:

Subreddit:

Toggle the checkbox below to show/hide unlabeled data:

Unlabled Data □

Unlabeled data cannot be accessed in the history table for projects using the Inter-Rater-Reliability feature.

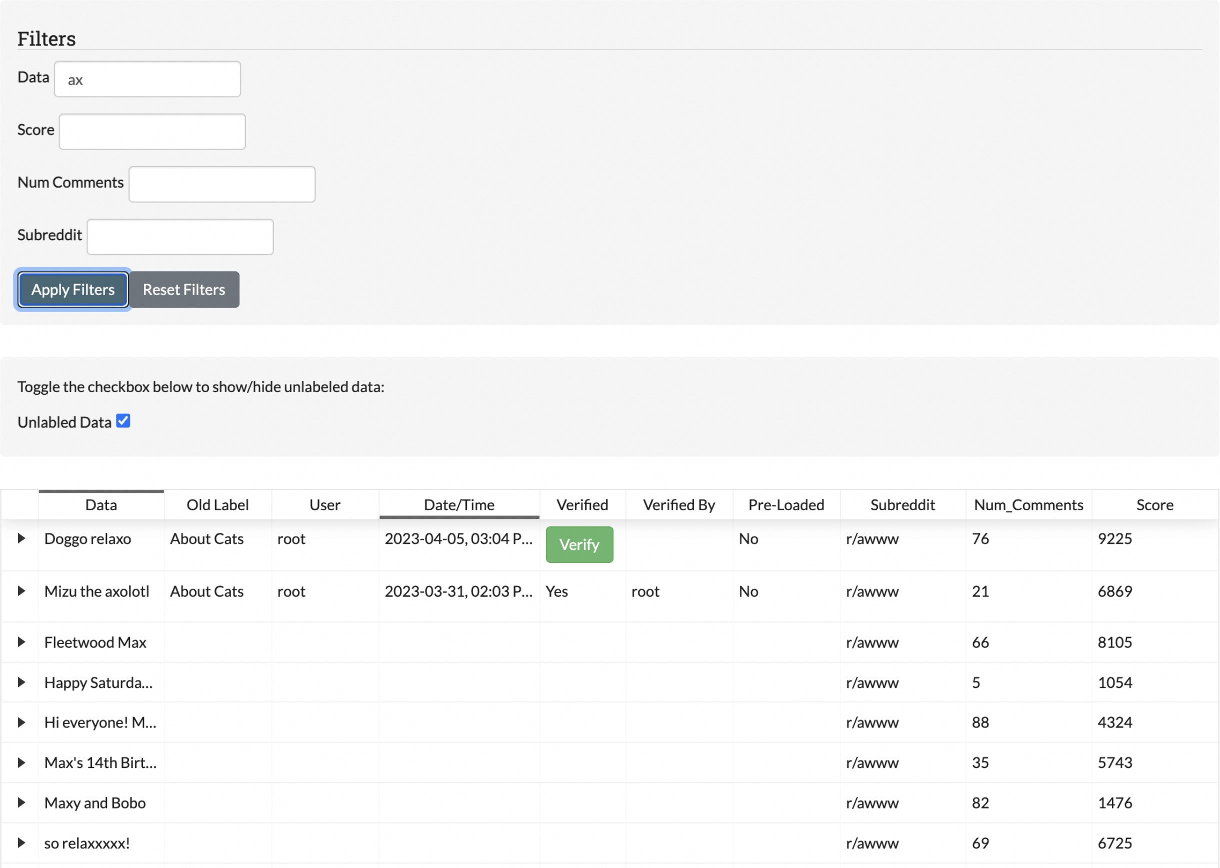
	Data	Old Label	User	Date/Time	Verified	Verified By	Pre-Loaded	Subreddit	Num_Comments	Score
▶	Bought a kitten...	Kitten	new_user	2023-04-05, 12:04 P...	Verify		No	r/awww	98	3499

Note: Filters are not case-sensitive, and return all examples where the filtered text is contained in the field of interest. This is also the case with numeric fields, so for example if you filter Num_Comments to "9," items with values 9, 89, 901, or 1239 would all be returned.

[NEW] Toggling Unlabeled Data (Non-IRR Projects Only)

By design, the History table primarily exists to allow users to view and change their past labels. But what do you do if you are trying to label new items, but require the context of how you labeled similar items in the past?

For these cases, SMART now allows users to toggle the History table to include data which is unlabeled and unassigned by checking the "Unlabeled Data" checkbox below the filter box. This data shows up with empty values for all label-related fields like "Old Label." They can then filter or sort the table to the data they want, and code items from there using the same workflow someone would use to change a previously assigned label.



Filters

Data: ax

Score:

Num Comments:

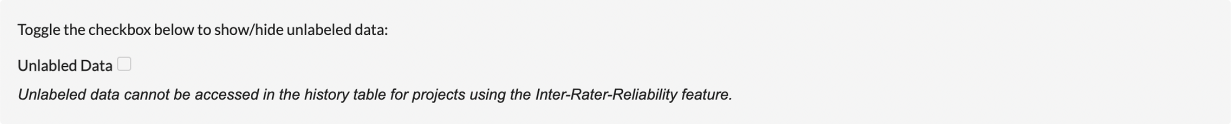
Subreddit:

Toggle the checkbox below to show/hide unlabeled data:

Unlabled Data ☑

	Data	Old Label	User	Date/Time	Verified	Verified By	Pre-Loaded	Subreddit	Num_Comments	Score
▸	Doggo relaxo	About Cats	root	2023-04-05, 03:04 P...	Verify		No	r/awww	76	9225
▸	Mizu the axolotl	About Cats	root	2023-03-31, 02:03 P...	Yes	root	No	r/awww	21	6869
▸	Fleetwood Max							r/awww	66	8105
▸	Happy Saturda...							r/awww	5	1054
▸	Hi everyone! M...							r/awww	88	4324
▸	Max's 14th Birt...							r/awww	35	5743
▸	Maxy and Bobo							r/awww	82	1476
▸	so relaxxxxx!							r/awww	69	6725

Because this feature essentially goes around the logic used to hand out IRR data to coders, it is disabled for projects where the percent IRR is greater than 0%. Instead, users will see the following message:



Toggle the checkbox below to show/hide unlabeled data:

Unlabled Data □

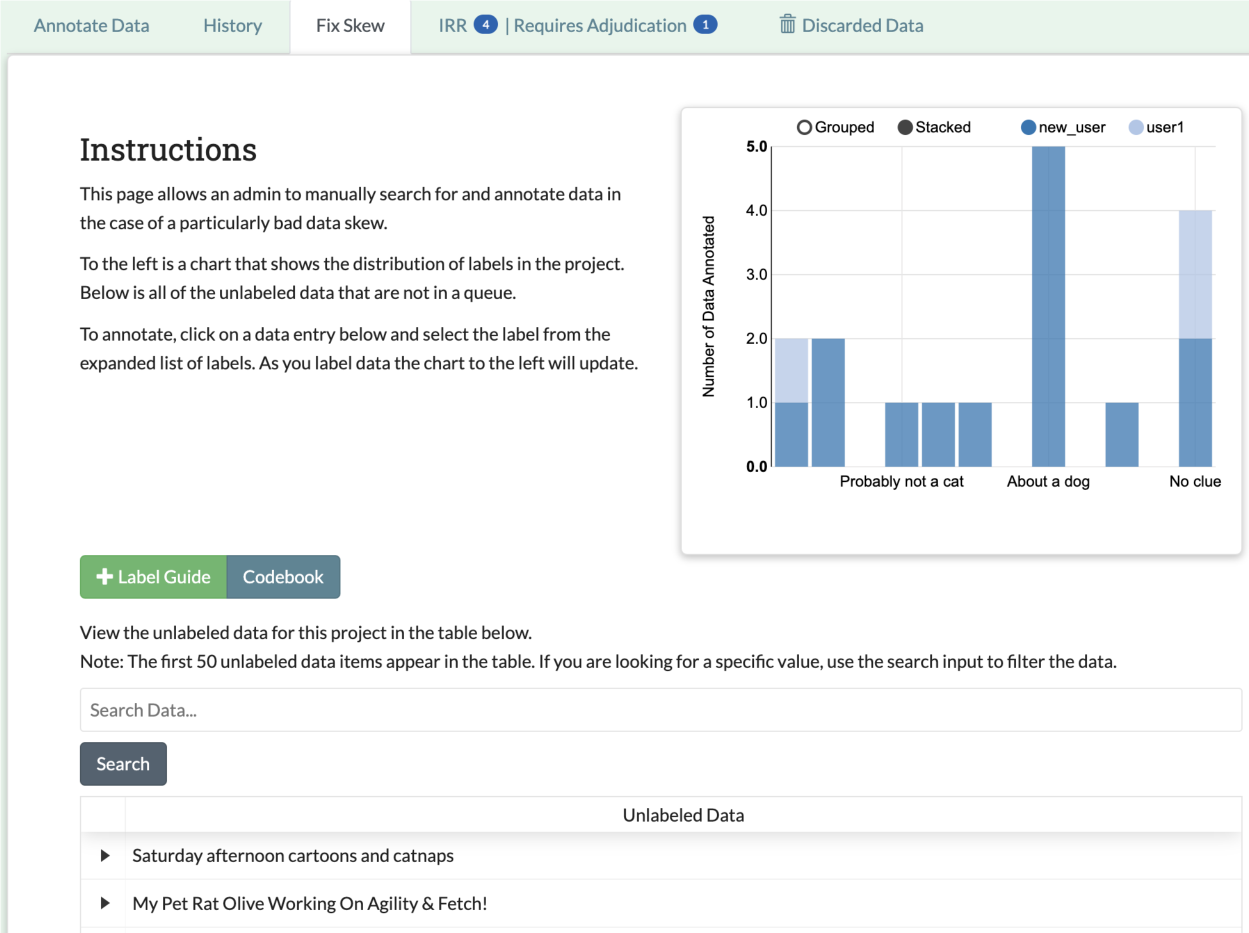
Unlabeled data cannot be accessed in the history table for projects using the Inter-Rater-Reliability feature.

> **Warning:** *For Active Learning Users:* While we don't explicitly prohibit projects with Active Learning from using this feature, it's important to note that the History table always presents all unassigned and unlabeled data in alphabetical order, and is not impacted by the ordering suggestions from Active Learning models. Users will need to annotate using the "Annotate Data" tab to benefit from Active Learning.

2.4.3 Fix Skew Page

User: ADMIN

In our "About Cats Labeling" project, the label set includes the labels "Cat" and "Kitten", but also "Wild cat" (since all cats are valid and we want to identify these specific ones). The only problem is that wild cat posts are fairly rare in your data, and nobody has seen one yet! You know your classifier won't even run until a wild cat post has been found (see *Refilling the Batch*), but you are worried that waiting for random selection to find a wild cat post might take a while. The "Fix Skew" page exists for this scenario. In this tab, users with admin privileges may search unlabeled data directly for examples of rare labels. The graph on the right side of the page shows the current counts for each label (see image below).



Instructions

This page allows an admin to manually search for and annotate data in the case of a particularly bad data skew.

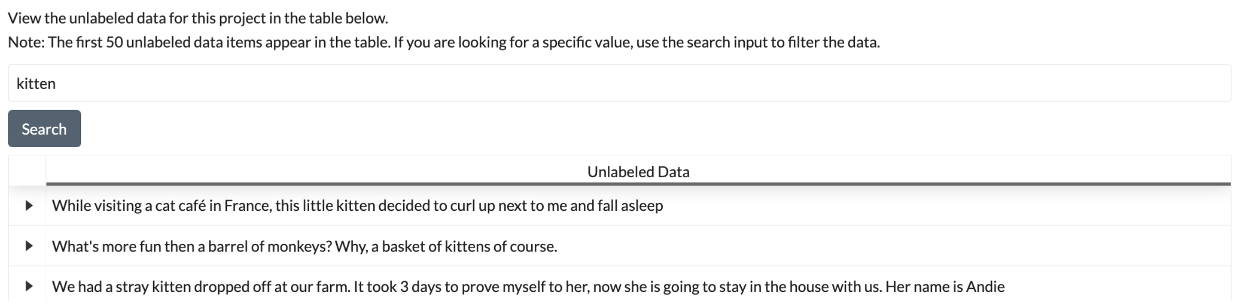
To the left is a chart that shows the distribution of labels in the project. Below is all of the unlabeled data that are not in a queue.

To annotate, click on a data entry below and select the label from the expanded list of labels. As you label data the chart to the left will update.

View the unlabeled data for this project in the table below.
Note: The first 50 unlabeled data items appear in the table. If you are looking for a specific value, use the search input to filter the data.

	Unlabeled Data
▸	Saturday afternoon cartoons and catnaps
▸	My Pet Rat Olive Working On Agility & Fetch!

The *Fix Skew Page* table has a separate text button and search bar above the table, as the skew page cannot load all of the unlabeled data at once, and will instead just load the top 50 data items that contain the searched text.



