NovaceneAI™ Platform
User manual

October 30, 2024

Contents

Disclaimers

By using the NovaceneAI™ Platform, you agree to the following disclaimers:

Exclusion of Consequential and Related Damages. IN NO EVENT SHALL NOVACENE OR YOU, HAVE ANY LIABILITY TO THE OTHER PARTY FOR ANY LOST PROFITS OR FOR ANY INDIRECT, SPECIAL, INCIDENTAL, PUNITIVE, OR CONSEQUENTIAL DAMAGES HOWEVER CAUSED AND, WHETHER IN CONTRACT, TORT, OR UNDER ANY OTHER THEORY OF LIABILITY, WHETHER OR NOT THE PARTY HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES.

Exclusions. THE LIMITATIONS OF LIABILITY SHALL NOT APPLY TO DAMAGES ARISING FROM A PARTY'S OBLIGATIONS WITH RESPECT TO (I) THE UPLOADING OF PERSONAL DATA TO THE SERVICES WITHOUT SUFFICIENT CONSENT OR AUTHORIZATION; (II) INFRINGEMENT OF A THIRD PARTY'S INTELLECTUAL PROPERTY RIGHTS; (III) ARISING FROM A PARTY'S GROSS NEGLIGENCE, RECKLESSNESS, INTENTIONAL OR WILLFUL MISCONDUCT; (IV) BREACH OF EITHER PARTY'S OBLIGATIONS OF CONFIDENTIALITY; OR (V) VIOLATION OF ANY APPLICABLE LAW.

Disclaimer of Warranty. EXCEPT AS EXPRESSLY PROVIDED HEREIN, NOVACENE MAKES NO WARRANTIES OF ANY KIND, WHETHER EXPRESS, IMPLIED, STATUTORY OR OTHERWISE, AND SPECIFICALLY DISCLAIMS ALL IMPLIED WARRANTIES, INCLUDING ANY WARRANTIES OF MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE, TO THE MAXIMUM EXTENT PERMITTED BY APPLICABLE LAW. WITH RESPECT TO THE USE OF THE SERVICES, NOVACENE MAKES NO EXPRESS OR IMPLIED WARRANTY THAT SERVICES ARE OR WILL BE ENTIRELY SECURE, UNINTERRUPTED, WITHOUT ERROR, OR FREE OF PROGRAM LIMITATIONS. YOU SHALL BE SOLELY RESPONSIBLE FOR ANY AND ALL BREACHES RESULTING FROM ITS OR ITS AUTHORIZED USERS' ACCESS TO THE SERVICES FROM AN UNSECURE PLACE OR NETWORK, OR FROM A JURISDICTION THAT MONITORS NATIONAL INTERNET USE.

What is the NovaceneAI Platform?

The NovaceneAI Platform enables developers, data scientists, research analysts, and non-technical domain experts to harvest insights from their data using artificial intelligence (AI) and machine learning (ML). The process involves three steps: ingesting data, enriching data, and visualizing the insights.

Quick start

A **typical analysis process** is comprised of the following steps:

1. Prepare your data following the data preparation steps
2. Upload your data file (“dataset”) following the data ingestion steps
3. Analyze your dataset using the following the autopilot enrichment steps or data enrichment steps
4. Once an enricher finishes running, load the enriched dataset (the dataset automatically created by the application when you applied the enricher) in the Studio and apply the next enricher in your enrichment sequence. Repeat this process until all enrichers in your sequence have been applied.
5. Visualize the results and create your own reports in the Stage following the data visualization steps.

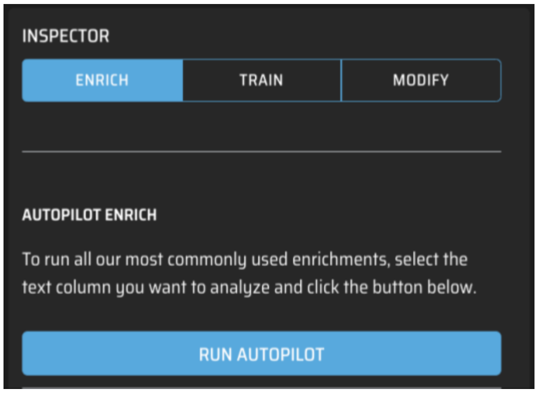
Data preparation steps

1. If you need to analyze data from different files, merge them into one single file
2. Remove any columns unnecessary to your analysis
3. Format cells containing text as ‘**Text**’ to prevent Excel from turning values into formulas
4. Format cells containing dates to use one of the supported date formats
5. Rename columns to use short and descriptive names
6. Save the file in one of the supported file formats, using detailed data preparation steps.

Autopilot Enrichment

To run all of our most commonly used enrichers with one click, take the following steps.

1. Select the column of text you want to analyze in the Studio
2. Hit the “Run Autopilot” button



AUTOPILOT ENRICH

To run all our most commonly used enrichments, select the text column you want to analyze and click the button below.

This will queue up the following enrichers. You can monitor the progress of them running on the Jobs page.

1. Language Translator
2. Clause Extraction
3. Text pre-processing
4. Clustering
5. Cluster theme extraction
6. Sentiment Analysis

You can choose to run further enrichments by selecting the enrichment manually from the dropdown in the Studio.

More Enrichment Sequences

Enrichment sequences describe which enrichers and in which order are to be applied to your data. The below is a sequence for analyzing open-ended feedback, such as survey open-ends, or reviews. **Note:** Before applying the next enricher in the sequence, you need to make sure that the last-enriched dataset is loaded in the Studio.

Step #	Enricher	Target column
1	Language Translator *(if the data is multilingual)*	The column containing the data to analyze
3	Social Media Content Cleanser *(if the text contains content from social media)*	The column containing the processed text
3	Clause Extraction	The column containing the processed text (or the cleansed text if the Social Media Content Cleanser was applied).
4	Sentiment Analysis	The column containing the clauses
5	Clustering	The column containing the clauses
6	Cluster Theme Extraction	The column containing the clauses

Platform overview

In this section:

1. Signing in
2. Overview of the web user interface (UI)

Signing in

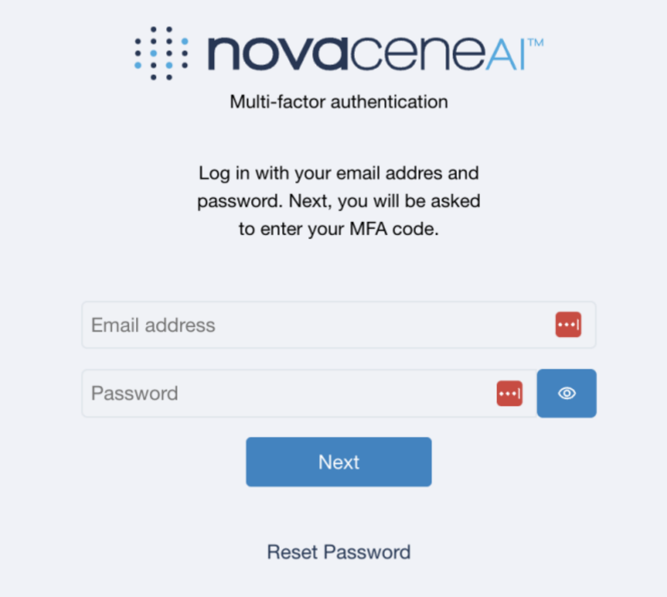
Authentication is available out-of-the-box on instances hosted by NovaceneAI, and as a paid add-on service available for air-gapped instances hosted on AWS. Authentication is not available on air-gapped environments hosted outside AWS.

With this manual, you should have received the URL to your instance and a set of user credentials. If you haven't received it or misplaced it, please contact your administrator.

You are required to set up multi-factor authentication to help keep you and your data secure. This change will require some setup steps.

We recommend you use the Microsoft Authenticator application, however you can also use Google Authenticator (Android and Apple) or LastPass Authenticator (Android and Apple) if you prefer.

Next, head to your landing page, and your login screen will look like the image below. Enter your credentials contained in the invitation email and click "Next".



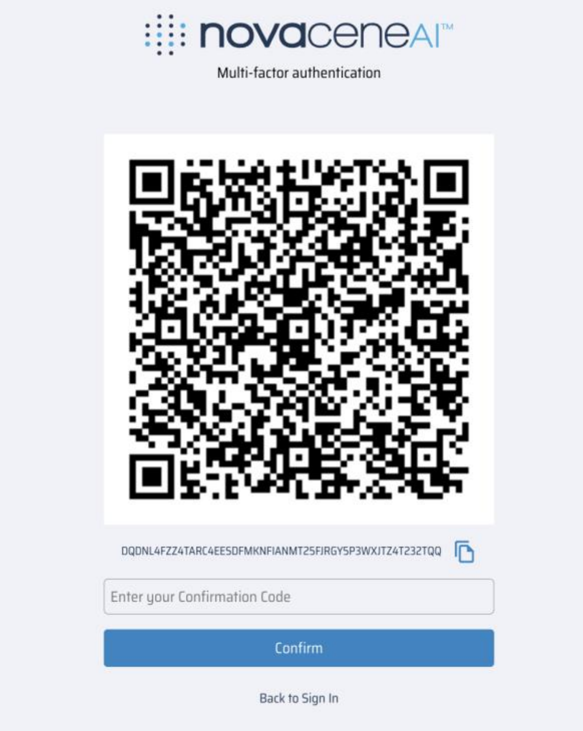
Multi-factor authentication

Log in with your email addres and password. Next, you will be asked to enter your MFA code.

Figure 1: NovaceneAI Platform Log In screen

Next, you'll be prompted to set a new password. Your new password must:

- Be at least 8 characters in length
- Contain at least 1 number
- Contain at least 1 special character (!@#$% etc.)
- Contain at least 1 uppercase and 1 lowercase letter

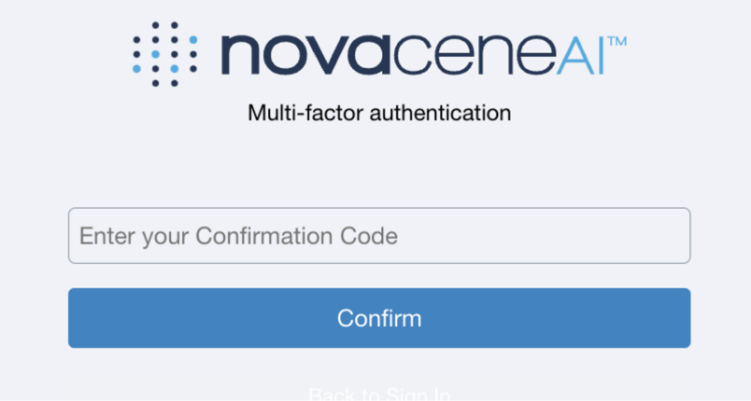


Multi-factor authentication

DQDNL4FZZ4TARC4EESDFMKNFIANMT25FJRGY5P3WXJTZ4T232TQQ

Figure 2: Multi-factor authentication setup screen

Once you select a password, you'll be presented with a QR code. Scan this QR code with your Authenticator app. This will give you a 6-digit code to enter into the platform. Once this step is complete, you're logged in! Now, each time you log in, you will need to get a 6 digit confirmation code from your Authenticator application. To balance security and usability, you will be required to log in again every 7 days.



Multi-factor authentication

Figure 3: Confirmation code screen on login

Overview of the web user interface (UI)

The web UI is organized into four screens:

1. **Data:** Ingest data and organize it into Projects. For information on how to ingest data, consult the

2. **Data** management section.
3. **Studio:** Review and enrich data using various enrichers. For information on how to apply and train enrichers, consult the **Data enrichment** section

Stage: Visualize the data using charts. For information on how to visualize data, consult the

4. Data visualization section.
5. **Lab:** Measure accuracy of trained models. For information on how to measure the accuracy of trained models, consult the **Measuring enricher accuracy** section.
6. **Jobs:** Monitor and review enrichment and training jobs. For information on how to monitor jobs, consult the **Monitoring enrichments** section.

Detailed Data Preparation Steps

Formatting cells

MS Excel or other text editors will allow you to prepare your data for processing in the NovaceneAI platform.

Text

To ensure that cells containing text remain formatted correctly, highlight the columns containing text and select “Text”.

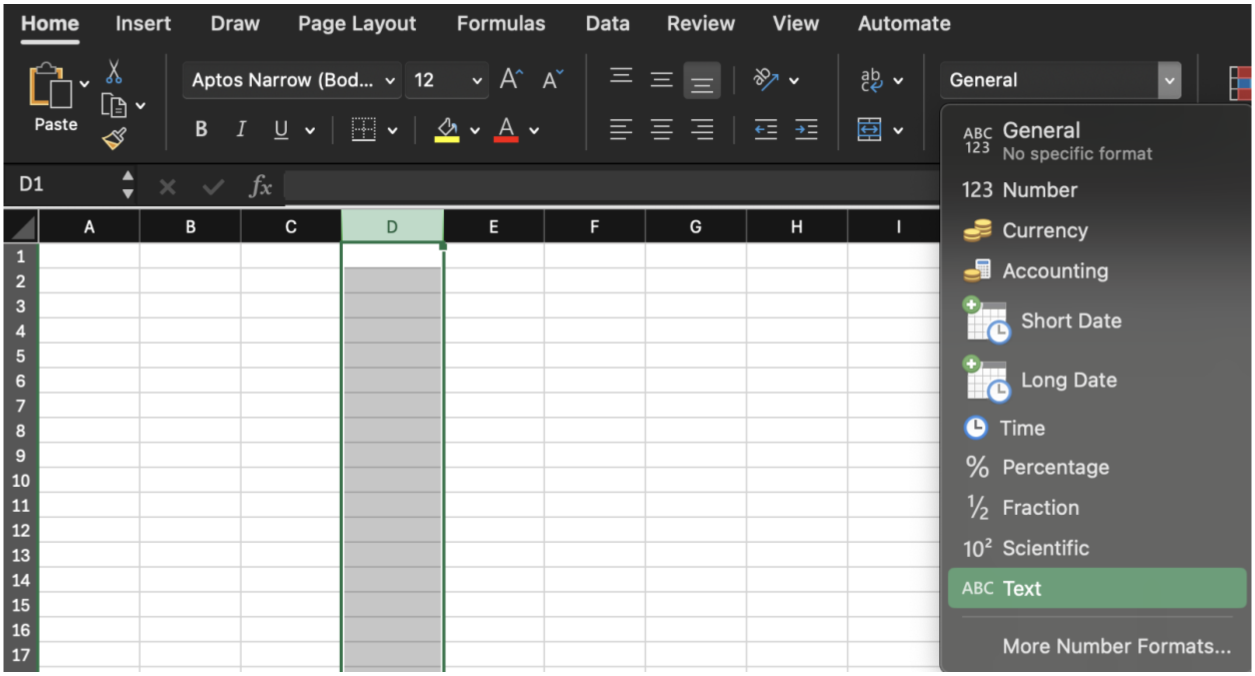


Fig. 4: Formatting a column in excel

Dates

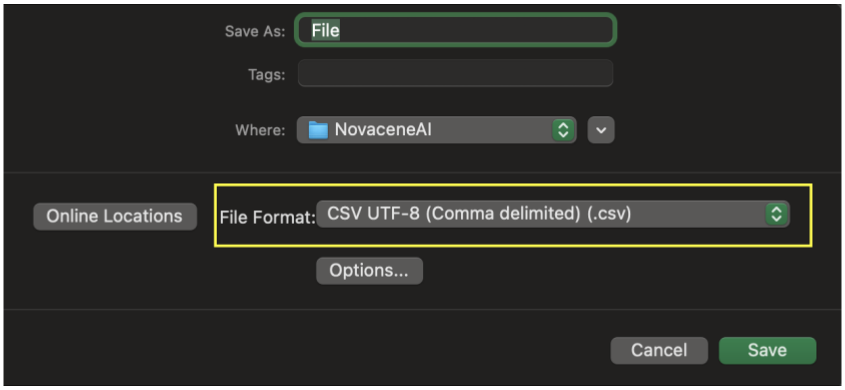
Ensure your dates are in one of the following formats:

- YYYY/MM/DD
- MM/DD/YYYY HH:MM:SS AM/PM
- YYYY-MM-DD HH:MM

MS Excel can help you automatically convert a column to this format.

Exporting data from excel into .CSV UTF-8

If your data is in .xlsx or .xls format, go to File → Save as…, which will open a dialogue where you can select .CSV UTF-8 under “File Type”, as shown below.



Save As: File

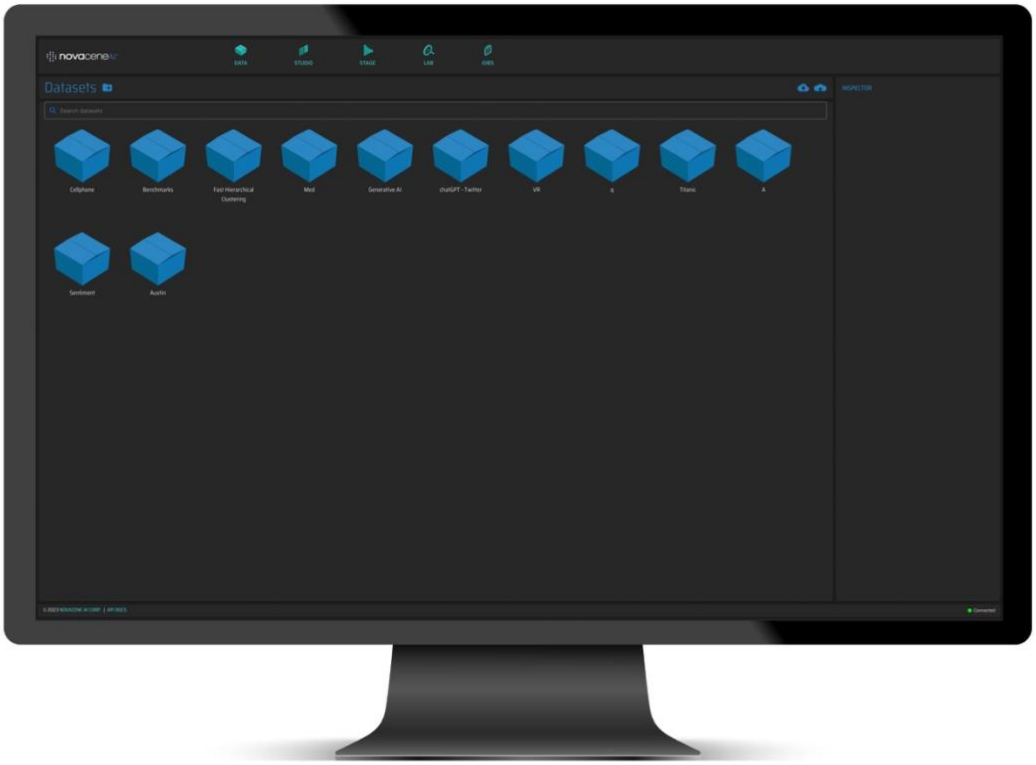
Tags:

Where: NovaceneAI

File Format: CSV UTF-8 (Comma delimited) (.csv)

Data management

The Data screen acts as a repository of all the data you upload to the application, as well as the data that is created within the application through enrichments. Each of these data objects are called *Datasets.*



In this section:

1. Projects
2. Ingestion
3. Export
4. Purging

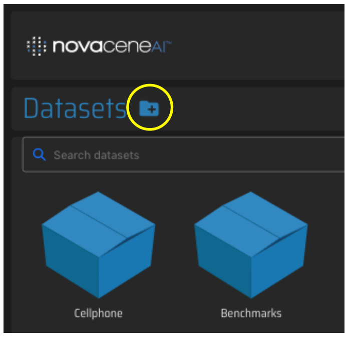
Projects

Projects act as dedicated environments where you can organize and analyze your datasets. These instructions will describe the creation, sharing, and management of these Projects, ensuring a collaborative space where team members can contribute, access, and control data analysis jobs in a secure and structured manner.

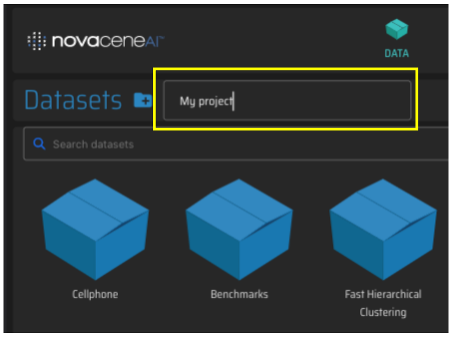
Creating a Project

Initiate your data exploration by creating a new Project. This is your first step towards building a shared workspace where you can collaborate with others to analyze your datasets.

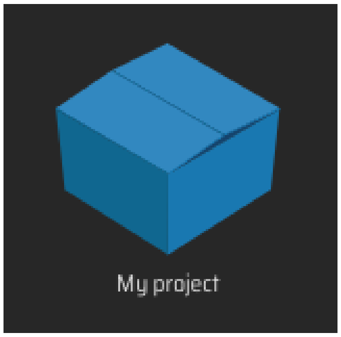
1) Go to the Datasets screen.
2) Click the “Add Project” icon, name your Project, and click enter. Your project will appear on the screen.



1 – Click the “Add Project” icon



2 – Name your Project



3 – The Project will appear on the screen

Once you’ve created the Project, you can enter it by double-clicking the Project box.

Project access and visibility

Maintaining the integrity and confidentiality of your data analysis projects is crucial. This section guides you on how to manage access to your projects, ensuring that only authorized users can view and interact with the sensitive datasets you are working with. Note that only Project owners—the users who created the Project—and Project Collaborators can see and access the datasets saved in the Project. This access restriction extends to saved reports.

Granting and managing collaborator access

Effective data analysis often requires a team effort. Here, we detail the process for inviting Collaborators to your Projects and controlling their access levels. You'll learn how to empower your team with the right access to contribute to data analysis jobs while safeguarding your datasets from unauthorized access.

1) **Granting and revoking access:** Within the Project settings, use the 'Collaborators' section to add users by checking the box next to their email address or username. Revoke a Collaborator's access by unchecking the box. Revoked collaborators will immediately lose access to the data housed in the Project. If they are actively in the Project during revocation, they will remain inside the Project until they leave, but they will not be able to view or interact with any datasets.