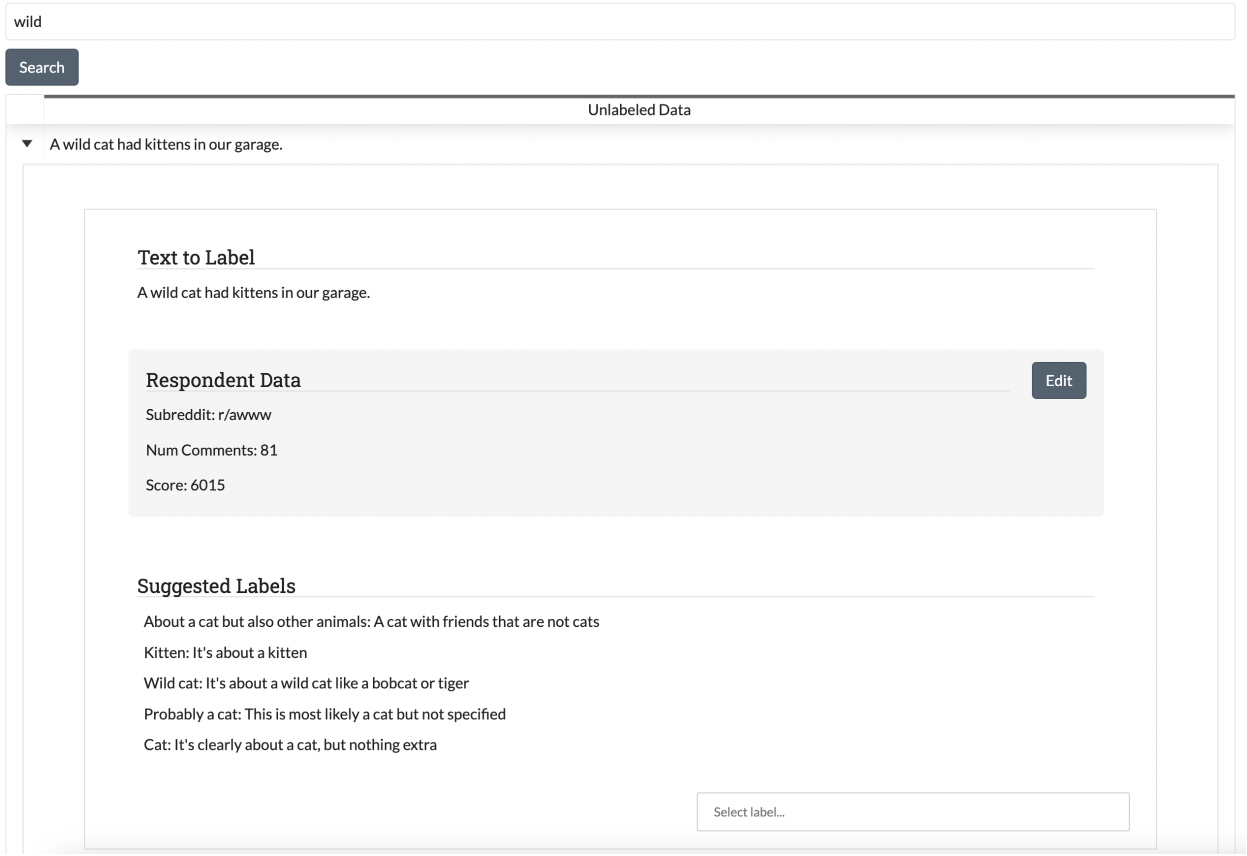
View the unlabeled data for this project in the table below.
Note: The first 50 unlabeled data items appear in the table. If you are looking for a specific value, use the search input to filter the data.

kitten

	Unlabeled Data
▸	While visiting a cat café in France, this little kitten decided to curl up next to me and fall asleep
▸	What's more fun then a barrel of monkeys? Why, a basket of kittens of course.
▸	We had a stray kitten dropped off at our farm. It took 3 days to prove myself to her, now she is going to stay in the house with us. Her name is Andie

To fix a skew, follow these steps:

1. Use the search bar above the table to search the data for keywords. The first 50 text items by closest match will be returned.
2. Click on the arrow to the left of the row to expand
3. Assign a label to the data



wild

	Unlabeled Data
▾	A wild cat had kittens in our garage.

Text to Label

A wild cat had kittens in our garage.

Respondent Data

Subreddit: r/awww

Num Comments: 81

Score: 6015

Suggested Labels

About a cat but also other animals: A cat with friends that are not cats

Kitten: It's about a kitten

Wild cat: It's about a wild cat like a bobcat or tiger

Probably a cat: This is most likely a cat but not specified

Cat: It's clearly about a cat, but nothing extra

Once data has been labeled, the graph at the top will show the change in label counts.

> **Warning:** The Fix Skew page is very similar to the History page's Unlabeled Data feature, in that it gives users the ability to code whatever they want in any order. This allows coders to both ignore any Active Learning model present, and any IRR requirements (data coded on this page will be assigned a final label without being shown to anyone else). As such, please use with caution if you are using either feature!

2.4.4 Admin Annotation Page

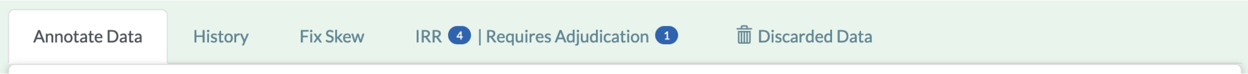
User: ADMIN

The Admin Annotation page lets users with admin User privileges resolve ambiguous data. There are two types of ambiguous data that could end up in this table.

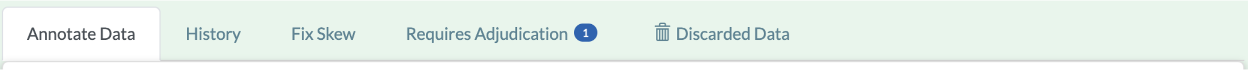
1. Normal (not *Inter-rater Reliability (IRR)*) data that was sent for Adjudication
2. *Inter-rater Reliability (IRR)* data that has been annotated/sent for adjudication by enough people, where there was either a disagreement between the assigned labels, or at least one coder sent it to adjudication (this counts as a disagreement).

Tip: Coders are not given any indication of which data is being used for IRR. If you are using IRR in your project, and cannot find a specific datum you sent for Adjudication in the admin table, it may be IRR data that has not been seen by enough people yet.

The Admin Annotation tab is marked with badges showing the total number of unaddressed items. For a project that uses IRR, it will look like the tab in the image below with two sections:

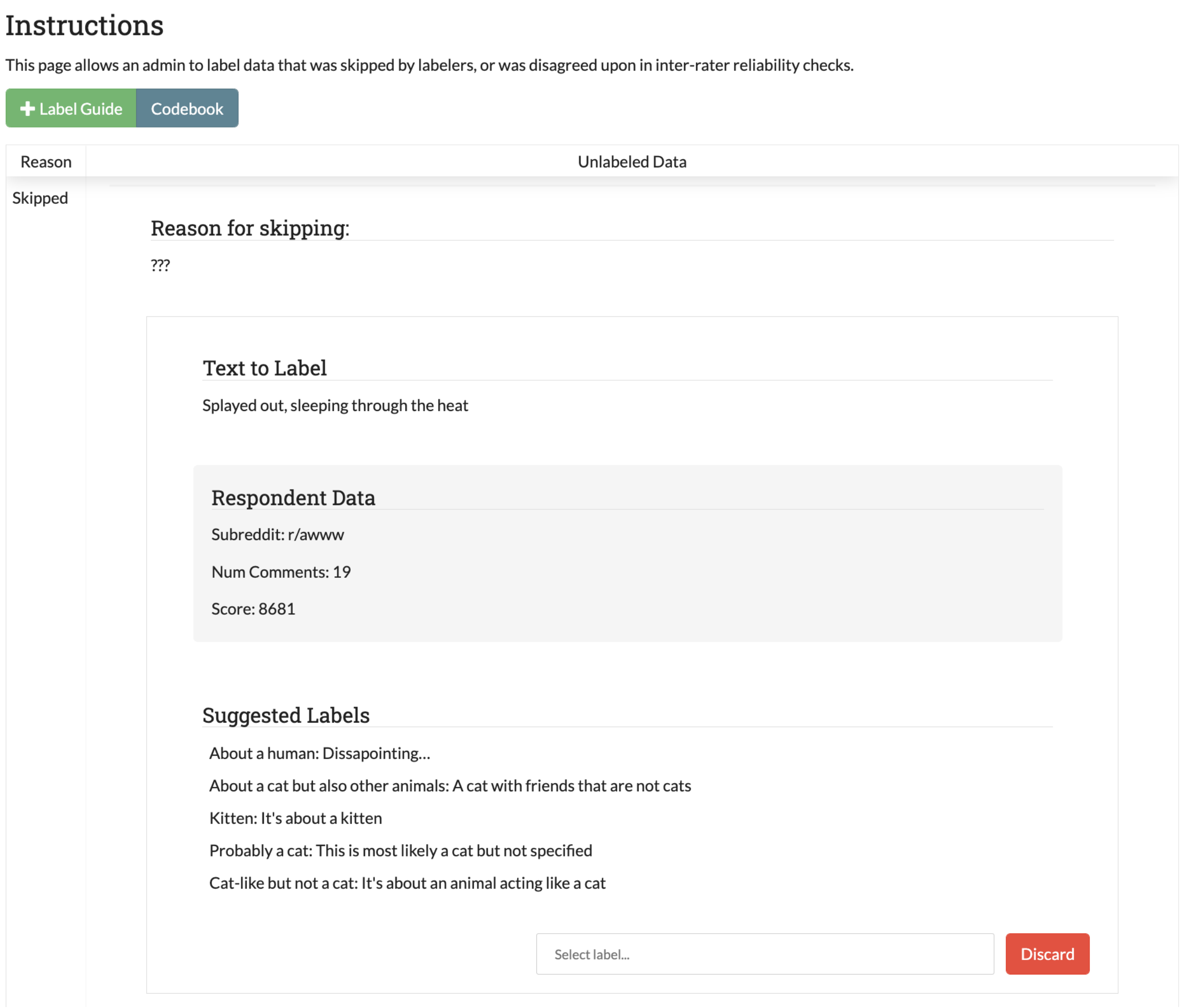


Projects that do not utilize IRR will only show the Requires Adjudication count:



The Admin Annotation page consists of a table with two columns. The first shows the reason data ended up in the table (IRR or Sent for Adjudication). The second gives the text for the data, the reason the coder gave for sending the data to Adjudication (if not IRR), and provides options for how the data should be processed. The admin has two options for any data in this table:

- **Label:**
  - By clicking on one of the label buttons, suggestions, or dropdowns, the data is assigned the selected label and becomes part of the training set. If this data was sent for adjudication, then it will also become available in the admin's *History Page* if they want to change it later. If the data is IRR, it will also appear in their history table, but will **NOT** be editable by any user.
- **Discard:**
  - This option exists for data that is simply un-codable and should not be included in the project. Clicking this option will remove the data from any IRR records, the *Fix Skew Page*, and any consideration for future batches. Note that the data can be restored on the *Recycle Bin Page*.



Instructions

This page allows an admin to label data that was skipped by labelers, or was disagreed upon in inter-rater reliability checks.

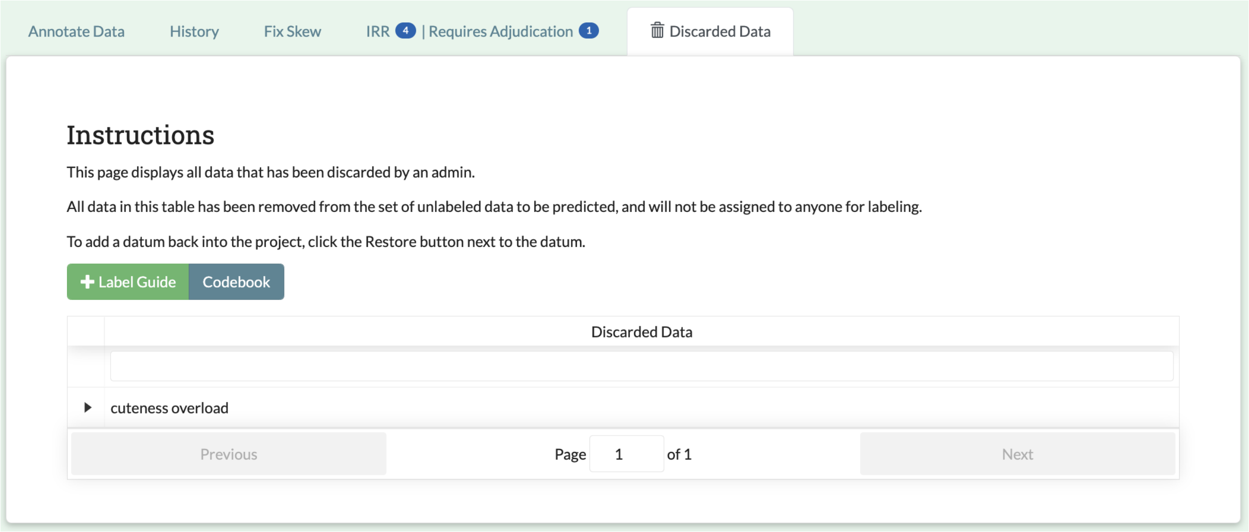
Reason	Unlabeled Data
Skipped	Reason for skipping: ??? Text to Label Splayed out, sleeping through the heat Respondent Data Subreddit: r/awww Num Comments: 19 Score: 8681 Suggested Labels About a human: Dissapointing... About a cat but also other animals: A cat with friends that are not cats Kitten: It's about a kitten Probably a cat: This is most likely a cat but not specified Cat-like but not a cat: It's about an animal acting like a cat

2.4.5 Recycle Bin Page

User: ADMIN

The Recycle Bin page acts much like a recycle bin or trash folder for most computers. Any data that was discarded in the *Admin Annotation Page* will appear on this page:

Tip: You can search the Recycle Bin table for specific data (see *Searching and Sorting Tables*).



Instructions

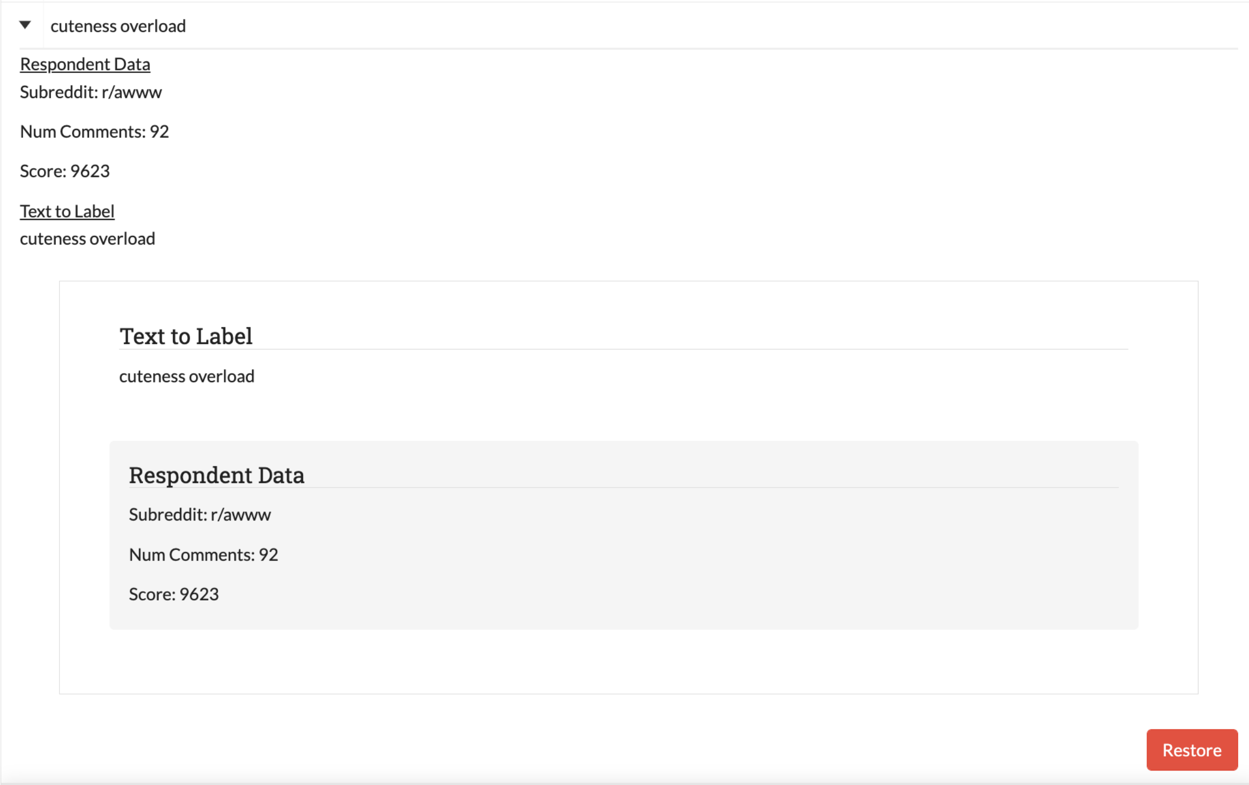
This page displays all data that has been discarded by an admin.

All data in this table has been removed from the set of unlabeled data to be predicted, and will not be assigned to anyone for labeling.

To add a datum back into the project, click the Restore button next to the datum.

	Discarded Data
▸	cuteness overload

Data in the table will only be shown up to the width of the page to maximize the number of rows shown on the screen. To expand data, click the arrow on the left of the row. This will open a subrow with the entire text and a "Restore" button. Clicking on this button will remove the data from the Recycle Bin and place it back in the pool of unlabeled data for consideration.



▾ cuteness overload

Respondent Data
Subreddit: r/awww

Num Comments: 92

Score: 9623

Text to Label
cuteness overload

Text to Label

cuteness overload

Respondent Data

Subreddit: r/awww

Num Comments: 92

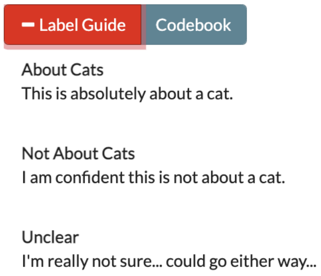
Score: 9623

Note: Restoring data will *not* restore any past records for this data. If data was marked for *Inter-rater Reliability (IRR)*, was discarded from the admin table, and then restored, any past labels or skips will not be restored with it and the data will not be marked for IRR unless it is chosen again later.

2.4.6 Label Guide (feature)

User: ALL

The label guide contains the list of possible labels and their descriptions as set by the project creator or updater. This guide is placed on every tab of the *Annotate Data Page* page for the user's convenience. To open the tab, click on the green `+ Label Guide` button (see *Annotate Data Page*). The button will turn red with a minus sign as long as the guide is open (as shown below). To close, click the button again.



About Cats
This is absolutely about a cat.

Not About Cats
I am confident this is not about a cat.

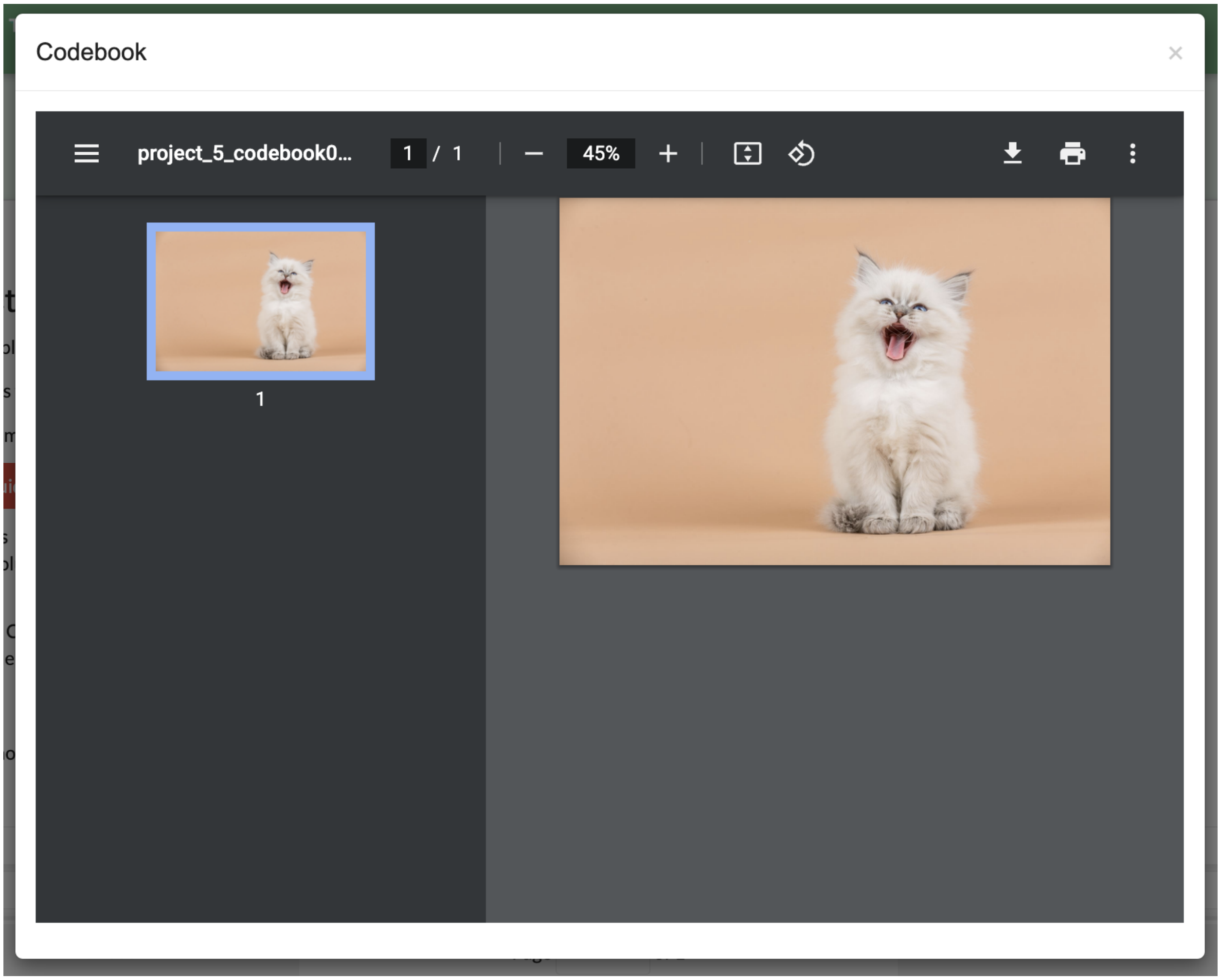
Unclear
I'm really not sure... could go either way...

2.4.7 Codebook (feature)

User: ALL

When creating or updating a project, a creator or admin has the option to add a codebook (see *Adding Codebook*). If a codebook has been uploaded, then in addition to the *Label Guide (feature)*, a codebook button will be available on each tab of the *Annotate Data Page* page. To open, click the `codebook` button. This will open a PDF viewer on the application with the file. To close, either click the `x` in the top right corner of the popup, or click anywhere on the screen outside of the codebook.

Below is our codebook for the "About Cats" projects.



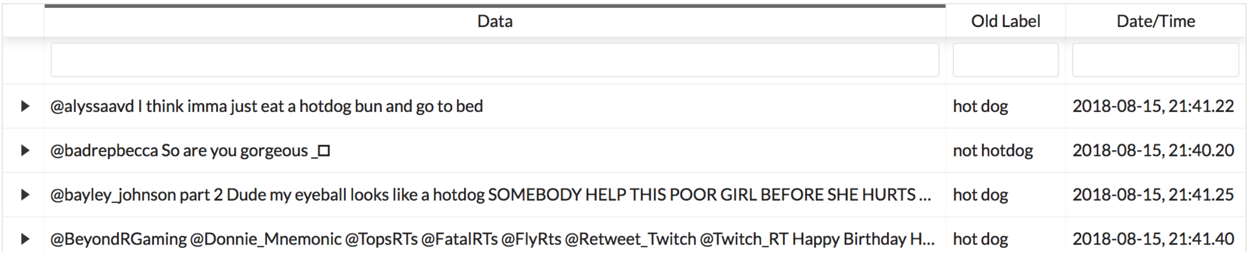
Warning: This feature makes use of the browser's built in PDF viewer. For most modern browsers like Firefox, Chrome, or Safari, this viewer will include a print or download button. However, if you are using an outdated browser, this might not be available.

2.4.8 Searching and Sorting Tables

User: ALL

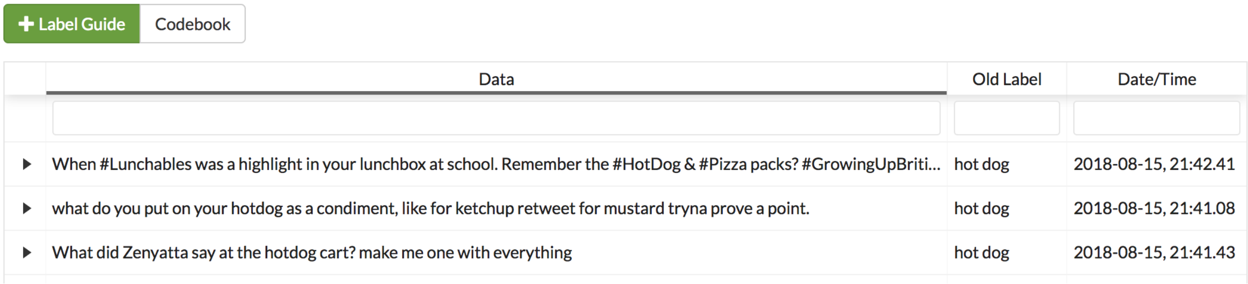
You can sort any table on an annotation page by a desired column by clicking on the column header.

One click will sort it in ascending order (indicated by a grey bar at the top of the column name).