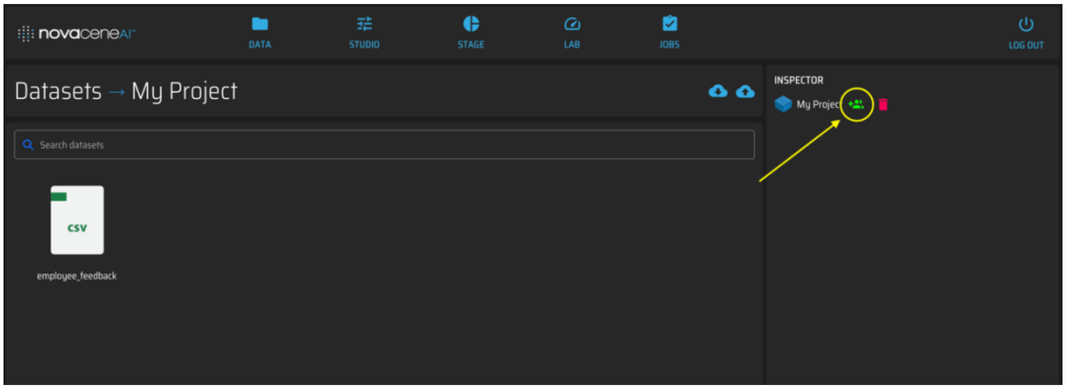
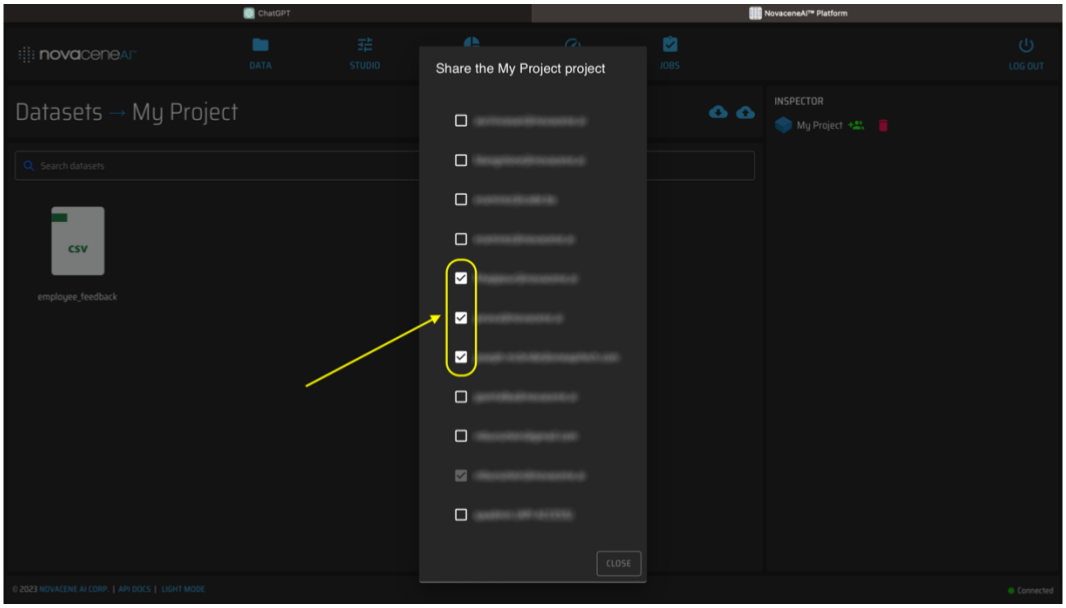


Figure 4: Click the Collaborators icon



Share the My Project project

☐ [redacted]
☐ [redacted]
☐ [redacted]
☐ [redacted]
☑ [redacted]
☑ [redacted]
☑ [redacted]
☐ [redacted]
☐ [redacted]
☑ [redacted]
☐ [redacted]

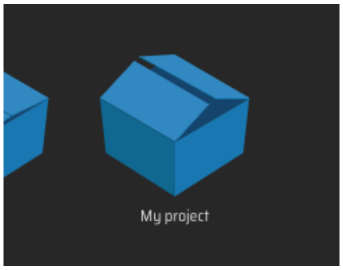
Figure 5: Select users to grant them Collaborator access.

Collaborator permissions:

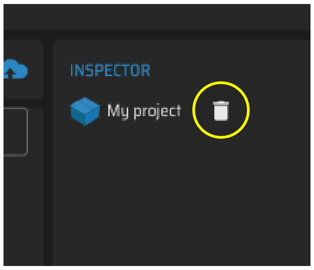
- Access datasets and saved reports associated with the Project.
- Further grant access to other users.
- Revoke the access of any other Collaborators. Note that Collaborators cannot revoke the access of the Project owner.
- Remove themselves from a Project. This action is irreversible through their own accord; they cannot access the Project again unless re-invited.
- Delete the Project.

Deleting Projects

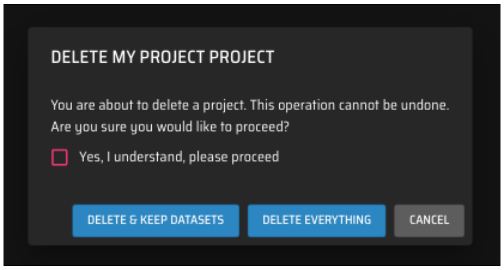
Follow these steps below to delete a Project.



1 – Select the Project you want to delete



2 – Click the Trash icon in the Inspector


DELETE MY PROJECT PROJECT

You are about to delete a project. This operation cannot be undone.
Are you sure you would like to proceed?

□ Yes, I understand, please proceed

3 – Provide confirmation and choose a delete option

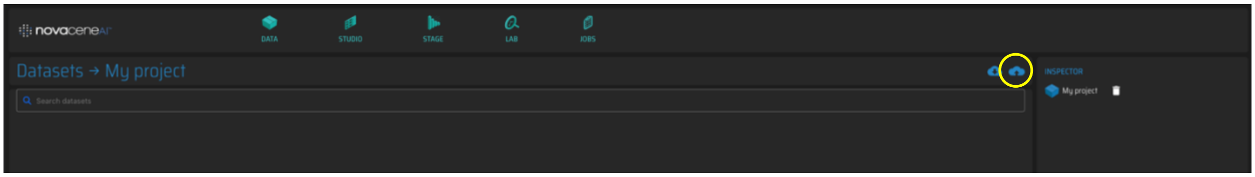
Ingestion

You can bring their data in one of two ways:

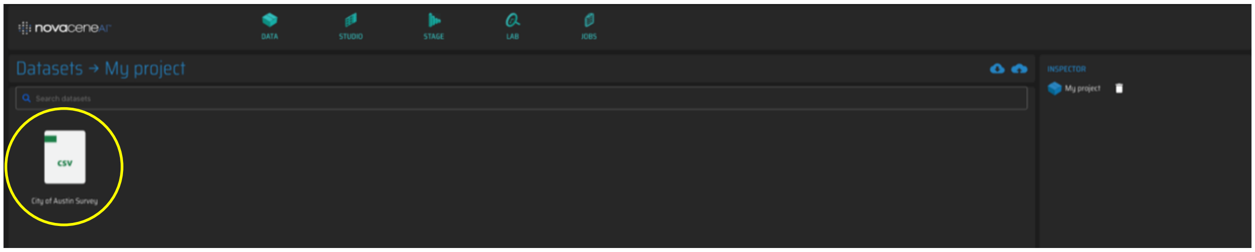
- Uploading files
- Fetching from external sources
- Add-on connectors

Upload

To upload a file, click the Upload icon located on the Data screen, and select a file from your computer.



Once uploaded, the file will appear on the screen.



File types

You can upload files to the application. Supported file formats include:

- CSV UTF-8 (Comma delimited) (.csv)
- Comma Separated values (.csv)
- Macintosh Comma Separated (.csv)
- MS-DOS Comma Separated (.csv)
- ZIP

File uploads can be performed in different ways:

- Using the web application
- Using the API
- Via SFTP

About character encodings: If your CSV file is of an unsupported encoding (i.e. "UTF-16", UTF-16 LE"), you can convert the file using MS Excel and selecting "UTF-8" as the target saved file encoding. Once the CSV file is in "UTF-8" encoding, you will be able to upload it to the application.

Recognized data types

The application automatically recognizes the following data types:

2. **Date:**
   - YYYY/MM/DD
   - MM/DD/YYYY HH:MM:SS AM/PM
   - YYYY-MM-DD HH:MM

   Note: Using any other format may result in inaccurate visualizations.

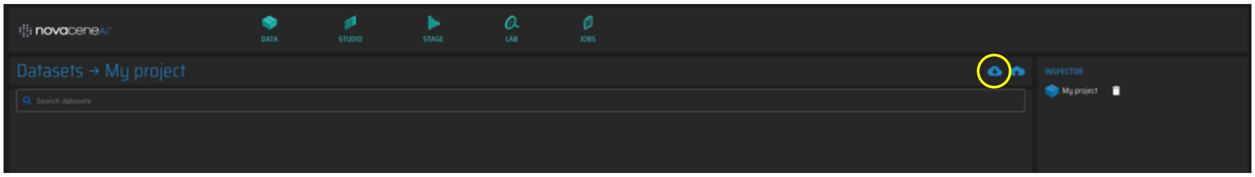
3. **Geolocation:** When the dataset contains two columns labelled: "Latitude" and "Longitude"

Fetch

The application allows you to fetch data from the following sources:

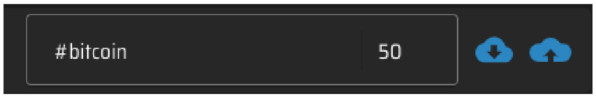
- Twitter
- The web

To fetch data, click on the Fetch icon and follow the instructions below specific to the source from which you are fetching.



Fetching from X (Twitter)

To fetch data, enter a search query in the search box. To fetch from Twitter, type in your query starting with “#”



#bitcoin 50

Note: the application will omit the '#' when querying Twitter. (e.g., #bitcoin will search for bitcoin)

The application will fetch 500 tweets per request. Note that there is a limit of 100 requests per month.

Operators:

Note: depending on your subscription, some operators **may not be available.**

-is:retweet: excludes tweets that were retweeted (includes only original tweets)

Example query: #bitcoin -is:retweet

Columns in generated dataset:

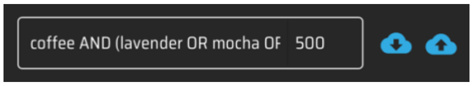
Column heading	Description
Tweet ID	The ID of the Tweet
Created At	The date and time of the Tweet
Author	The Tweet’s author
Followers	The Tweet’s author’s number of followers

Text	The Tweet text
Favourites	The number of times the Tweets was favourited
Retweets	The number of times the Tweets was retweeted
Author Location	The author’s location as specified by the author in their profile
Tweet Country	The Tweet’s country, if available
Tweet Bounding Box	The Tweet’s bounding box, if available
Referenced Tweet ID	The Tweet being retweeted or “Original Tweet” of the Tweet is original and not a retweet.

Table 1: Twitter dataset columns

Fetching from the Web

The Web connector allows you to fetch content found in news sources on the web. To fetch data, enter a search query in the search box, and hit the Enter key.



coffee AND (lavender OR mocha OF 500

The application will search the web for results matching the query, and for each result, it will scrape the content.

Add-on connectors

The following connectors are provided outside of the platform. Contact us to use these connectors.

Reddit

The Reddit connector will fetch posts (“submissions”), comments, and information about posters. To use the connectors, you must provide keywords, the subreddit to search, and the timeframe.

Option	Description
Keywords	See https://support.reddithelp.com/hc/en-us/articles/19696541895316-Available-search-features#h_01HBS25GVN59BBF1TMQCRZT3GA
Subreddit	The name of the subreddit to search (e.g., 3Dprinting)
Timeframe	Must be one of: hour, day, week, month, year, or all

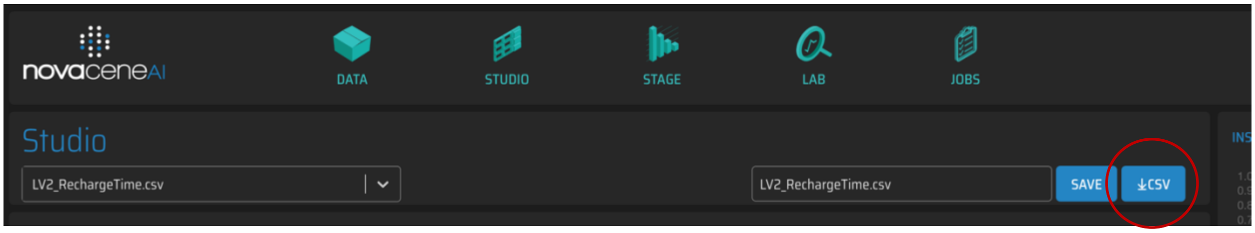
The following table shows the output from this connector:

Option	Description
ID	The ID of the submission or comment (together, “content”)
Subreddit	The Subreddit where the content was posted
Keyword	The keyword used to fetch the submission
Date	The date the content was posted
Title	The title of the submission
Text	The text of the content
Votes	The number of votes on the content
Author	The author who posted the content
Karma	The author’s comment karma
Parent submission	The submission against which the comment was posted
Type	Whether the content is a submission or a comment

Limits: The current connector returns up to a maximum of 250 submissions.

Export

You can export data from the application in CSV format. To export the data, go to the Studio, load the dataset you wish to export, and click the “CSV” button.