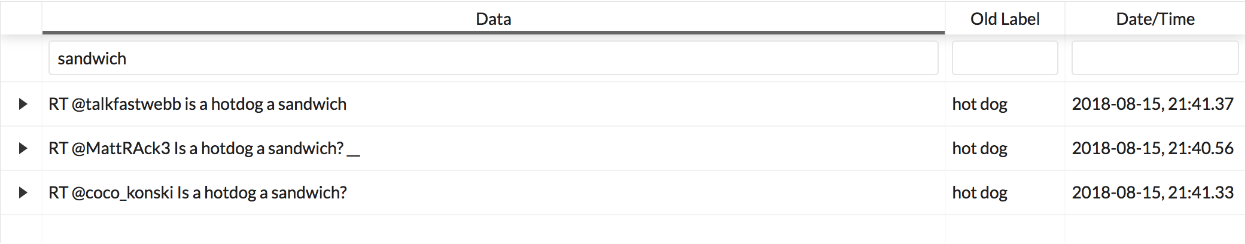
	Data	Old Label	Date/Time
▸	@alyssaavd I think imma just eat a hotdog bun and go to bed	hot dog	2018-08-15, 21:41.22
▸	@badrepbecca So are you gorgeous _□	not hotdog	2018-08-15, 21:40.20
▸	@bayley_johnson part 2 Dude my eyeball looks like a hotdog SOMEBODY HELP THIS POOR GIRL BEFORE SHE HURTS ...	hot dog	2018-08-15, 21:41.25
▸	@BeyondRGaming @Donnie_Mnemonic @TopsRTs @FatalRTs @FlyRts @Retweet_Twitch @Twitch_RT Happy Birthday H...	hot dog	2018-08-15, 21:41.40

A second click will sort it in descending order (indicated by the grey bar below the text).



	Data	Old Label	Date/Time
▶	When #Lunchables was a highlight in your lunchbox at school. Remember the #HotDog & #Pizza packs? #GrowingUpBriti...	hot dog	2018-08-15, 21:42.41
▶	what do you put on your hotdog as a condiment, like for ketchup retweet for mustard tryna prove a point.	hot dog	2018-08-15, 21:41.08
▶	What did Zenyatta say at the hotdog cart? make me one with everything	hot dog	2018-08-15, 21:41.43

The tables on the *History Page* and *Recycle Bin Page* can be filtered using the text boxes under each column header. When text is entered in one of these boxes, only the rows containing the entered text will be displayed.

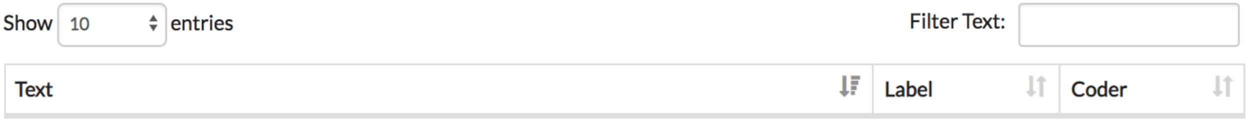


	Data	Old Label	Date/Time
	sandwich		
▶	RT @talkfastwebb is a hotdog a sandwich	hot dog	2018-08-15, 21:41.37
▶	RT @MattRAck3 Is a hotdog a sandwich? _	hot dog	2018-08-15, 21:40.56
▶	RT @coco_konski Is a hotdog a sandwich?	hot dog	2018-08-15, 21:41.33

2.5 Part 5: Administrator Dashboard

If you are a project admin, you may want some way to keep track of how your project is doing. The administrator dashboard allows users with admin privileges to track their project's progress. Depending on the project settings, this page will have one, two, or three tabs to let them track different aspects of the project. Each of the sections below specifies what projects they apply to.

Each of the Administrator Dashboard tabs includes a table. These tables can be filtered using the text box at located at the top right. They can also be sorted by column by clicking on the column header (exactly as with the *annotation page tables* but instead of a grey bar there is an arrow and stair icon).



Text	Label	Coder

2.5.1 Labeled Data Page

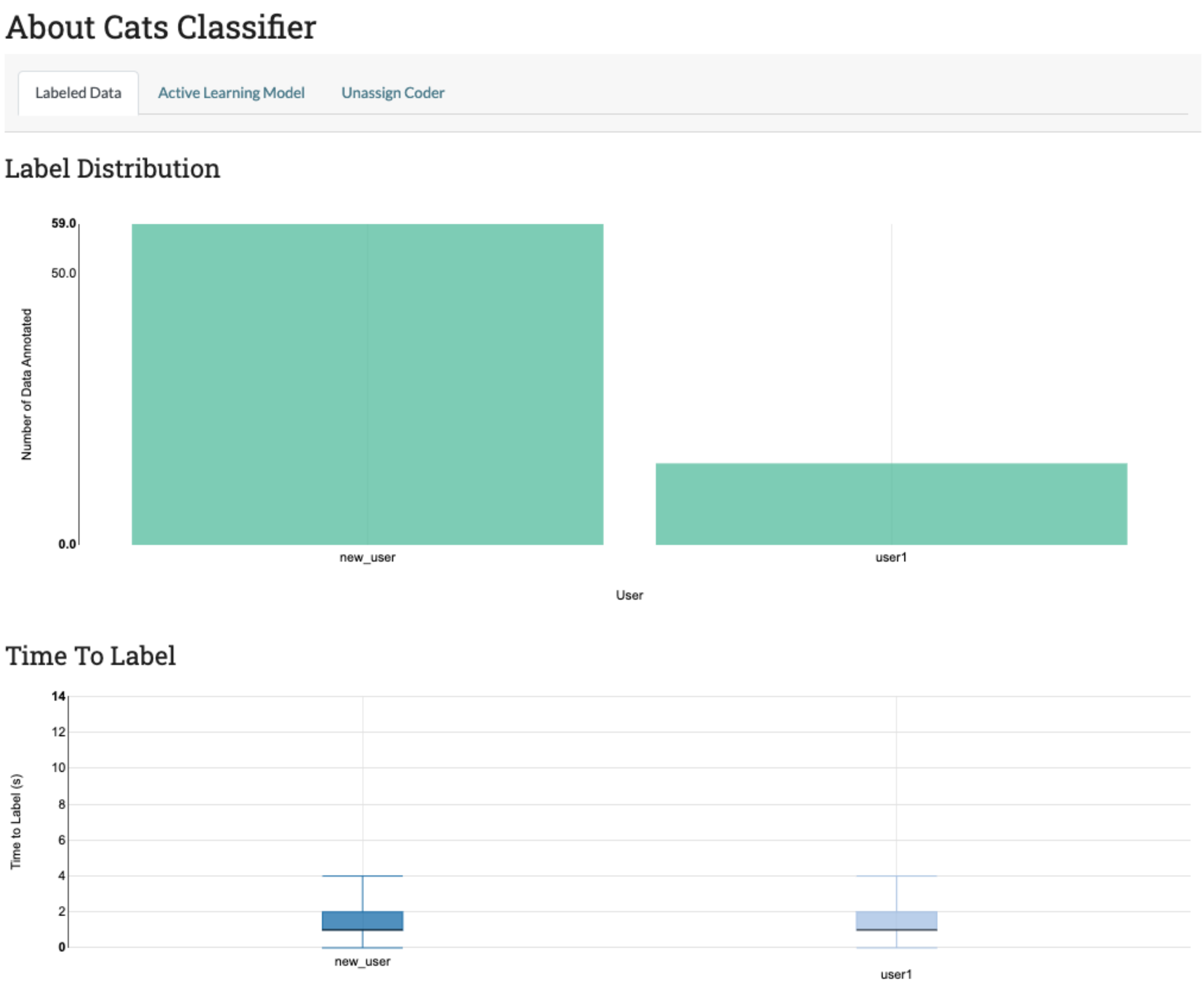
Visible for all projects

The labeled data page is designed to provide a summary of how the coders are doing comparatively in terms of speed and label distribution. There are three main features of the page:

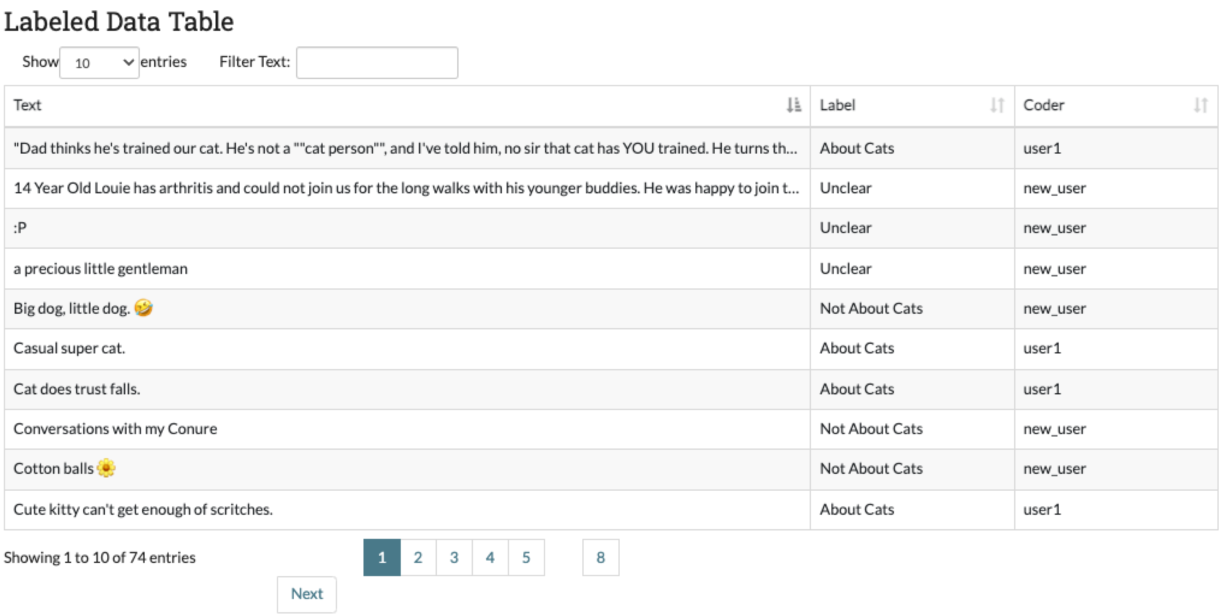
- The bar chart on the top permits project admin to compare at a glance how many items coders have labeled. This lets admin see which users are labeling more.
- The box and whisker chart on the bottom lets project admin see how long each coder is taking to annotate the data. This helps admin detect coders who may be having trouble or coders who are simply clicking through their data.

Note: As of SMART 3.0.0, the bar chart no longer stratifies by label, as this quickly becomes impractical for many coders or labels.

Below is the admin page for the “About Cats Classifier” project:



- The Labeled Data table at the bottom of the page contains all data that has been officially assigned a label. The table includes a snippet of the text, the assigned label, and the user responsible for the label. This lets admin see how much data has been collected in total so far, and get a sense for their labeled data without *Part 6: Downloading Labeled Data and/or Model*.



Labeled Data Table

Text	Label	Coder
"Dad thinks he's trained our cat. He's not a ""cat person"", and I've told him, no sir that cat has YOU trained. He turns th...	About Cats	user1
14 Year Old Louie has arthritis and could not join us for the long walks with his younger buddies. He was happy to join t...	Unclear	new_user
:P	Unclear	new_user
a precious little gentleman	Unclear	new_user
Big dog, little dog. 🤣	Not About Cats	new_user
Casual super cat.	About Cats	user1
Cat does trust falls.	About Cats	user1
Conversations with my Conure	Not About Cats	new_user
Cotton balls 🌼	Not About Cats	new_user
Cute kitty can't get enough of scritches.	About Cats	user1

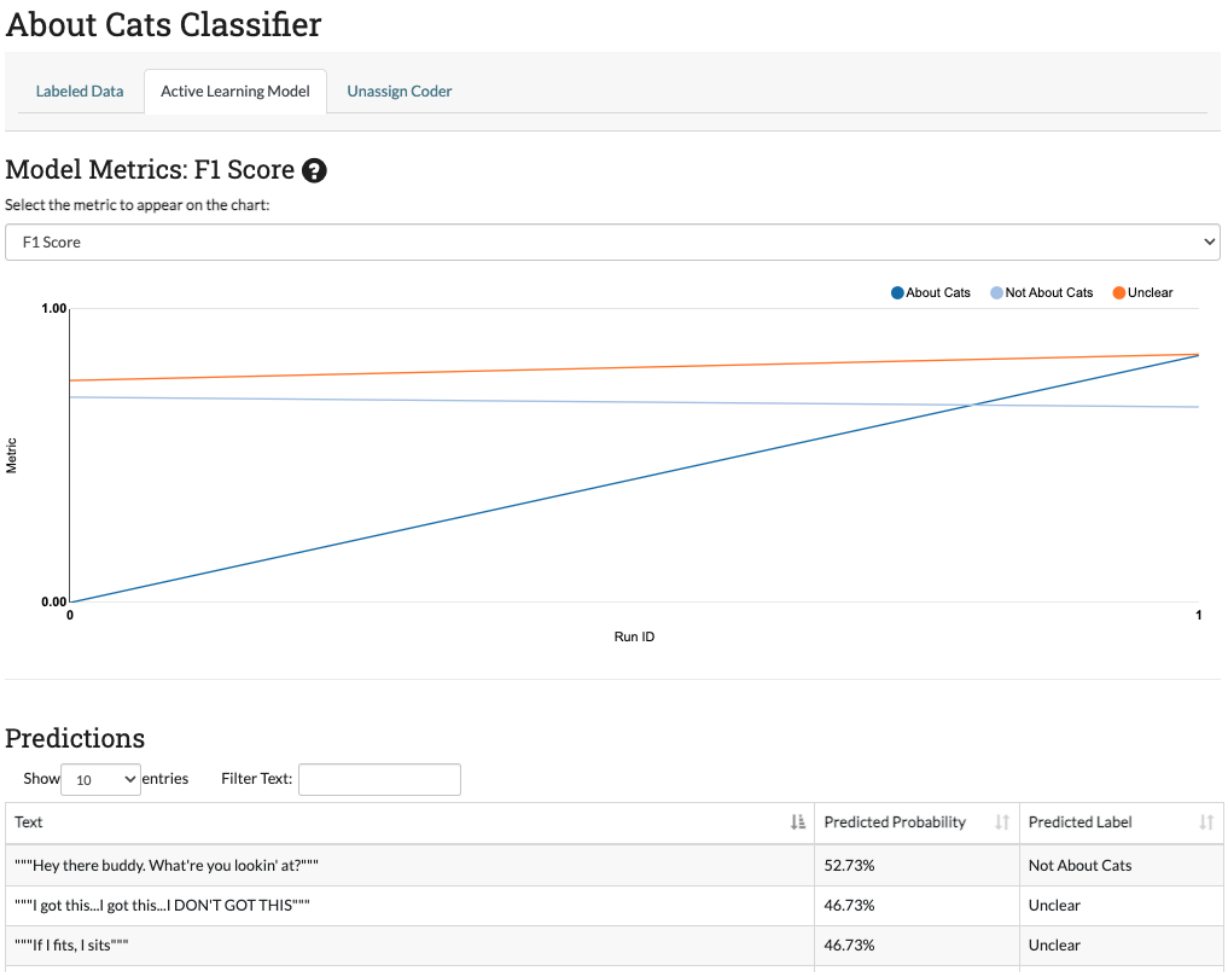
Showing 1 to 10 of 74 entries

2.5.2 Active Learning Model Page

Visible for projects with a model

The model page lets project admin track how well the classifier trained on their labeled data is performing. After each batch of data is labeled, the model retrains on the entire labeled data set. The model page has two main components:

- **The model metrics chart:** This chart shows the change in model accuracy, F1 score, Precision, or Recall (see *Active Learning Metrics* for more information) after each successive batch is labeled. These scores are calculated by running five-fold cross validation on the labeled data. You can change which metric is being displayed using the dropdown above the chart. You can also get a formal definition of the displayed metric by hovering over the (?) symbol next to the title.



About Cats Classifier

Model Metrics: F1 Score

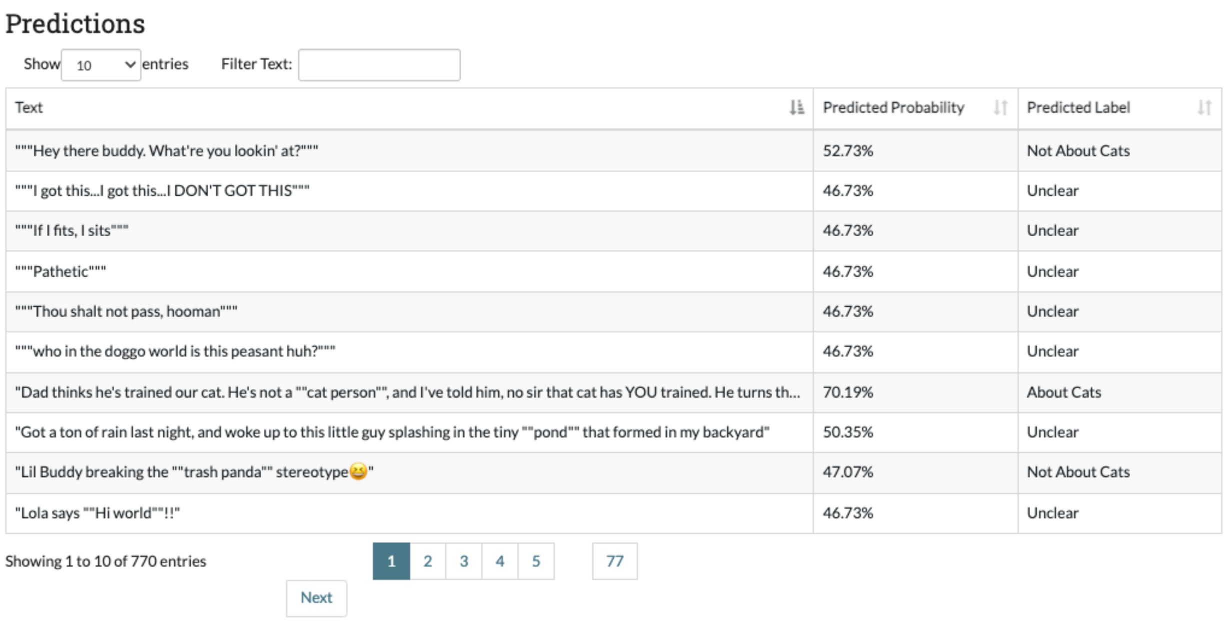
Select the metric to appear on the chart: F1 Score

Predictions

Show 10 entries Filter Text:

Text	Predicted Probability	Predicted Label
"""Hey there buddy. What're you lookin' at?"""	52.73%	Not About Cats
"""I got this...I got this...I DON'T GOT THIS"""	46.73%	Unclear
"""If I fits, I sits"""	46.73%	Unclear

- **The prediction table:** (*only for projects using active learning*) Each time a model is run, SMART then predicts the likelihood of the unlabeled data belonging to each class. The Predictions table shows the label with the highest probability for each unlabeled piece of data. If your project uses Uncertainty-based Active Learning (entropy, margin, or least confident), then the data in the table with lower probabilities (the data where the model is the most "uncertain") is more likely to be chosen for the next batch.



Predictions

Show 10 entries   Filter Text:

Text	Predicted Probability	Predicted Label
"""Hey there buddy. What're you lookin' at?"""	52.73%	Not About Cats
"""I got this...I got this...I DON'T GOT THIS"""	46.73%	Unclear
"""If I fits, I sits"""	46.73%	Unclear
"""Pathetic"""	46.73%	Unclear
"""Thou shalt not pass, hooman"""	46.73%	Unclear
"""who in the doggo world is this peasant huh?"""	46.73%	Unclear
"Dad thinks he's trained our cat. He's not a ""cat person"", and I've told him, no sir that cat has YOU trained. He turns th...	70.19%	About Cats
"Got a ton of rain last night, and woke up to this little guy splashing in the tiny ""pond"" that formed in my backyard"	50.35%	Unclear
"Lil Buddy breaking the ""trash panda"" stereotype😆"	47.07%	Not About Cats
"Lola says ""Hi world""!!"	46.73%	Unclear

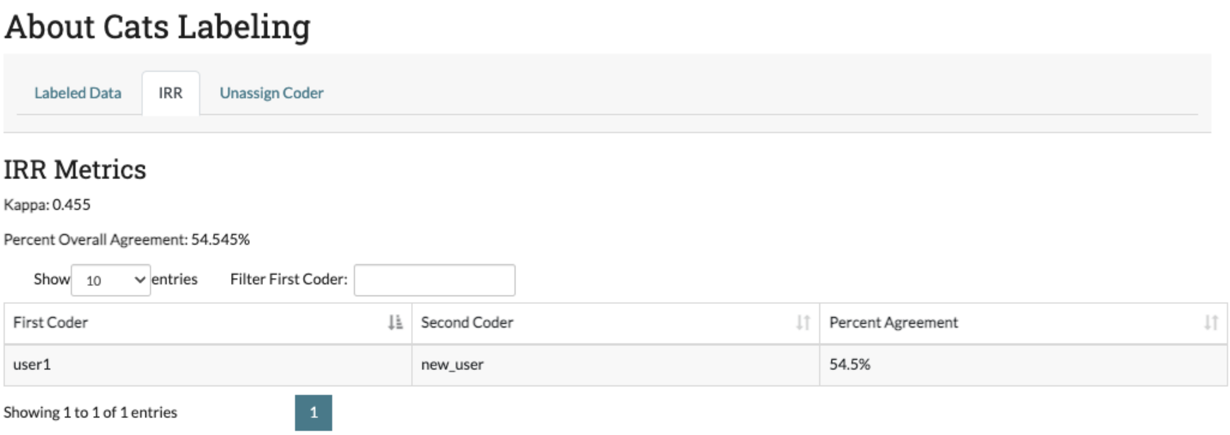
Showing 1 to 10 of 770 entries

2.5.3 IRR Page

Visible for projects that are using IRR

The Inter-Rater Reliability (IRR) page lets admin explore the results of having multiple users label the same data (see *Inter-rater Reliability (IRR)* for a full explanation of IRR). The IRR tab includes four parts:

- **Kappa:** The first value below the IRR Metrics title is a kappa score. This is a common metric for evaluating IRR. This score is calculated using *Cohen's kappa* if the number of required coders is two, and *Fleiss's kappa* if the number of required coders is higher than two.
- **Percent Overall Agreement:** The next value below the kappa gives the percent of IRR data where all coders agreed (note that skipping does not count toward agreement).
- **Pairwise Percent Agreement Table:** Below the numeric metrics, a table is provided with the percent agreement between each pair of coders. In the case where a particular pair has never coded the same IRR data (since there may be more coders on a project then required for IRR), the message "No samples" is displayed.



About Cats Labeling

IRR Metrics

Kappa: 0.455

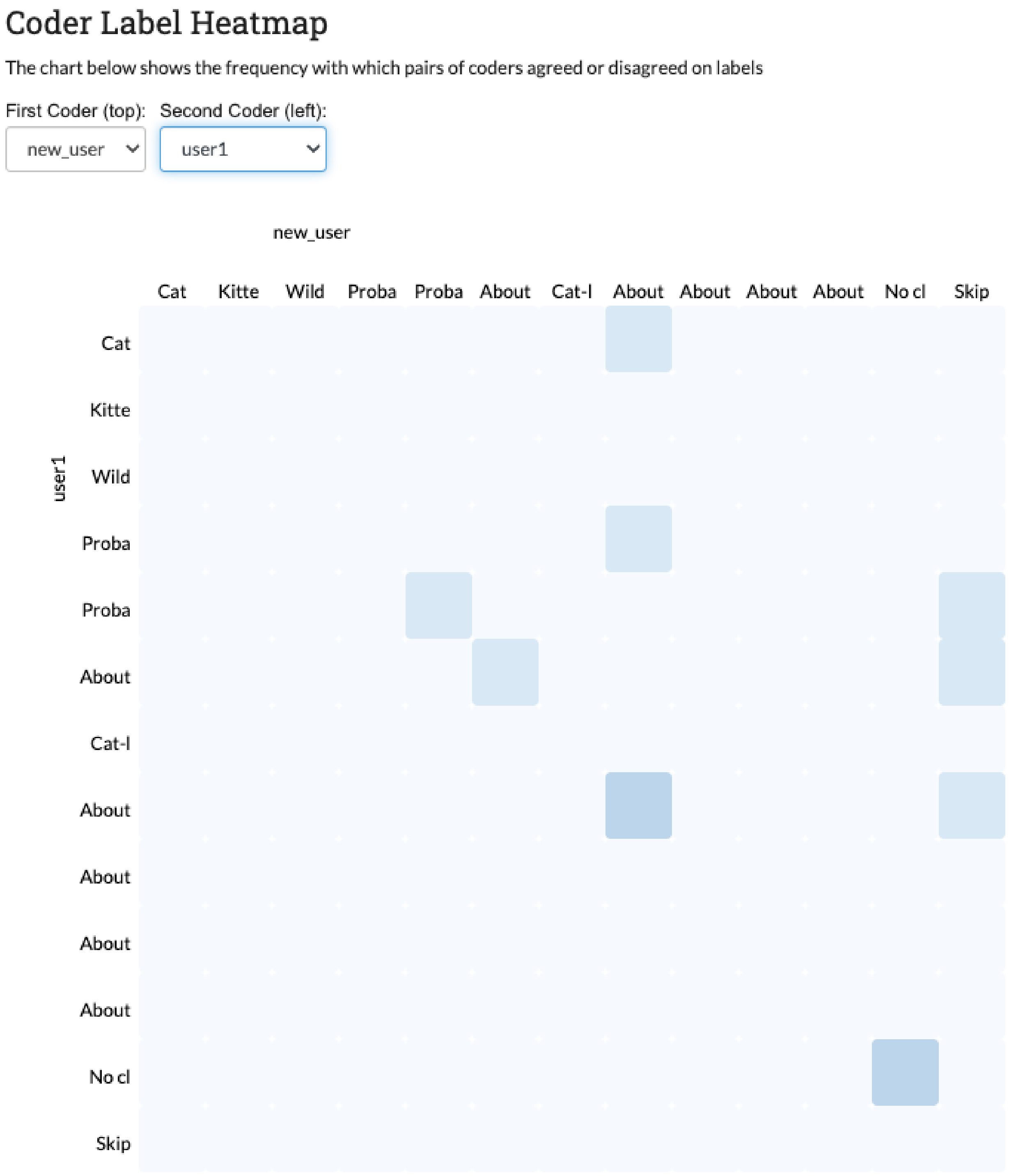
Percent Overall Agreement: 54.545%

Show 10 entries   Filter First Coder:

First Coder	Second Coder	Percent Agreement
user1	new_user	54.5%

Showing 1 to 1 of 1 entries

- **Coder Label Heatmap:** The pairwise relationship in the pairwise percent agreement table can be explored in more detail in the Coder Label Heatmap. An admin can examine how often two coders agreed or disagreed on a label and pinpoint areas of disagreement between coders. You can select two coders to compare using the two dropdowns labeled `First Coder (left)` and `Second Coder (top)` above the chart. The legend on the bottom of the chart corresponds to the number of observations involved.