Opening exported CSV files in Excel

For best results, **import** the file instead of simply opening. When importing the file, select:

a) Comma as the delimiter
b) “UTF-8” as the “File Origin” option
c) “Text” as “Column data format” option for all columns

Purging

Manual data purging

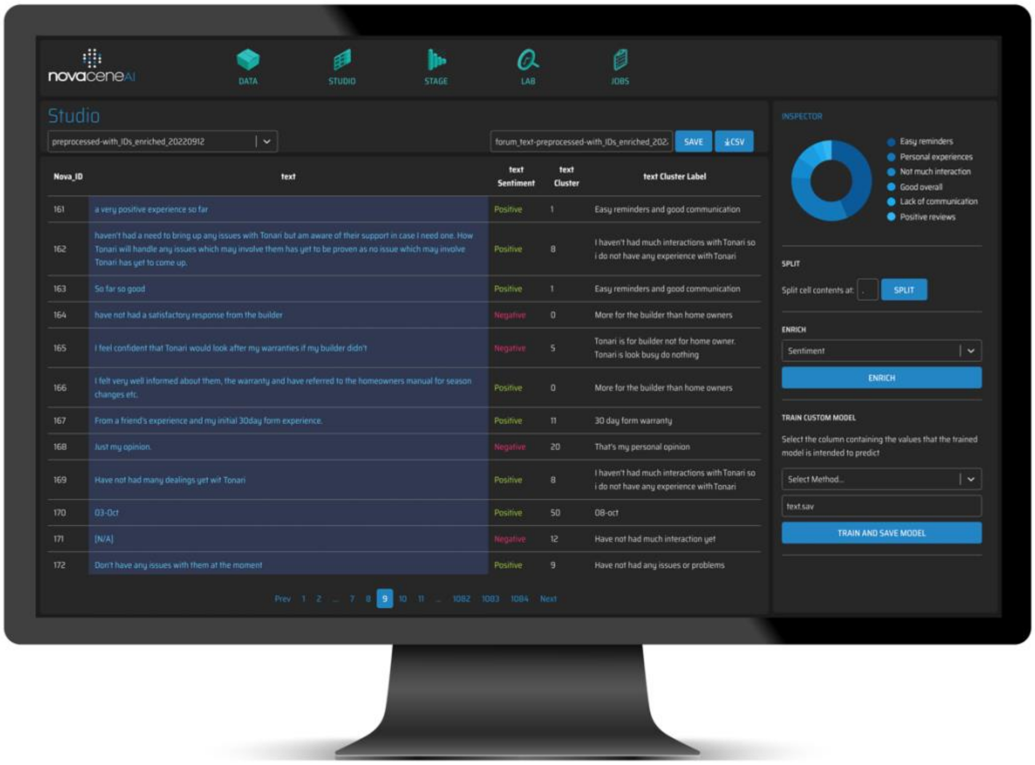
You can delete datasets manually via the web application, or by calling the appropriate API method.

Automated data purging

During initial provisioning, the application is configured to purge datasets automatically after a certain number of days (the retention period). If you process many daily jobs, it is possible that your instance will accumulate several datasets, which may degrade the speed with which the application responds. In this case, please contact support to reduce the retention period, or increase the compute power of your instance.

Data enrichment

Data enrichment represents the core functionality of the application. Enriching data is the act of inferring new data from your existing data using various enrichers. All enrichment operations are performed in the Studio screen.



preprocessed-with_IDs_enriched_20220912

forum_text-preprocessed-with_IDs_enriched_202

Nova_ID	text	text Sentiment	text Cluster	text Cluster Label
161	a very positive experience so far	Positive	1	Easy reminders and good communication
162	haven't had a need to bring up any issues with Tonari but am aware of their support in case I need one. How Tonari will handle any issues which may involve them has yet to be proven as no issue which may involve Tonari has yet to come up.	Positive	8	I haven't had much interactions with Tonari so i do not have any experience with Tonari
163	So far so good	Positive	1	Easy reminders and good communication
164	have not had a satisfactory response from the builder	Negative	0	More for the builder than home owners
165	I feel confident that Tonari would look after my warranties if my builder didn't	Negative	5	Tonari is for builder not for home owner. Tonari is look busy do nothing
166	I felt very well informed about them, the warranty and have referred to the homeowners manual for season changes etc.	Positive	0	More for the builder than home owners
167	From a friend's experience and my initial 30day form experience.	Positive	11	30 day form warranty
168	Just my opinion.	Negative	20	That's my personal opinion
169	Have not had many dealings yet wit Tonari	Positive	8	I haven't had much interactions with Tonari so i do not have any experience with Tonari
170	03-Oct	Positive	50	08-oct
171	[N/A]	Negative	12	Have not had much interaction yet
172	Don't have any issues with them at the moment	Positive	9	Have not had any issues or problems

Split cell contents at: .

Sentiment

Select the column containing the values that the trained model is intended to predict

Select Method...

text.sav

In this section:

1. Built-in enrichers
2. Available built-in enrichers
3. Custom enrichers
4. Monitoring enrichments

Enrichers

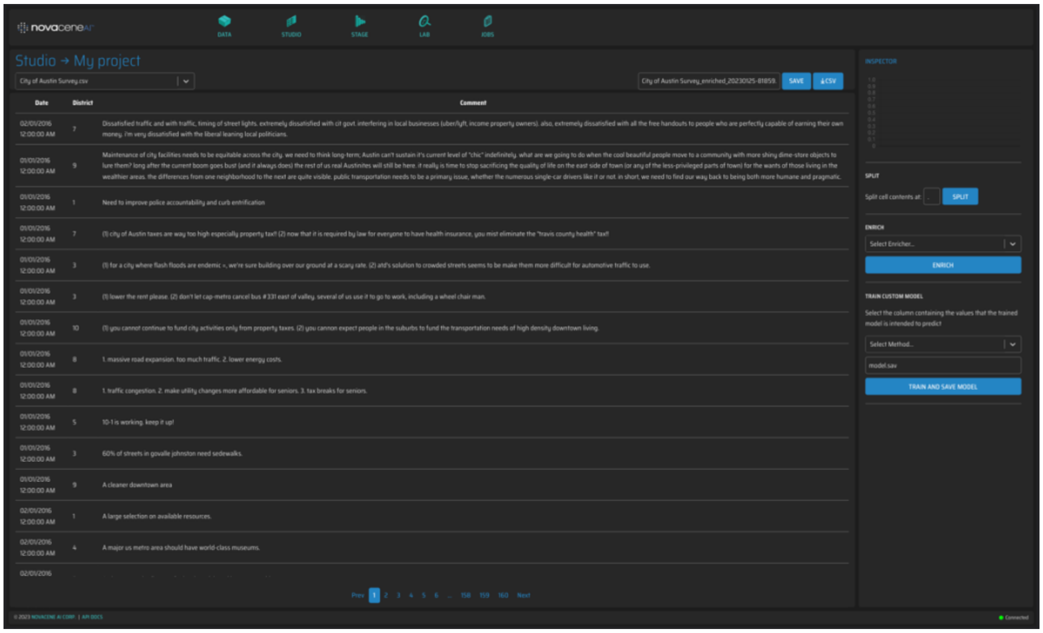
Enrichers are a combination of AI & ML models and algorithms available to you for performing inference and transformations on your data. The application provides built-in enrichers, and you can also train your own.

Built-in enrichers

Built-in enrichers do not require to be trained and can be used out-of-the-box. Trained enrichers are enrichers you trained using your own training data and are therefore adapted to your specific need.

To apply built-in enrichers:

1. Open your dataset in the Studio
2. Select the input column
3. Select the enricher, and click the “Enrich” button



Studio → My project

City of Austin Survey.csv

City of Austin Survey_enriched_20230125-81859.

Date	District	Comment
02/01/2016 12:00:00 AM	7	Dissatisfied traffic and with traffic, timing of street lights. extremely dissatisfied with cit govt. interfering in local businesses (uber/lyft, income property owners). also, extremely dissatisfied with all the free handouts to people who are perfectly capable of earning their own money. i'm very dissatisfied with the liberal leaning local politicians.
01/01/2016 12:00:00 AM	9	Maintenance of city facilities needs to be equitable across the city. we need to think long-term; Austin can't sustain it's current level of "chic" indefinitely. what are we going to do when the cool beautiful people move to a community with more shiny dime-store objects to lure them? long after the current boom goes bust (and it always does) the rest of us real Austinites will still be here. it really is time to stop sacrificing the quality of life on the east side of town (or any of the less-privileged parts of town) for the wants of those living in the wealthier areas. the differences from one neighborhood to the next are quite visible. public transportation needs to be a primary issue, whether the numerous single-car drivers like it or not. in short, we need to find our way back to being both more humane and pragmatic.
01/01/2016 12:00:00 AM	1	Need to improve police accountability and curb entrification
01/01/2016 12:00:00 AM	7	(1) city of Austin taxes are way too high especially property tax!! (2) now that it is required by law for everyone to have health insurance, you mist eliminate the "travis county health" tax!!
01/01/2016 12:00:00 AM	3	(1) for a city where flash floods are endemic =, we're sure building over our ground at a scary rate. (2) atd's solution to crowded streets seems to be make them more difficult for automotive traffic to use.
01/01/2016 12:00:00 AM	3	(1) lower the rent please. (2) don't let cap-metro cancel bus #331 east of valley. several of us use it to go to work, including a wheel chair man.
01/01/2016 12:00:00 AM	10	(1) you cannot continue to fund city activities only from property taxes. (2) you cannon expect people in the suburbs to fund the transportation needs of high density downtown living.
01/01/2016 12:00:00 AM	8	1. massive road expansion. too much traffic. 2. lower energy costs.
01/01/2016 12:00:00 AM	8	1. traffic congestion. 2. make utility changes more affordable for seniors. 3. tax breaks for seniors.
01/01/2016 12:00:00 AM	5	10-1 is working. keep it up!
01/01/2016 12:00:00 AM	3	60% of streets in govalle johnston need sedewalks.
01/01/2016 12:00:00 AM	9	A cleaner downtown area
02/01/2016 12:00:00 AM	1	A large selection on available resources.
02/01/2016 12:00:00 AM	4	A major us metro area should have world-class museums.
02/01/2016		

INSPECTOR

SPLIT

Split cell contents at: .

ENRICH

Select Enricher...

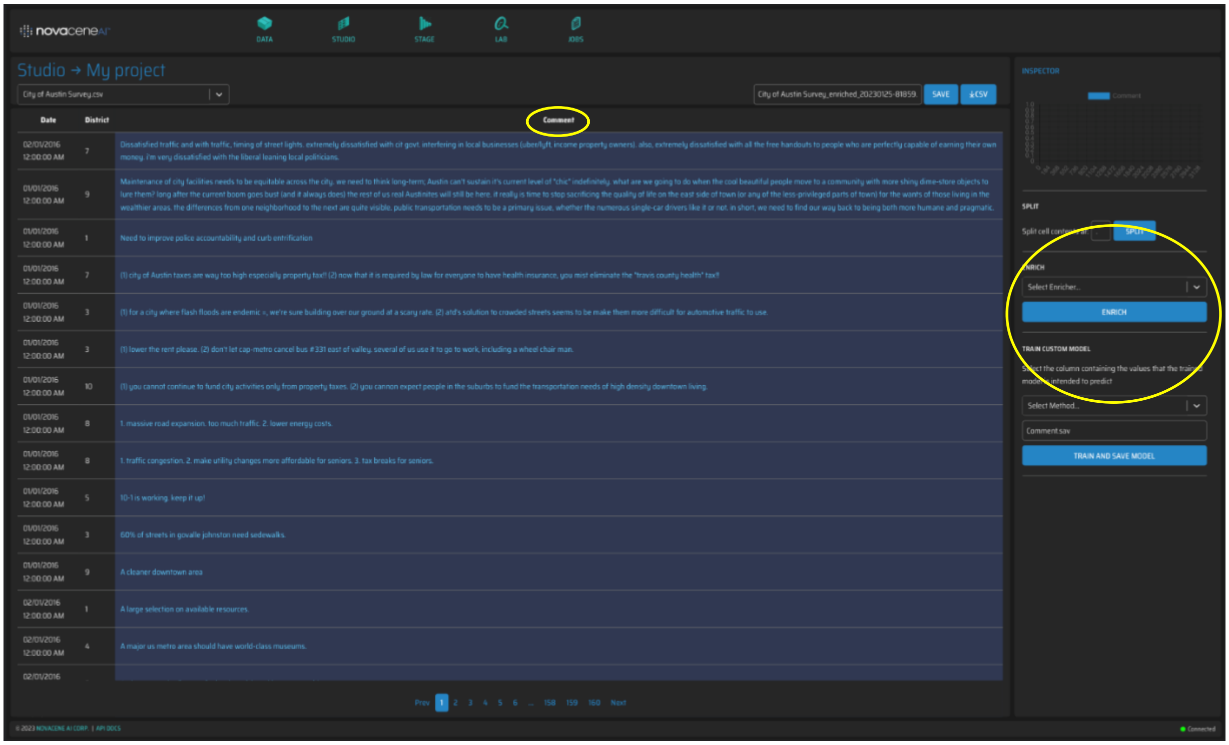
TRAIN CUSTOM MODEL

Select the column containing the values that the trained model is intended to predict

Select Method...

model.sav

Figure 3: To open a dataset in the Studio, either click on the dataset icon on the Data screen and then click the Studio icon on the top navigation bar; or go to the Studio and select the dataset from the drop down menu on the top-left area of the screen



Studio → My project

City of Austin Survey.csv

City of Austin Survey_enriched_20230125-81859.