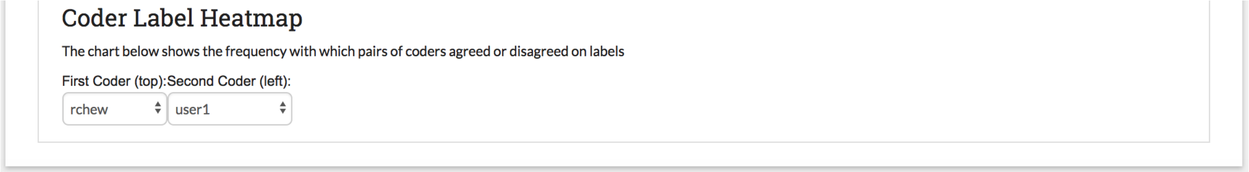
Coder Label Heatmap

The chart below shows the frequency with which pairs of coders agreed or disagreed on labels

First Coder (top): new_user

Second Coder (left): user1

If you select two coders with no samples between them, the heat map will not display:



Coder Label Heatmap

The chart below shows the frequency with which pairs of coders agreed or disagreed on labels

First Coder (top):Second Coder (left):

rchew user1

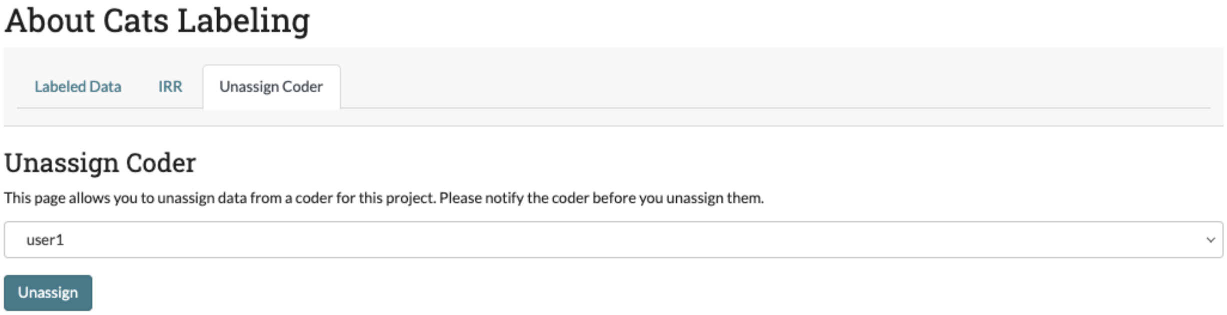
2.5.4 Unassign Coders [NEW]

SMART now provides a screen for admin users to unassign cards which have been assigned to coders. This can be useful for several reasons:

- A coder closed the browser without signing out, leaving their cards assigned.
- A coder has left the annotate tab open on their browser and do not intend to go back to it.
- A coder is leaving the project and the admin wants to reassign their remaining cards across team members.

Note: Cards should be automatically unassigned from users when they go to the project list page, the details page for their project, or sign out. In most cases, you will not need to manually unassign them.

To unassign the cards assigned to user1, we will select them in the dropdown and click "Unassign."



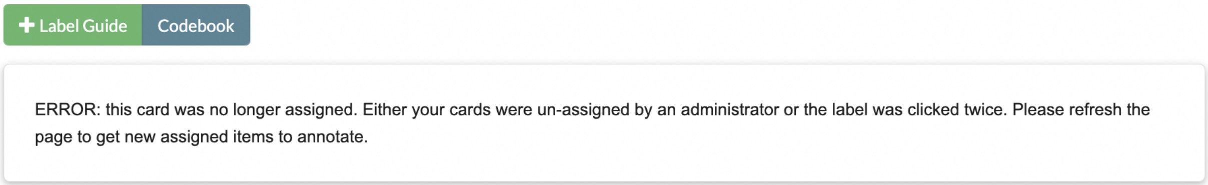
About Cats Labeling

Unassign Coder

This page allows you to unassign data from a coder for this project. Please notify the coder before you unassign them.

user1

When `user1` then goes back to try and annotate the card they were looking at (which may have been passed out to someone else now) they will see the following message:



ERROR: this card was no longer assigned. Either your cards were un-assigned by an administrator or the label was clicked twice. Please refresh the page to get new assigned items to annotate.

2.6 Part 6: Downloading Labeled Data and/or Model

So you have been working hard labeling your data and have accumulated a respectable amount. How do you get the data out of the application and onto your computer? SMART provides a download function that works one of three ways depending on the state and settings of your project:

1. If your project has no data labeled, then the download button does nothing and will display "No Labeled Data to Download".

2. If your project is not using a model or the requirements for a model to run have not yet been met (see *Refilling the Batch*), then the download button will display "Download Labeled Data" and output a comma separated value (.csv) file of the labeled data with the columns `ID` (for the unique ID of the data), `Text`, and `Label`. The data is sorted by `Label`.

3. If your project has a model, then the download button will display "Download Model and Labeled Data". This will output a zip file with:

   a. The labeled data file (see number 2)

   b. A csv with the labels and their internal ID's assigned by the application

   c. A pickle (.pkl) file with the preprocessed version of your input data as a TFIDF matrix

   d. A pickle file with the trained classifier model

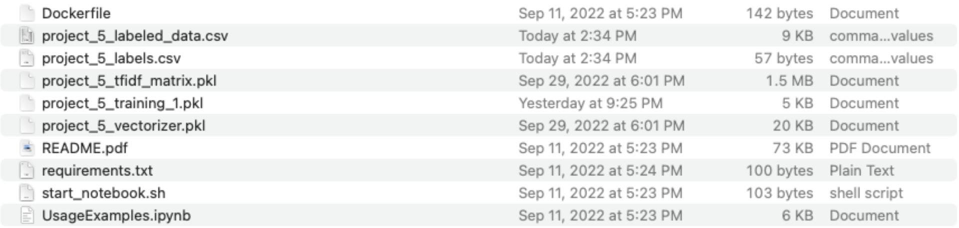
   e. A pickle file with the trained Vectorizer used to preprocess data into the TFIDF format

   f. A README with detailed descriptions of the files and sample code on how to preprocess new data and predict it with your trained model.

   g. A Dockerfile which can be used to setup an environment similar to that of the application.

   h. A script to startup a jupyter notebook server. Not meant to be run outside of the docker container.

   i. A Jupyter Notebook which demonstrates usage of the files and model (see section 4 of the README).

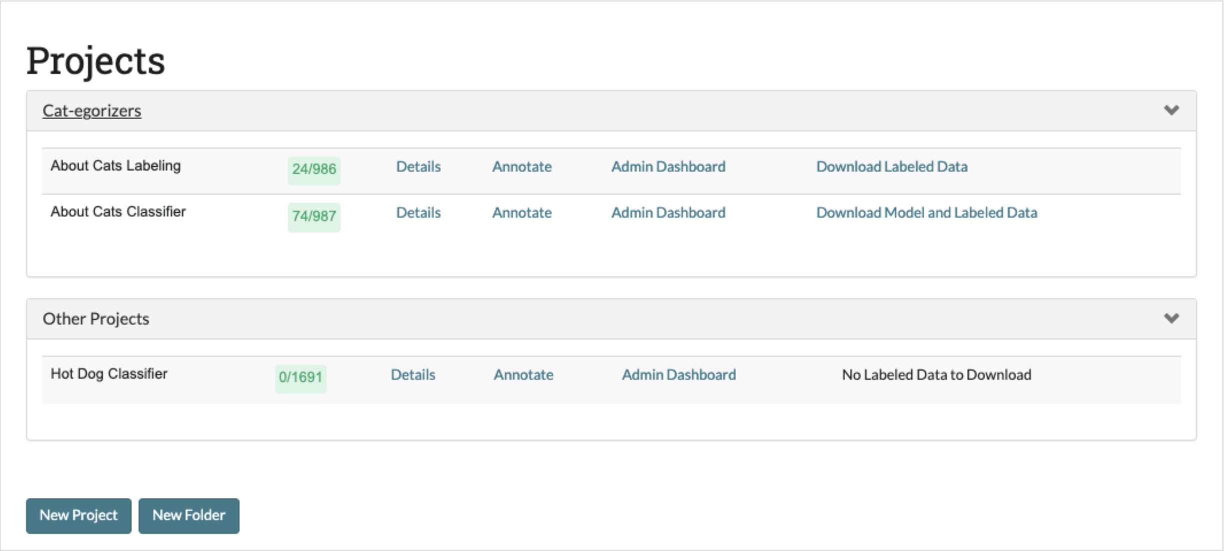


Dockerfile	Sep 11, 2022 at 5:23 PM	142 bytes	Document
project_5_labeled_data.csv	Today at 2:34 PM	9 KB	comma...values
project_5_labels.csv	Today at 2:34 PM	57 bytes	comma...values
project_5_tfidf_matrix.pkl	Sep 29, 2022 at 6:01 PM	1.5 MB	Document
project_5_training_1.pkl	Yesterday at 9:25 PM	5 KB	Document
project_5_vectorizer.pkl	Sep 29, 2022 at 6:01 PM	20 KB	Document
README.pdf	Sep 11, 2022 at 5:23 PM	73 KB	PDF Document
requirements.txt	Sep 11, 2022 at 5:24 PM	100 bytes	Plain Text
start_notebook.sh	Sep 11, 2022 at 5:23 PM	103 bytes	shell script
UsageExamples.ipynb	Sep 11, 2022 at 5:23 PM	6 KB	Document

4. If your project has both verified and unverified labeled data, you will see an additional button which says "Download Only Verified Labeled Data" or "Download Model and Only Verified Labeled Data" depending on if you have a trained model to download. Using this button restricts the data downloaded to just data which has been verified (see *Details Page* for more information).

This button is available in one of two places.

- The Projects Page:



Projects

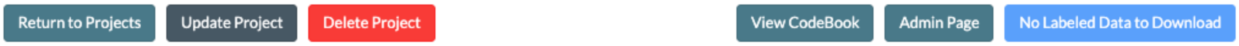
Cat-egorizers

About Cats Labeling	24/986	Details	Annotate	Admin Dashboard	Download Labeled Data
About Cats Classifier	74/987	Details	Annotate	Admin Dashboard	Download Model and Labeled Data

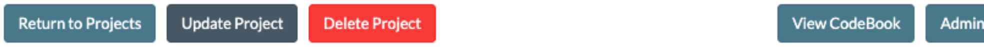
Other Projects

Hot Dog Classifier	0/1691	Details	Annotate	Admin Dashboard	No Labeled Data to Download

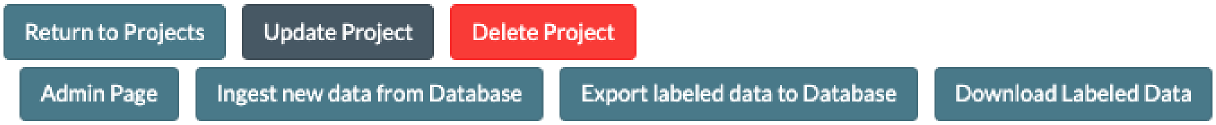
- The bottom of the *Details Page*:



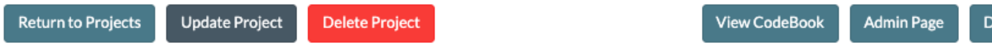
OR (if you have labeled data but no model)



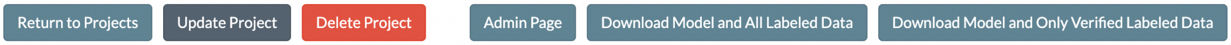
OR (if you have the same project above with a database connection set up)



OR (if you have labeled data and a model)



OR (if you have the same project as above but you also have some verified labels)


Return to Projects
Update Project
Delete Project
Admin Page
Download Model and All Labeled Data
Download Model and Only Verified Labeled Data

2.7 Advanced Feature Details

2.7.1 Active Learning

What is Active Learning?

The process of creating annotated training data for supervised machine learning models is often expensive and time-consuming. **Active Learning** is a branch of machine learning that seeks to minimize the total amount of data required for labeling by strategically sampling observations that provide new insight into the problem. In particular, *Pool-based Active Learning* algorithms seek to select diverse and informative data for annotation (rather than random observations) from a pool of unlabeled data. Active learning algorithms are a cornerstone of the SMART platform, allowing users to utilize these methodologies with minimal costs to projects.

Due to the lack of a universal "one size fits all" active learning algorithm, SMART provides a number of options, allowing users to select the configuration that works best for their situation. Additionally, the first batch is always chosen randomly, both because there must be sufficient training data for Active Learning to work and to mitigate initial bias.

To learn more about active learning, read Settles (2010)[1] for an excellent survey of the active learning literature.

Enabling Active Learning in SMART

Users can enable active learning and select the method for measuring uncertainty in the Advanced Settings page when creating a new project (see *Advanced Settings* for more details).