Date	District	Comment
02/01/2016 12:00:00 AM	7	Dissatisfied traffic and with traffic, timing of street lights. extremely dissatisfied with cit govt. interfering in local businesses (uber/lyft, income property owners). also, extremely dissatisfied with all the free handouts to people who are perfectly capable of earning their own money. i'm very dissatisfied with the liberal leaning local politicians.
01/01/2016 12:00:00 AM	9	Maintenance of city facilities needs to be equitable across the city. we need to think long-term; Austin can't sustain it's current level of "chic" indefinitely. what are we going to do when the cool beautiful people move to a community with more shiny dime-store objects to lure them? long after the current boom goes bust (and it always does) the rest of us real Austinites will still be here. it really is time to stop sacrificing the quality of life on the east side of town (or any of the less-privileged parts of town) for the wants of those living in the wealthier areas. the differences from one neighborhood to the next are quite visible. public transportation needs to be a primary issue, whether the numerous single-car drivers like it or not. in short, we need to find our way back to being both more humane and pragmatic.
01/01/2016 12:00:00 AM	1	Need to improve police accountability and curb entrification
01/01/2016 12:00:00 AM	7	(1) city of Austin taxes are way too high especially property tax!! (2) now that it is required by law for everyone to have health insurance, you mist eliminate the "travis county health" tax!!
01/01/2016 12:00:00 AM	3	(1) for a city where flash floods are endemic =, we're sure building over our ground at a scary rate. (2) atd's solution to crowded streets seems to be make them more difficult for automotive traffic to use.
01/01/2016 12:00:00 AM	3	(1) lower the rent please. (2) don't let cap-metro cancel bus #331 east of valley. several of us use it to go to work, including a wheel chair man.
01/01/2016 12:00:00 AM	10	(1) you cannot continue to fund city activities only from property taxes. (2) you cannon expect people in the suburbs to fund the transportation needs of high density downtown living.
01/01/2016 12:00:00 AM	8	1. massive road expansion. too much traffic. 2. lower energy costs.
01/01/2016 12:00:00 AM	8	1. traffic congestion. 2. make utility changes more affordable for seniors. 3. tax breaks for seniors.
01/01/2016 12:00:00 AM	5	10-1 is working. keep it up!
01/01/2016 12:00:00 AM	3	60% of streets in govalle johnston need sedewalks.
01/01/2016 12:00:00 AM	9	A cleaner downtown area
02/01/2016 12:00:00 AM	1	A large selection on available resources.
02/01/2016 12:00:00 AM	4	A major us metro area should have world-class museums.
02/01/2016		

INSPECTOR

SPLIT

Split cell contents at: .

ENRICH

Select Enricher...

TRAIN CUSTOM MODEL

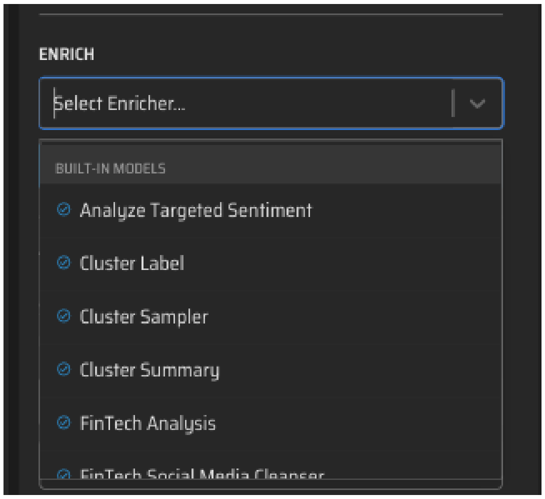
Select the column containing the values that the trained model is intended to predict

Select Method...

Comment.sav

© 2023 NOVACENE AI CORP. | API DOCS

Figure 4: To select the column to which you wish to apply the enricher, click on the column heading

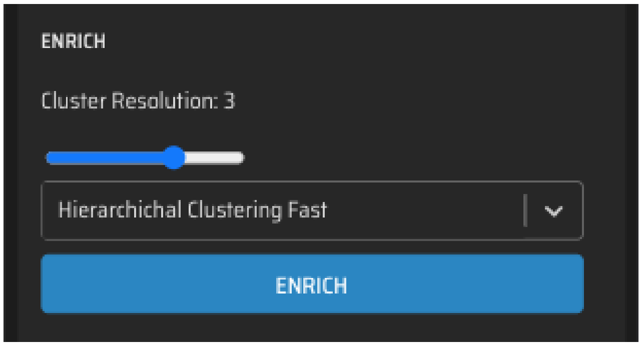

ENRICH

Select Enricher...

BUILT-IN MODELS

- Analyze Targeted Sentiment
- Cluster Label
- Cluster Sampler
- Cluster Summary
- FinTech Analysis
- FinTech Social Media Cleanser

Figure 5: Select the enricher you wish to apply


ENRICH

Cluster Resolution: 3

Hierarchichal Clustering Fast

Figure 6: Adjust settings if applicable, and click the Enrich button.

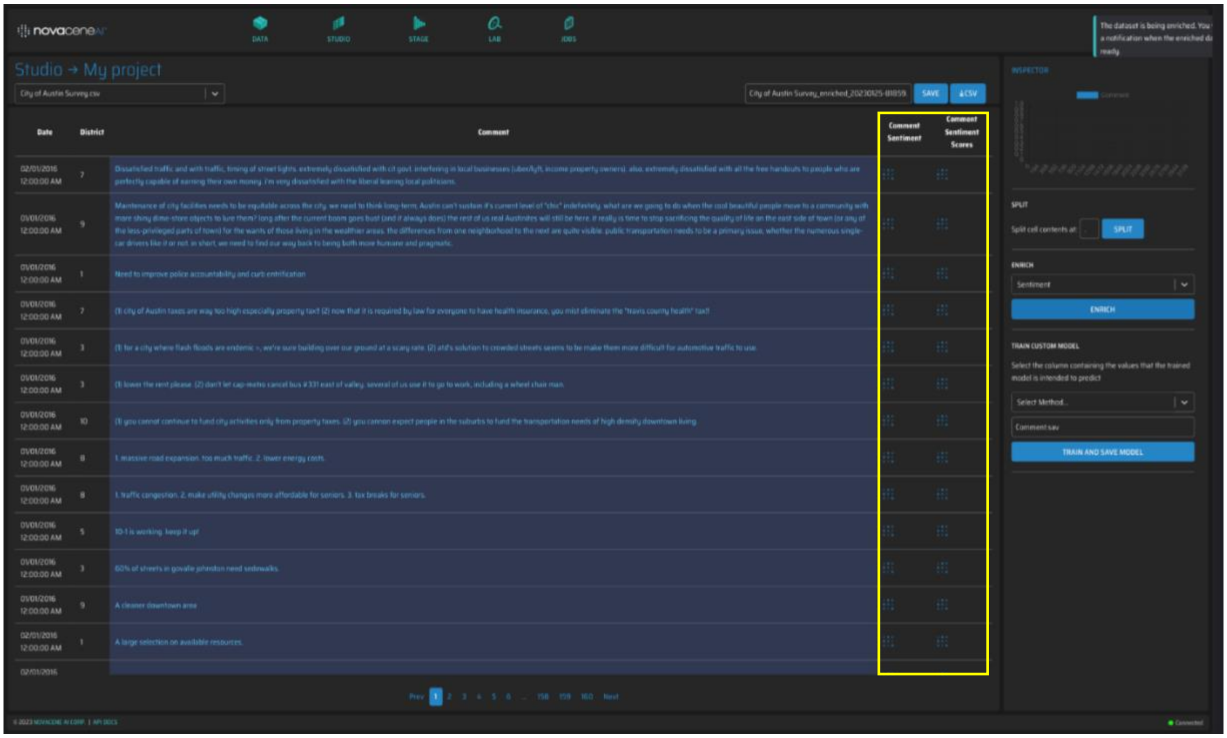
Enriched datasets

Each time you apply an enrichment to a dataset, the application clones the original dataset and adds the results from the enrichment to the clone. This is done to preserve the integrity of datasets and provide you with a trail of datasets as they have been enriched.

Viewing enrichment results

You can see the results from enrichments while they're being executed. Note that the progress of enrichers that act on your entire dataset at one, like clustering enrichers for example, cannot be tracked while the enrichment runs.

As soon as you click the Enrich button, the application will add columns to your dataset where the results from the enrichment will be displayed.



The dataset is being enriched. You a notification when the enriched d ready

Studio → My project

City of Austin Survey.csv

City of Austin Survey_enriched_20230125-181859

Date	District	Comment	Comment Sentiment	Comment Sentiment Scores
02/01/2016 12:00:00 AM	7	Dissatisfied traffic and with traffic, timing of street lights. extremely dissatisfied with cit govt. interfering in local businesses (uber/lyft, income property owners). also, extremely dissatisfied with all the free handouts to people who are perfectly capable of earning their own money. I'm very dissatisfied with the liberal leaning local politicians.		
01/01/2016 12:00:00 AM	9	Maintenance of city facilities needs to be equitable across the city. we need to think long-term. Austin can't sustain it's current level of "chic" indefinitely. what are we going to do when the cool beautiful people move to a community with more shiny dime-store objects to lure them? long after the current boom goes bust (and it always does) the rest of us real Austinites will still be here. it really is time to stop sacrificing the quality of life on the east side of town (or any of the less-privileged parts of town) for the wants of those living in the wealthier areas. the differences from one neighborhood to the next are quite visible. public transportation needs to be a primary issue, whether the numerous single-car drivers like it or not. in short, we need to find our way back to being both more humane and pragmatic.		
01/01/2016 12:00:00 AM	1	Need to improve police accountability and curb entrification		
01/01/2016 12:00:00 AM	7	(1) city of Austin taxes are way too high especially property tax!! (2) now that it is required by law for everyone to have health insurance, you mist eliminate the "travis county health" tax!!		
01/01/2016 12:00:00 AM	3	(1) for a city where flash floods are endemic -, we're sure building over our ground at a scary rate. (2) atd's solution to crowded streets seems to be make them more difficult for automotive traffic to use.		
01/01/2016 12:00:00 AM	3	(1) lower the rent please. (2) don't let cap-metro cancel bus #331 east of valley. several of us use it to go to work, including a wheel chair man.		
01/01/2016 12:00:00 AM	10	(1) you cannot continue to fund city activities only from property taxes. (2) you cannon expect people in the suburbs to fund the transportation needs of high density downtown living		
01/01/2016 12:00:00 AM	8	1. massive road expansion. too much traffic. 2. lower energy costs.		
01/01/2016 12:00:00 AM	8	1. traffic congestion. 2. make utility changes more affordable for seniors. 3. tax breaks for seniors.		
01/01/2016 12:00:00 AM	5	10-1 is working. keep it up!		
01/01/2016 12:00:00 AM	3	60% of streets in govalle johnston need sedewalks.		
01/01/2016 12:00:00 AM	9	A cleaner downtown area		
02/01/2016 12:00:00 AM	1	A large selection on available resources.		
02/01/2016				

Split cell contents at: SPLIT

Sentiment

ENRICH

Select the column containing the values that the trained model is intended to predict

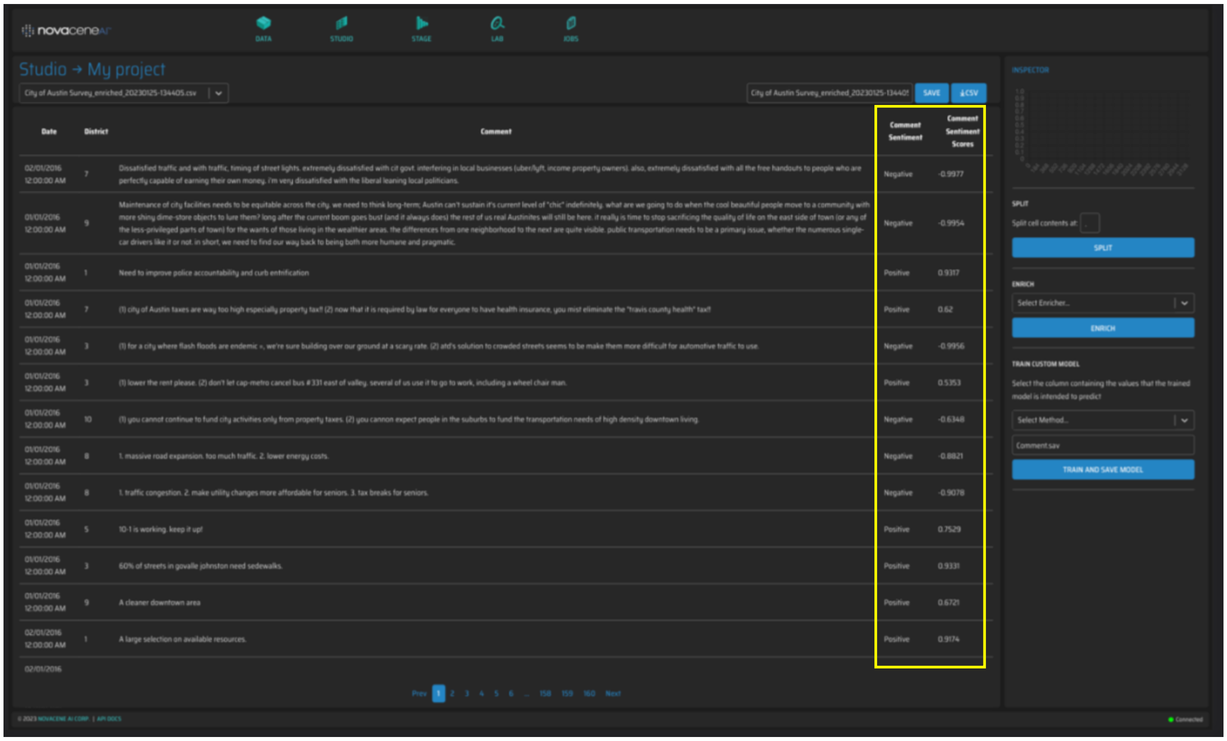
Select Method...

Comment.sav

TRAIN AND SAVE MODEL

Figure 7: Additional columns added by the application to hold the results from the enrichment

As results become available, the application will "push" the results to the Studio:



Studio → My project

City of Austin Survey_enriched_20230125-134405.csv

City of Austin Survey_enriched_20230125-13440!

Date	District	Comment	Comment Sentiment	Comment Sentiment Scores
02/01/2016 12:00:00 AM	7	Dissatisfied traffic and with traffic, timing of street lights. extremely dissatisfied with cit govt. interfering in local businesses (uber/lyft, income property owners). also, extremely dissatisfied with all the free handouts to people who are perfectly capable of earning their own money. i'm very dissatisfied with the liberal leaning local politicians.	Negative	-0.9977
01/01/2016 12:00:00 AM	9	Maintenance of city facilities needs to be equitable across the city. we need to think long-term; Austin can't sustain it's current level of "chic" indefinitely. what are we going to do when the cool beautiful people move to a community with more shiny dime-store objects to lure them? long after the current boom goes bust (and it always does) the rest of us real Austinites will still be here. it really is time to stop sacrificing the quality of life on the east side of town (or any of the less-privileged parts of town) for the wants of those living in the wealthier areas. the differences from one neighborhood to the next are quite visible. public transportation needs to be a primary issue, whether the numerous single-car drivers like it or not. in short, we need to find our way back to being both more humane and pragmatic.	Negative	-0.9954
01/01/2016 12:00:00 AM	1	Need to improve police accountability and curb entrification	Positive	0.9317
01/01/2016 12:00:00 AM	7	(1) city of Austin taxes are way too high especially property tax!! (2) now that it is required by law for everyone to have health insurance, you mist eliminate the "travis county health" tax!!	Positive	0.62
01/01/2016 12:00:00 AM	3	(1) for a city where flash floods are endemic », we're sure building over our ground at a scary rate. (2) atd's solution to crowded streets seems to be make them more difficult for automotive traffic to use.	Negative	-0.9956
01/01/2016 12:00:00 AM	3	(1) lower the rent please. (2) don't let cap-metro cancel bus #331 east of valley. several of us use it to go to work, including a wheel chair man.	Positive	0.5353
01/01/2016 12:00:00 AM	10	(1) you cannot continue to fund city activities only from property taxes. (2) you cannon expect people in the suburbs to fund the transportation needs of high density downtown living.	Negative	-0.6348
01/01/2016 12:00:00 AM	8	1. massive road expansion. too much traffic. 2. lower energy costs.	Negative	-0.8821
01/01/2016 12:00:00 AM	8	1. traffic congestion. 2. make utility changes more affordable for seniors. 3. tax breaks for seniors.	Negative	-0.9078
01/01/2016 12:00:00 AM	5	10-1 is working. keep it up!	Positive	0.7529
01/01/2016 12:00:00 AM	3	60% of streets in govalle johnston need sedewalks.	Positive	0.9331
01/01/2016 12:00:00 AM	9	A cleaner downtown area	Positive	0.6721
02/01/2016 12:00:00 AM	1	A large selection on available resources.	Positive	0.9174
02/01/2016				

Split cell contents at: .

Select the column containing the values that the trained model is intended to predict

Select Method...

Comment.sav

Figure 8: Results from the enrichment job are pushed to the Studio as they become available. Note that the predictions shown are only a preview, and that you will need to load the enriched dataset in order to apply further enrichments

Available built-in enrichers

Clause Extraction

The Clause Extraction enricher splits a paragraph into clauses, allowing you to apply other enrichers to your text in a more granular way.

Why doesn't the Clause Extraction enricher add a column to my dataset on the Studio screen when I click the Enrich button? The Clause Extraction enricher adds both multiple rows and multiple columns to your dataset. Currently, the interface only able to add multiple columns only.

Clustering

The Clustering enricher organizes similar text into categories. Each category is represented by a number, starting with 0.

The *Cluster Resolution* setting affects the number of clusters that are formed. A higher resolution results in many clusters, while a lower resolution results in fewer clusters.

This enricher does not group unclustered documents into a common class (like "Other") but rather it forces all documents to be placed in clusters, even if some clusters end up containing one single document.

Cluster Label[1]

Outputs the single most representative sample in a cluster, providing an understanding of the topic of the samples in the cluster.

Cluster Theme Extraction[1]

Outputs a label based on extractive samples of text in the cluster to provide an idea of the themes in a given cluster. Outputs five pieces of text separated by a bar symbol.

Cluster Summary[1]

Outputs a few samples formatted as a paragraph, providing context as to what the samples in the cluster are about. Requires that the input data has already been clustered.

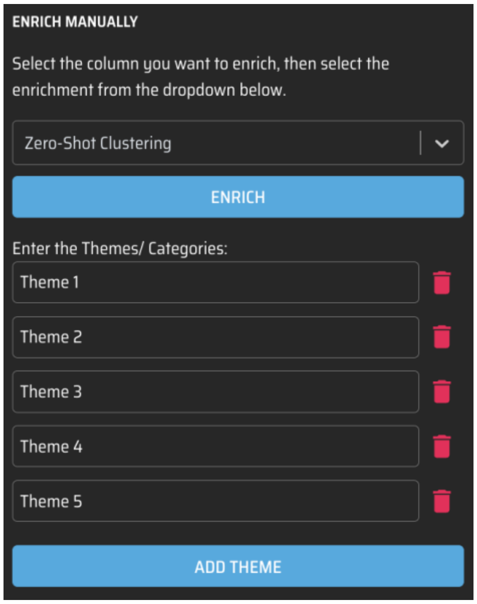
Cluster Summary Generator[1]

Generates a summary of the contents of the cluster using the OpenAI GPT API. Note that to overcome limits imposed by the API, the summary might based on a subset of the contents of each cluster. Requires that the input data has already been clustered.

[1] Requires that the input data has already been clustered.

Zero Shot Clustering

This enricher uses an LLM to categorizes text according to themes input by the user. Users can add or remove themes in the Inspector.



ENRICH MANUALLY

Select the column you want to enrich, then select the enrichment from the dropdown below.

Zero-Shot Clustering

Enter the Themes/ Categories:

Theme 1

Theme 2

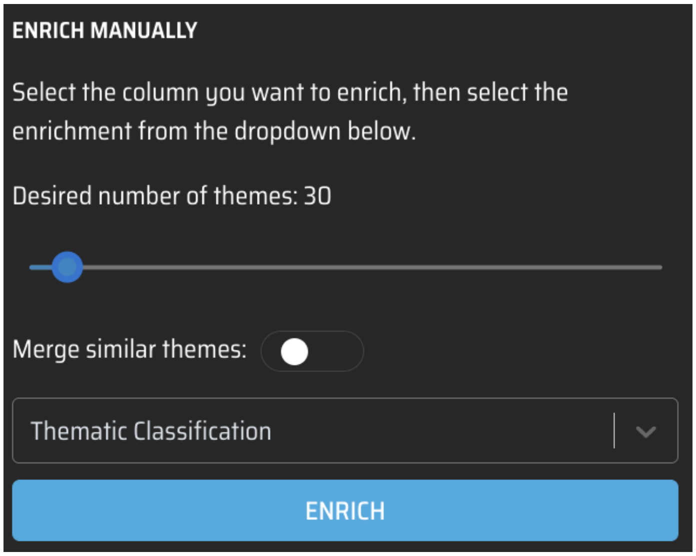
Theme 3

Theme 4

Theme 5

Thematic Classification

This enricher is designed for large datasets. When using Thematic Classification, you can choose the maximum amount of themes you’d like to receive. You can also choose to merge similar themes with a toggle when running this enricher.



ENRICH MANUALLY

Select the column you want to enrich, then select the enrichment from the dropdown below.

Desired number of themes: 30

Merge similar themes:

Thematic Classification

Sentiment Analysis

This enricher analyzes the sentiment of text, and classifies results as positive, neutral, or negative. This enricher also outputs a confidence score of the classification (i.e. a number between 0.00 and 0.99 indicating how certain the model is of the classification)

Sentiment Analysis GPT

This enricher outputs three columns of enrichment. In the first column, it analyzes the sentiment of text, and classifies results as positive, neutral, or negative. In the second column, it describes the emotion of the text. In the third column, it outputs whether the text is sarcastic or sincere.