As of this release, SMART supports **Uncertainty Sampling** with three different measures of uncertainty: *Least Confident*, *Margin*, and *Entropy*. Uncertainty Sampling works by training the model on the existing labeled data and then calculating the probability that each piece of unlabeled data belongs to each possible label. The algorithm returns the most "uncertain" data to be correctly labeled by the coders. The algorithm uses one of three methods to select the unlabeled data that the classifier is the most "uncertain" about:

- **Least Confident** (default): The algorithm chooses the data with the lowest probability for the most likely label using the equation:

$$*arg\,maxx_{LC}^{*} =_{x} 1 - P_{\theta}(\hat{y}|x)$$

where:

$$\hat{y} =_{y} P_{\theta}(y|x)$$

- **Margin Sampling**: The algorithm chooses the data with the smallest difference between the probability of the most likely and least likely labels using the equation:

$$x_{M}^{*} =_{x} [P_{\theta}(\hat{y}_{2}|x) - P_{\theta}(\hat{y}_{1}|x)]$$

where:

`yhat_1` and `yhat_2` are the first and second most likely predictions under the model.

- **Entropy**: The algorithm chooses the most uncertain or "disordered" data by taking the data with the highest score for the entropy equation:

$$x_{H}^{*} =_{x} -\sum_{y} P_{\theta}(\hat{y}|x) * \log P_{\theta}(\hat{y}|x)$$

where:

`y` ranges over all possible labelings of `x`.

Active Learning Metrics

An important consideration in active learning is model performance. To assess your models as your team labels data, SMART provides the following classification model evaluation metrics in the Active Learning Model page of the Admin dashboard:

- **Accuracy**: proportion of observations that the model correctly labeled.

[1] Settles, B. (2012). Active learning. Synthesis Lectures on Artificial Intelligence and Machine Learning, 6(1), 1-114.

$$\frac{TP + TN}{TP + FN + TN + FP}$$

- **Precision**: Indicates how precise the model is at correctly predicting a particular category.

$$\frac{TP}{TP + FN}$$

- **Recall**: Indicates how comprehensive the model is at identify observations of a particular category.

$$\frac{TP}{TP + FP}$$

- **F1-score**: the harmonic mean of Precision and Recall.

$$2 \cdot \frac{(Precision \cdot Recall)}{Precision + Recall}$$

where:

`TP` = True Positive, `FP` = False Positive, `TN` = True Negative, `FN` = False Negative

2.7.2 Inter-rater Reliability (IRR)

What is IRR?

SMART is designed to support labeling projects that may utilize many labelers. When many coders are working on a project, it becomes crucial that coders agree on what labels should apply to what data. Inter-rater Reliability (IRR) is a set of metrics that measures how consistently coders agree with each other, and it is common for a labeling project to require a minimum score for a particular IRR metric for the data to be deemed usable. IRR metrics are calculated from having coders label the same data and examining the results.

Enabling IRR in SMART

Project creators can enable IRR in their SMART projects through the *Advanced Settings* page of project creation. Once IRR is enabled, two additional settings are available:

1. *The percentage of a batch that will be IRR* – This number signifies how much of the data per batch will be used to calculate IRR metrics. This data must be either labeled or skipped by a minimum number of coders before it can be processed.
2. *The minimum number of coders participating in IRR activities* – This number signifies the minimum number of coders that would need to either skip or annotate a piece of IRR data before it can be processed.

As an example, if a project creator chooses 100% for the percentage of the batch that will be IRR and 3 for the minimum number of coders participating in IRR activities, all data in each batch would be required to be labeled by three coders before it could be processed.

Tip: Setting the percentage to 0% is the same as disabling IRR.

IRR Data Flow

If the project creator has enabled IRR, additional steps are added to the data pipeline. First, when the project provides a batch of unlabeled data to label, the previously specified IRR percentage is taken out and marked as IRR. When a user opens the annotation page to begin labeling, SMART first checks if there is any IRR data that SMART has not yet seen. This data is pulled first, and the rest of the deck is filled with non-IRR data. This deck is then shuffled before being presented to the user to make it harder to know what data is IRR. SMART tracks what IRR data has been labeled/sent for adjudication by which users. "Sent to adjudication" is automatically recorded in the internal IRR Log table, while labels are placed in the same label table as non-IRR data (though the training set will not incorporate them as they are marked IRR). Once IRR data has enough people either code it or send it to adjudication, two outcomes can happen:

1. If everyone labeled the datum and these labels were the same, then the datum is added with the agreed upon label to the training set.
2. If any coder sent the datum for adjudication, or coders disagreed on the label, the datum is sent to the admin table for the final label.

After a datum is processed, the labels from all coders are recorded in the IRR Log table.

Note:

- If an admin chooses to discard an IRR datum as unusable, all records of this datum will be flushed from the IRR Log table.

IRR Metrics

To evaluate the reliability of coders, several metrics are calculated for the project admins. This includes percent overall agreement (how many often did everyone give the same label), pairwise percent agreement (how much did two users in particular agree), and a heat map showing the frequency where one coder chose label A and another chose label B (see *IRR Page* for more information). In addition, SMART provides a kappa score, which is a common IRR metric. The kappa score comes from one of the two types below:

Cohen's kappa

This metric is used when there are two coders. Cohen's kappa is most commonly used for categorical items[2]. The general formula is:

$$\kappa = \frac{p_o - p_e}{1 - p_e}$$

where

$$p_o = accuracy$$
$$p_e = \frac{1}{N^2}\sum_k n_{k1}n_{k2}$$

and where `N` is the number of data points, `k` represents the number of possible labels, and `n` is a matrix of label counts of category by labeler (or how many times did each coder choose each label)[2].

p_e is the hypothetical probability of agreeing by chance.

Fleiss's kappa

This metric is the counterpart to Cohen's kappa for more than three coders. The formula is the ratio between the degree of agreement that is attainable above chance, and the degree of agreement actually achieved[3]. The general formula is:

$$\kappa = \frac{\hat{P} - \hat{P}_e}{1 - \hat{P}_e}$$

[2] https://en.wikipedia.org/wiki/Cohen%27s_kappa

[3] https://en.wikipedia.org/wiki/Fleiss%27_kappa

$$\hat{P} = \frac{1}{N}\sum_{i=1}^{N}(\frac{1}{l(l-1)} * \sum_{j=1}^{k} n_{ij}(n_{ij}-1))$$

$$\hat{P_e} = \sum_{j=1}^{k}(\frac{1}{l(N)} * \sum_{i=1}^{N} n_{ij})$$

Where `N` is the number of data points, `k` represents the number of possible labels, `l` is the number of labels for each piece of data, and `n` is a matrix of data points by the number of votes per label[3].

2.7.3 Fix Skewed Label Distributions

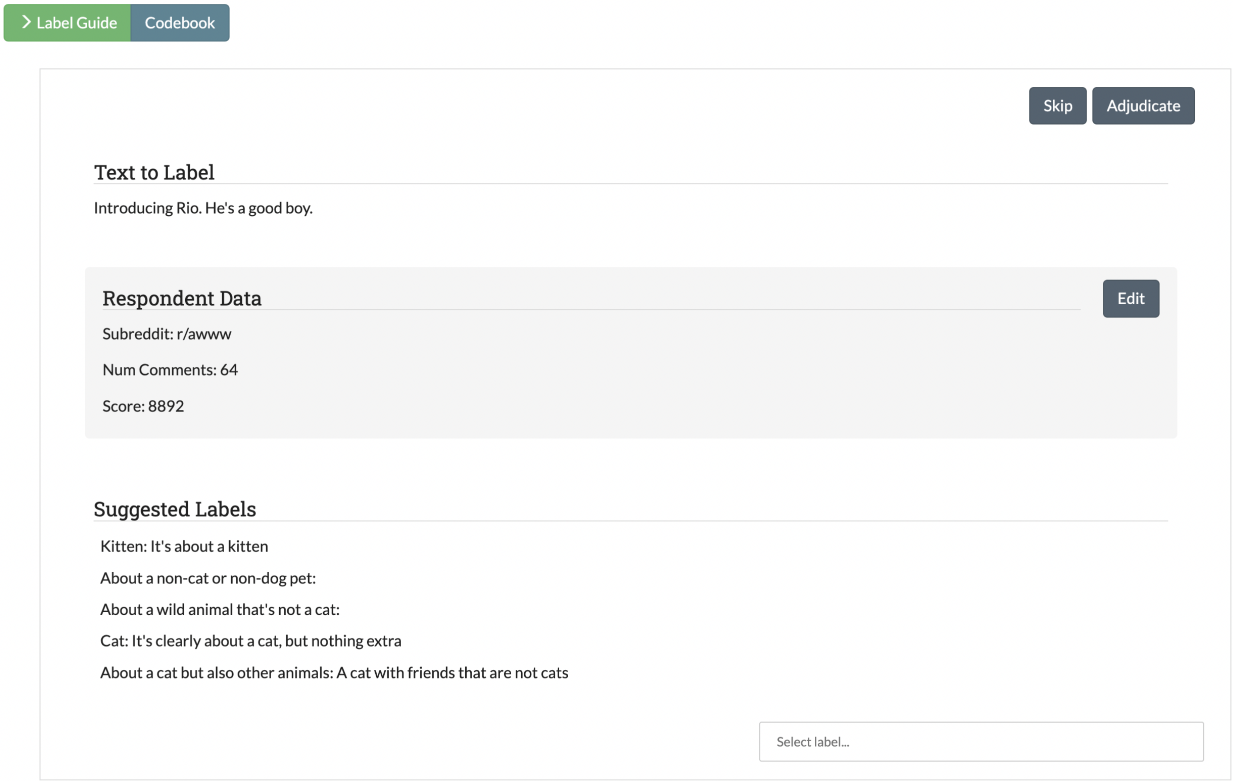
In many applied settings, the distribution of categories the user may be interested in labelling is not well balanced. In particular, if one or more categories of interest occur rarely, labeling observations at random will be particularly inefficient and can quickly exhaust a project's labelling budget. To help combat this issue, SMART implements a version of the *guided learning* strategy outlined in Attenberg and Provost (2010)[4]. This approach treats active learning as a search problem, allowing the user to utilize prior context to identify relevant observations of the rare category, effectively initializing the training batch with a set of relevant rare examples. Findings in Attenberg and Provost (2010)[4] indicate an 8x reduction of real annotation cost per instance using this method on imbalanced data sets when compared to other active learning strategies studied.

See *Fix Skew Page* for more information on using this feature.

2.7.4 Label Embeddings

For projects with more than 5 labels, SMART automatically generates embeddings of the labels and their descriptions. When a user goes to code items, SMART will present the top five label categories based on the cosine similiarty between the text and label embeddings.

[4] Attenberg, J., & Provost, F. (2010). Why label when you can search?: Alternatives to active learning for applying human resources to build classification models under extreme class imbalance. In Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 423-432). ACM.



Text to Label

Introducing Rio. He's a good boy.

Respondent Data

Subreddit: r/awww

Num Comments: 64

Score: 8892

Suggested Labels

Kitten: It's about a kitten

About a non-cat or non-dog pet:

About a wild animal that's not a cat:

Cat: It's clearly about a cat, but nothing extra

About a cat but also other animals: A cat with friends that are not cats

What are Text Embeddings?

A text embedding is a numerical representation of text which can be used for many downstream use cases. At its most basic, a text embedding could be a vector of length N where each dimension is the number of times a specific word appears in the text (i.e., a bag-of-words model). However, more advanced deep learning methods that do not rely solely on term counts have been shown to be able to effectively capture the semantic meaning of text. For example, two sentences could be deemed similiar if they convey similiar meaning, even if they use completely different words.

SMART uses a version of MPNet model[5] from the sentence-transformers library to generate embeddings, mapping input documents and label text to 384 dimensional dense vectors.

2.8 Frequently Asked Questions (FAQs)

2.8.1 General Questions

What does SMART stand for?

SMART stands for “**S**marter **M**anual **A**nnotation for **R**esource-constrained collection of **T**raining data”, an acronym so contrived that it qualifies as both a backronym and a recurvise acronym.

[5] Song, K., Tan, X., Qin, T., Lu, J., & Liu, T. Y. (2020). Mpnet: Masked and permuted pre-training for language understanding. Advances in Neural Information Processing Systems, 33, 16857-16867.

What was the genesis for SMART?

The creation of SMART stems directly from a data-focused business problem that the core development team encounter repeatedly at RTI International; projects that could greatly benefit from having a well trained supervised ML classifier often don't yet have the labelled data to build one. From our experience as practicing data scientists and researchers, the data annotation effort can be particularly arduous in the social science and public health domains where the underlying categories are difficult for humans to categorize and/or are ambiguous. Several of our projects utilizing social media, online news articles, legaslative texts, etc. require careful coordination with project staff or crowdsourced workers to get labelled data that is timely and reliable. SMART was a means to help make this process more efficient, especially in cases where using a crowdsourcing platform isn't possible (e.g., when the data is proprietary or sensitive in nature).

More generally, building labeled data sets is also a commonly reported bottleneck for groups doing applied data science and machine learning in both research and industry. With generous support from the National Consortium for Data Science and RTI International, we were able to build SMART as an open source project to share with the larger data science and applied ML community.

Can I add or remove new class labels after a project is created?

SMART does not allow users to add, remove, or modify class labels after a project is created. This is mainly to prevent awkward interactions with the active learning models. For example, the model evaluation metrics and their associated visualizations become difficult to compare if class definitions are frequently changing.

That being said, we recognize that determining meaningful categories for a new labeling project can be non-trivial. To help users iterate during early-stage labeling projects when model categories are still being decided, we support exporting labeled data from any existing projects and recommend creating a new project (with modified codebook, label descriptions, etc.) to provide a clean annotation experience for your coding team.

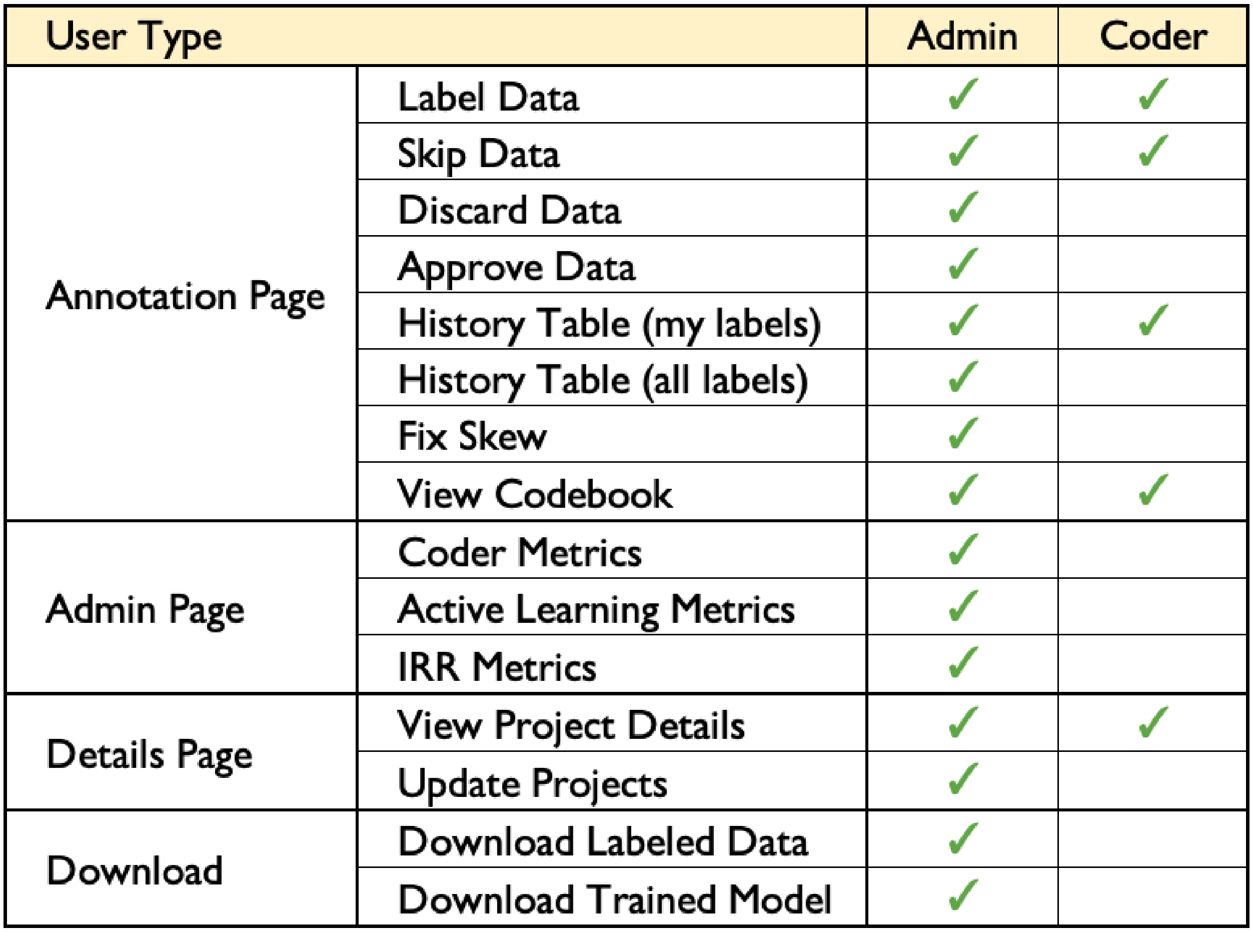
I accidentally mislabeled an observation. How do I correct my mistake?

If you accidentally mislabel a document during the coding process, you can edit the observation in the "History" tab of the annotation page. Label editling is only unavailable if the data was used for IRR and was either resolved due to coder agreement, or an Admin provided a final label after coder disagreement.

> **Warning:** When using active learning, data labels modified on the History tab will not change the model accuracy metrics of past batches displayed on the Active Learning tab of the Admin page, but instead, will update the data for the next model re-training.

What functionality do I get as a coder? Admin?

SMART has two levels of user: `coder` and `admin`. The following chart summarizes what operations are accessible to what level of user:



User Type		Admin	Coder
Annotation Page	Label Data	✓	✓
	Skip Data	✓	✓
	Discard Data	✓	
	Approve Data	✓	
	History Table (my labels)	✓	✓
	History Table (all labels)	✓	
	Fix Skew	✓	
	View Codebook	✓	✓
Admin Page	Coder Metrics	✓	
	Active Learning Metrics	✓	
	IRR Metrics	✓	
Details Page	View Project Details	✓	✓
	Update Projects	✓	
Download	Download Labeled Data	✓	
	Download Trained Model	✓	

2.8.2 Technical Questions

What kinds of supervised ML tasks does SMART support?

As of this version, SMART only supports text classification tasks. However, we hope to extend the platform to support other types of media (images, video, etc.), and perhaps other types of modeling tasks beyond classification in future releases.

What features underly the active learning models?

Currently, SMART uses a term frequency inverse document frequency (TF-IDF) matrix to structure text data. Other popular representations (embeddings, pre-trained language models, etc.) are not currently available in SMART but are welcome additions in future releases.

Can I code for multiple text classification tasks at the same time?

The only way currently to annotate for multiple modeling tasks is to create multiple projects (one for each task). Though multi-task active learning (learning how to select observations that best jointly learn multiple modeling tasks simultaneously) is an exciting area of research, there are not plans for supporting it in the near future.

Do I have to use Active Learning, IRR, etc.?

Depending on your modeling goals, many of the options provided in SMART (active learning, IRR, etc.) may be unnessary or overkill for your use case. To allow users to customize their data labelling experience, users are encouraged to add or remove project features in the Advanced Settings page during project creation.

What are the metrics on the Active Learning page?

The model evaluation metrics presented on the Active Learning section can help you and your team diagnose how a model is performing as more data is labelled. Definitions for the classification evaluation metrics can be found in the *Active Learning* section of the *Advanced Feature Details* page.

What active learning strategies does SMART support?

The active learning strategies implemented in SMART can be found in the *Active Learning* section of the *Advanced Feature Details* page.

Why support labeling data in batches?

We implemented an option to label data in batches due to its practicality. While many active learning strategies assume a sequential back-and-forth between the model and the labeller, waiting for the model to train and predict new examples after every new labeled observation can be prohibitively slow when models are complex or when the underlying data set is large. Additionally, labeling observations in batches more easily allows the labeling process to be spread out among multiple people working on a batch in parallel.

To provide assistance for just this scenario, researchers have developed *batch-mode active learning* algorithms that help assemble batches containing both informative and diverse examples, reducing the chance that observations within a batch will provide redundant information. While effective on large batch sizes, initial tests comparing batch-mode active learning models against simpler non-batch active learning strategies showed similar performance on more modest batch sizes **[link to notebook]**. Due to the complexity of many batch-mode active learning models and similar performance on smaller batch sizes, we forego including batch-mode active learning models in the initial release.

Is the model used to generate embeddings the same as the classifier SMART iteratively trains?

No, while a model is also used to generate embeddings, they are static and do not update as more items are labeled.

Can I customize the label embeddings?

Yes you can! SMART saves the embeddings model it uses in the smart_embeddings_model folder. Depending on the subject domain, you might want to update your model to associate certain phrases as being similar or dissimilar if they are uncommonly used outside of your field.

You can update the SMART embeddings model using the csv to embeddings model repository.

What's the tech stack used to build SMART?

It consists of a Django web application backed by a Postgres database and a Redis store. The application uses Docker in development to aid in dependency management.

2.9 Release Notes and Change Log

2.9.1 Release v.3.0.0

This version has several substantial updates from the previous release. Most of these are quality-of-life improvements or bug fixes, with a few new features added. Specifically, SMART now better supports projects which primarily use SMART as a labeling platform but without using the model features.

> **Warning:** This version has several new features, some of which are better tested than others. We will try and post new updates regularly as bugs are discovered and fixed.

Changes from 2.0.1

New Features

- Metadata fields can now be provided with data and options have been added for deduplication based on this (see "Metadata [NEW]"[TODO link]).
- A database connection can now be set-up for ingesting new data and exporting labeled data. Currently only supports Microsoft SQL Server (MSSQL) databases [TODO verify and link].
- Labels can now be loaded in from files during project creation (see *Part 2: Creating a New Project*).
- Projects can not be added to Groups which mainly just affects the project page (see "Project Grouping [NEW]").
- For projects with more than 5 labels, the top 5 most likely labels are now provided when annotating using label embeddings (see "Most Likely Label Prediction [NEW]").
- The production build has been updated to be more functional (see "Production Settings [NEW]").
- The history table is now searchable and project admin can see and edit the historic labels of all coders.
- The history table can now be filtered by either text or metadata fields, and is paginated by 100 result batches.
- Unlabeled data can now be viewed in the history table for projects that do not use IRR.
- The skew page is now searchable more directly and returns the top 50 items.
- Project admin can now un-assign coders on the admin page, to free up items for other coders to label.
- There is now a timeout in affect for the lock on the admin tables. If an admin has not done anything in the last 15 minutes in a project, and another admin requests the project page, the lock will be given to the new admin.
- The project page now lists counts for the data left to code. This excludes items in the recycle bin.
- The details page now provides counts for the data in the project and where the data is in the coding or IRR pipeline.
- The Skip button has been changed so it merely un-assigns the card to be labeled later. Instead, there is a separate "Adjudicate" button which has the old skip functionality of sending the card to the Admin table. This button now also requires a reason for sending to the Admin.
- Metadata values can now be updated by clicking the "edit" button on any of the cards.

Removed Features

- The progress bar on the coding page has been removed due to both the underlying package not being supported and changes to how SMART assigns data making it less neccessary.

Bug Fixes

- **The email functionality has been fixed so that users who are set an email for their account can use this email to reset their**
    - Please check spam and quarantine folders if unable to find the emails.
    - By default the password reset emails will say they are from "example.com." This can be changed in a deployed SMART instance through the Django admin interface see instructions here.
- The annotate page can now refill itself when there are no cards assigned, instead of relying on other processes like the model build to call a refill. This helps in cases where those other processes fail to refill the queue for some reason.
- The leave-coding-page functionality has been fixed for Chrome, after a recent update disabled it. When broken, Chrome users signing out of SMART would not free up admin tables or un-assign their cards for other users.
- Many small frontend bugs to do with getting long content to render properly have been fixed.
- A button was added to the project create Codebook page to help with removing an uploaded codebook file.
- The project permissions page has been updated to prevent duplicate or conflicting permission assignments.
- The IRR queue now is filled proportionally to how much the non-IRR queue is filled in cases where IRR is not 100% or 0%. Previous to this fix the IRR queue would always add batch size x percent IRR new items to itself each time fill_queue was called even if no items were added to the normal queue, causing the number of items classified as IRR to be far larger than the expected proportion.
- Broken tests have been fixed.

Other Changes

- We are now using pip-tools for backend feature building and maintenance. See the project README section "Dependency management in Python" for instructions on upgrading packages.
- The timezone for frontend-facing date/time output like downloaded labeled data now defaults to EST (see the project README section "Timezones" for instructions for changing the default timezone for the frontend).
- An empty Label column is no longer required for uploaded data with no labels.
- Various charts in SMART have been updated to make them more practical for projects with many labels.
- **The defaults in the Advanced Project Settings page have been updated:**
    - Batch size defaults to 30 instead of the number of labels times 10.
    - By default the model and active learning are turned off and have to be enabled.
    - IRR is disabled by default and must be enabled.
- The steps on the project creation page have been re-arranged so Advanced Settings is last.
- The annotate page has been updated to make things more readable and work with the new Metadata options. In addition projects with many labels will see them appear in a dropdown instead of as individual buttons.
- Frontend dependencies have been updated so that they pull in new bug fix versions.
- Messages for admin lockout or when there are no cards to assign have been updated for clarity.
- Some small GUI changes were made based on feedback from a UX designer.

2.9.2 Release v.2.0.1

This version has no changes from the initial functionality of SMART, but makes the following maintenance updates:

Changes from 1.0.0

- Upgrades frontend and backend packages to maintain usability and patch potential security issues found in older package versions. Several packages had not been maintained, and so had to be removed. The most notable is the package responsible for the loading bar which appears when the user is loading large data files into the software.
- Adds in pre-commit hooks and automated formatting options to make the code cleaner and more readable.
- Replaces the default data from sentiment data challenge with new cleaner dataset.
- *Bug fix*: deck of cards for labeling will not duplicate itself if someone flips through tabs during annotation.
- *Bug fix*: admin charts now automatically resize to fit window when tab changes.
- *Bug fix*: IRR admin table search bar now functions for filtering the first coder field.

2.9.3 Release v.1.0.0

Contributors

- Durk Steed
- Peter Baumgartner
- Rob Chew
- Emily Hadley
- Caroline Kery
- Lucy Liu
- Joey Morris
- Jason Nance
- Keith Richards
- Michael Wenger
- Souliya Chittarath
- Alex Harding

2.10 License