Emotion Analysis

This enricher analyzes the emotion of text, and outputs an emotion label.

FinTech Analysis

Classifies text into one of these categories: ["mining", "price", "scam", "project", "wallet"]

FinTech Social Media Cleanser

Strips tweets from potentially distracting content such as broken URLs, special symbols, ReTweet (RT) tags, etc.

Hierarchical Clustering

Clusters text by grouping similar content together into categories. Further documentation available at: https://demo.novacene.ai/docs/novacene-api/redoc/#operation/studio_hierarchical_clustering_social_list

Ideas and Comments Classifier

The *Ideas and Comments Classifier* is an enricher in which input text is classified as an *idea* or *comment.* This is useful for idea extraction and analysis. This enricher was trained using an expert-annotated open dataset related to energy generation. Details on the training dataset can be found in **Deprecated enrichers**

Enricher name	Description	Deprecation reason
Sentiment Sampler	Outputs the top 5 most negative samples in the dataset in decreasing order of negative sentiment score.	Ranking to be treated as a function on the visualization system and not as part of the classification algorithm.
Cluster Summary[1]	Performs a one sentence summary using extractive summarization.	Generative methods outperform this extractive method.

Analyze Targeted Sentiment		
FinTech Analysis		
FinTech Social Media Cleanser	Strips tweets from potentially distracting content such as broken URLs, special symbols, ReTweet (RT) tags, etc.	Replaced by HTML cleanser and Social Media Content Cleanser
Hierarchical Clustering Responses	Clusters text by grouping similar content together into categories. (Optimized for some types of survey responses).	Inflexible for different data formats
Quantum Classifier	Binary classification algorithm that runs on a Quantum backend.	
Hierarchical Clustering Social	Clusters text by grouping similar content together into categories. (Optimized for short texts or social media updates).	
Peer Clustering[2] [3]	Clusters records that share many similar attributes.	
Sentence Segmentation	Expands the input text by segmenting input text into separate sentences.	Replaced by Clause Extraction
Sentiment Pre-Processing		Replaced by Sentiment Analysis and Emotion Analysis.
Sentiment Sampler	[TBD]	Please contact us for instructions.
Sentiment Analysis (Retail)	Specific for retail data, classifies the tone of the text as being *positive, negative, or neutral.*	Replaced by Sentiment Analysis
Topic Modelling and Clustering M1	Clustering for text	Earlier versions of Clustering
Topic Modelling and Clustering M2	Clustering for text	Earlier versions of Clustering

Appendix: Supported Languages

Arabic	Dutch	Greek	Italian	Russian
Catalan	Esperanto	Hebrew	Japanese	Spanish
Chinese	Finnish	Hindi	Korean	Swedish
Czech	French	Hungarian	Persian	Turkish
Danish	German	Indonesian	Portuguese	Ukrainian

Appendix: Ideas and Comments Classifier.

Language Translator

The Language Translator enricher detects text in languages other than English and translates it into English. Our API-based language translation service has been updated.

Why do I sometimes get different translations for the same text? Due to the nature of the automatic language translation technology, you may see slightly different translations for the same text each time you apply the enricher. This is normal, and the differences should not affect the meaning of the translation.

Online Language Translator

We support the following languages for our online Language translator:

Supported languages

Afrikaans, Albanian, Amharic, Arabic, Armenian, Azerbaijani, Basque, Belarusian, Bengali, Bosnian, Bulgarian, Catalan, Cebuano, Chichewa, Chinese (Simplified), Chinese (Traditional), Corsican, Croatian, Czech, Danish, Dutch, English, Esperanto, Estonian, Filipino, Finnish, French, Frisian, Galician, Georgian, German, Greek, Gujarati, Haitian Creole, Hausa, Hawaiian, Hebrew, Hindi, Hmong, Hungarian, Icelandic, Igbo, Indonesian, Irish, Italian, Japanese, Javanese, Kannada, Kazakh, Khmer, Korean, Kurdish (Kurmanji), Kyrgyz, Lao, Latin, Latvian, Lithuanian, Luxembourgish, Macedonian, Malagasy, Malay, Malayalam, Maltese, Maori, Marathi, Mongolian, Myanmar (Burmese), Nepali, Norwegian, Pashto, Persian, Polish, Portuguese, Punjabi, Romanian, Russian, Samoan, Scots Gaelic, Serbian, Sesotho, Shona, Sindhi, Sinhala, Slovak, Slovenian, Somali, Spanish, Sundanese, Swahili, Swedish, Tajik, Tamil, Telugu, Thai, Turkish, Ukrainian, Urdu, Uzbek, Vietnamese, Welsh, Xhosa, Yiddish, Yoruba, Zulu

Offline Language Translator

We support the following languages for our offline Language Translator

Supported languages

Arabic, Azerbaijani, Catalan, Chinese, Czech, Danish, Dutch, English, Esperanto, Finnish, French, German, Greek, Hebrew, Hindi, Hungarian, Indonesian, Italian, Japanese, Korean, Persian, Polish, Portuguese, Russian, Spanish, Swedish, Turkish, Ukrainian

Named Entity Recognition

The Named Entity Recognition (NER) enricher extracts entities from text. The following entities are recognized:

Entity Type	Description
LOC	Location name

PER	Person name
ORG	Organization name
MISC	Other name

Tips for best results when using NER

When extracting organizations (ORG) from a dataset containing social media data like tweets, it is recommended that you first clean the text using the Social Media Content Cleanser enricher, followed by the Language Translator enricher to translate any text that isn't English.

Peer Clustering

The *Peer Clustering* enricher groups together cases based on the number of shared attributes amongst them. For example, it can be used to group universities by the type of software tools they use, or to group restaurant orders by the type of dishes that have been ordered.

Preparing your data

This enricher requires that data be provided in the following format:

Case id	Type	Value

For example, if you are trying to group together restaurant orders, your data may look like this:

Order id	Meal type	Dish ordered
1	Appetizer	Soup
1	Main	Chicken
1	Dessert	Cake
2	Main	Chicken
2	Main	Lasagna
3	Appetizer	Soup
3	Dessert	Ice cream

Note that you need to supply a minimum of 100 unique **case ids** to use this enricher.

Applying the enricher

To apply the enricher, load the dataset using the format shown above, open the dataset in the Studio, select the Peer Clustering enricher from the dropdown and click the Enrich button. Note that you do not need to select a column to apply this enricher.

Interpreting the results

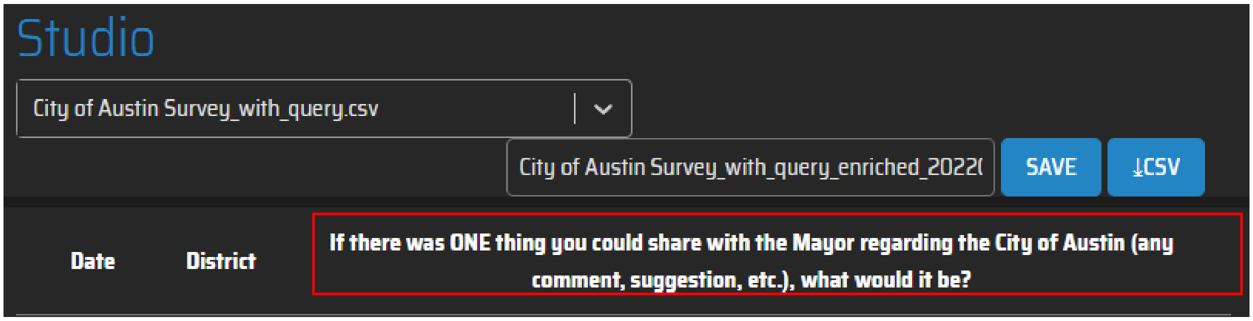
The enricher will output a flattened dataset with a number of rows matching the number of unique case ids. Each row will contain all the information related to each case, and the last column will contain a number representing the cluster. You can then filter by the this Peer Cluster number to see the cases that have been grouped together.

Term Frequency

The *Term* Frequency enricher can output two types of grams with their corresponding counts, 1) Unigrams and 2) Noun Phrases.

1) Unigrams – an n-gram that consists of a single item from a sequence (input text)
2) Noun Phrases – a part of speech pattern that contains nouns, adjectives, or verbs.

In addition, the enricher includes a parameter that can be set, that is, the use of *context-aware stop words*. This feature is beneficial when a *query* is set as the column name for the input data.



City of Austin Survey_with_query.csv

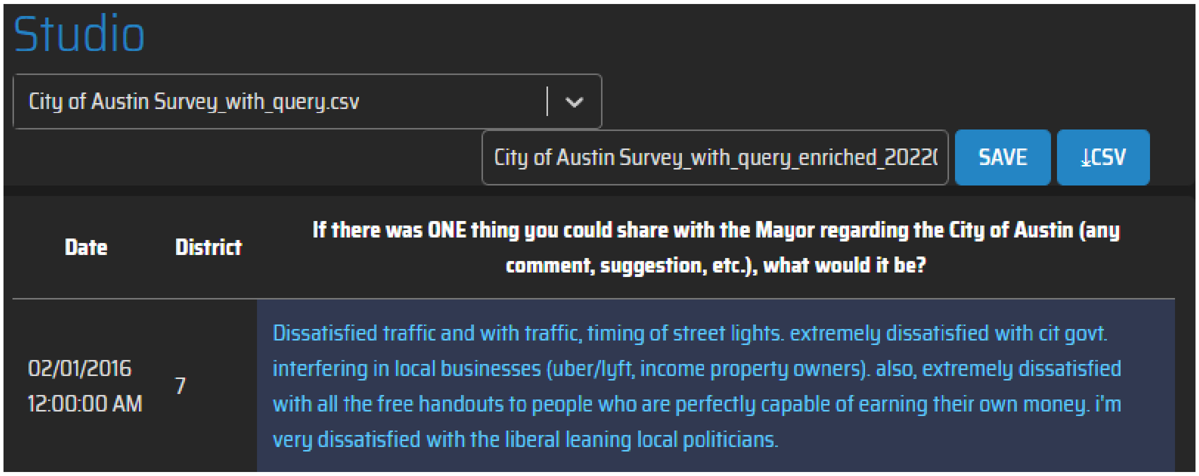
City of Austin Survey_with_query_enriched_2022(

Date	District	If there was ONE thing you could share with the Mayor regarding the City of Austin (any comment, suggestion, etc.), what would it be?

By enabling *context-aware* stop words, the enricher will remove words based on the *query* that may not be useful to the user such as *Mayor, city,* and *Austin.*

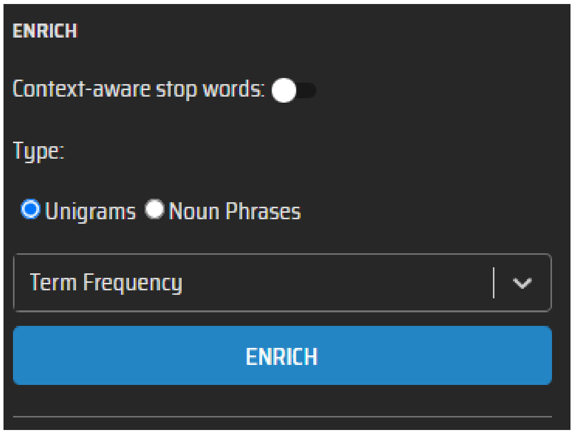
Usage

1. Select desired column


Studio
City of Austin Survey_with_query.csv
City of Austin Survey_with_query_enriched_2022(
SAVE
⤓CSV

Date	District	If there was ONE thing you could share with the Mayor regarding the City of Austin (any comment, suggestion, etc.), what would it be?
02/01/2016 12:00:00 AM	7	Dissatisfied traffic and with traffic, timing of street lights. extremely dissatisfied with cit govt. interfering in local businesses (uber/lyft, income property owners). also, extremely dissatisfied with all the free handouts to people who are perfectly capable of earning their own money. i'm very dissatisfied with the liberal leaning local politicians.

2. Select parameters based on the desired output



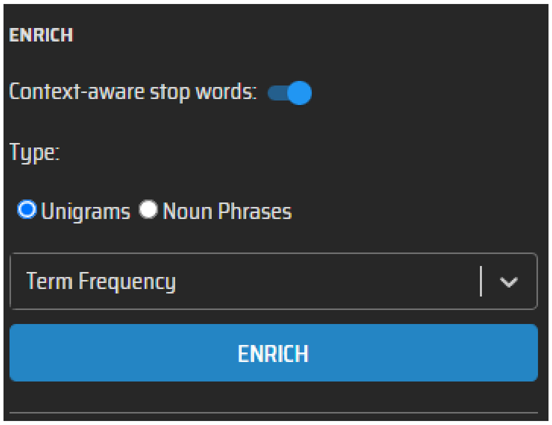
Context-aware stop words: ○

Type:

● Unigrams ○ Noun Phrases

Term Frequency

Example settings: ***Unigrams*** and **disabled** *context-aware stop words*



Context-aware stop words: ●

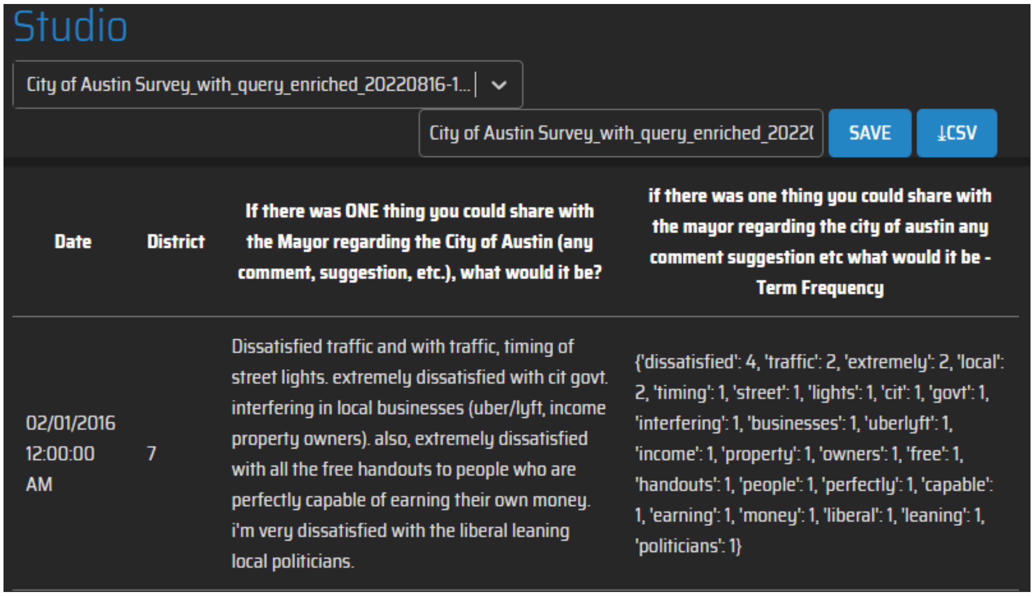
Type:

● Unigrams ○ Noun Phrases

Term Frequency

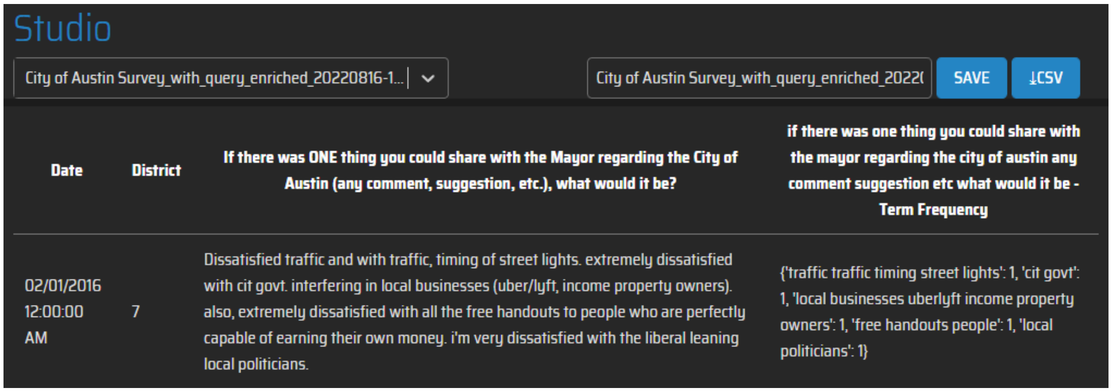
Example settings: ***Unigrams*** and **enabled** *context-aware stop words*

3. Click the *Enrich* button.


Studio
City of Austin Survey_with_query_enriched_20220816-1...
City of Austin Survey_with_query_enriched_2022(
SAVE
↓CSV

Date	District	If there was ONE thing you could share with the Mayor regarding the City of Austin (any comment, suggestion, etc.), what would it be?	if there was one thing you could share with the mayor regarding the city of austin any comment suggestion etc what would it be - Term Frequency
02/01/2016 12:00:00 AM	7	Dissatisfied traffic and with traffic, timing of street lights. extremely dissatisfied with cit govt. interfering in local businesses (uber/lyft, income property owners). also, extremely dissatisfied with all the free handouts to people who are perfectly capable of earning their own money. i'm very dissatisfied with the liberal leaning local politicians.	{'dissatisfied': 4, 'traffic': 2, 'extremely': 2, 'local': 2, 'timing': 1, 'street': 1, 'lights': 1, 'cit': 1, 'govt': 1, 'interfering': 1, 'businesses': 1, 'uberlyft': 1, 'income': 1, 'property': 1, 'owners': 1, 'free': 1, 'handouts': 1, 'people': 1, 'perfectly': 1, 'capable': 1, 'earning': 1, 'money': 1, 'liberal': 1, 'leaning': 1, 'politicians': 1}

Example output of ***Unigrams*** checked


Studio
City of Austin Survey_with_query_enriched_20220816-1...
City of Austin Survey_with_query_enriched_2022(
SAVE
↓CSV

Date	District	If there was ONE thing you could share with the Mayor regarding the City of Austin (any comment, suggestion, etc.), what would it be?	if there was one thing you could share with the mayor regarding the city of austin any comment suggestion etc what would it be - Term Frequency
02/01/2016 12:00:00 AM	7	Dissatisfied traffic and with traffic, timing of street lights. extremely dissatisfied with cit govt. interfering in local businesses (uber/lyft, income property owners). also, extremely dissatisfied with all the free handouts to people who are perfectly capable of earning their own money. i'm very dissatisfied with the liberal leaning local politicians.	{'traffic traffic timing street lights': 1, 'cit govt': 1, 'local businesses uberlyft income property owners': 1, 'free handouts people': 1, 'local politicians': 1}

Example output of ***Noun Phrases*** checked

To visualize the term frequency data. Please see the Widgets section below.

Text Summary

The *Text Summary* enricher summarizes text, cutting down the original text by about 60% while still preserving its most salient ideas.

To use it, select the column that contains the text to summarize, select the enricher, and click the Enrich button. In cases where you’re trying to summarize a document with various distinct sections, consider having each section as a separate row in the CSV in order to preserve the meaning of each, control which sections need to be summarized vs. which ones don’t, and to prevent the merging of sections in the resulting summary.

JSON to CSV

The *Json to CSV* enricher aims at flattening a JSON file consisting of nested features at different levels into a CSV file with a tabular format where all the attributes of the JSON file are converted to different columns.

The enricher begins by flattening one attribute at a time. Whenever it encounters a nested JSON feature (parent), the nested attribute (child) is flattened and added as a new column to the root table. The parent attribute's name is then appended as a prefix to this new attribute. This prefix addition is applied to any attribute at any level of nesting, including cases where the parent attribute itself is nested.

Note: A tilde (~) is used to separate the parent attribute name from the nested attribute names during prefixing. Hence, it is imperative to make sure that the existing features in the input JSON file don’t use ‘~’ in their feature names.

To use the enricher, simply select the enricher and click on the ‘**Enrich’** button to apply it.

Issues with some JSON files: Not all JSON files are created equal. If the platform fails to convert a JSON file, please contact us and we will assist you.

Custom enrichers

Creating a custom enricher

You can create your own enrichers using your own data and choosing to either train a specific base model or use the AutoML feature to let the platform choose the best algorithm for you automatically. The following methods are available for training a custom model:

Method name	**Description**
Category (M1-M2)	Use this method to categorize data. Both features and target variable must exhibit categorical values. The M2 version transforms empty values, while M1 does not.
Forecasting	Use this method to create predictions that output a quantifiable value, like revenue, temperature, distance, and time for example.
Text Classification (M1-M5)	Use this method to classify text using two or more classes. M1: optimized for unbalanced datasets. M2: optimized for balanced datasets for binary classification and reducing false negatives.

	M3: unoptimized and uses a different algorithm from M1 and M2 methods. M4: optimized for social media content and speed. M5: optimized for social media content and accuracy but slower speed than M4.
AutoML	Use this method to make predictions on tabular data and give control to the platform in choosing the best performing machine learning algorithm to make binary or multiclass classifications on the target variable. The platform tries Gradient Boosting, Instance-Based Learning, Deep Learning, and Ensemble Learning methods. AutoML accepts numerical, Boolean, categorical or textual data.

Creating a custom enricher via the Studio:

To train a model like Text Classification M2, follow these steps:

1. Load the training dataset. **Important:** the training dataset must only contain two columns: the column containing the text and the column containing the class.
2. In the Studio, select the column that you will want the model to predict (the column containing the class)
3. Select the base model (Text Classification or Text Classification M2)
4. Name your model
5. Click the "Save Model" button

Creating a custom enricher via the Models screen:

1. Upload a dataset that will be used to train and test the model.
2. Go to the Models screen and select the options in the inspector to train the model.

   Note: when using any of the text classification methods, you will be asked to select the input data that will be used to test the trained model in the dropdown labelled *“Select column for test analysis”*.
3. Click the “Create Model” button.

At this point, the platform will begin training the model. Once the training is complete, a “Trained” label will be shown next to the model. At this point, you must reload the screen to trigger the testing of the newly trained model. Once tested, the platform will provide performance metrics in the Inspector section when selecting the model.

Model statuses

Status	**Description**
Initialized	Initial state

Training	Model in training
Trained	Model training complete
Testing	Trained model being verified
Failed	Model training failed

Training a model using AutoML

The AutoML feature on the platform is a step towards automation that lets the platform choose the algorithm or model to be trained for tabular classification tasks. This feature enables the user to upload tabular datasets consisting of different data types. The feature supports the following data types - Numeric, Boolean, Categorical or Text. The platform can ingest the data while making sure to transform it into a format appropriate for the Machine Learning algorithms to process the data.

Note: This feature is available only through the Models screen.

Creating a custom AutoML enricher via the Models screen:

1. Upload a dataset that will be used to train and test the model.

   > **NOTE:** The training data should exclude columns/features which are irrelevant such as unique identifiers, time-related features, features that won't be available at the time of inference, etc.

2. Select the AutoML option from the model drop down Menu.
3. Choose the target variable that you want to train your model on.
   a. The platform automatically detects whether the task is one of binary or multiclass classification based on the number of classes in the target variable.
4. In the Performance requirement drop-down menu, choose the appropriate target performance metric to optimize the model based on your preference:
   a. Fewer False Negatives -> Optimizes the ML models for better Recall scores.
   b. Fewer False Positives -> Optimizes the ML models for better Precision scores.
   c. Default Metric (Balanced) -> Optimizes the ML models for better F1 scores.
5. Click the 'Create Model' button.

The platform will begin training multiple ML models and choose the best performing on to serve for inferencing. Once the training is complete, the platform will test the newly created model and provide performance metrics for the best performing model in the Inspector section of the Models screen.

You will also be able to download the feature importance rankings for all the features that were considered important during training by clicking the **'Download Train Features'** button. Upon clicking the download button, a zip file containing two files will get downloaded:

- train_features.txt - This file provides the list of features that were fed to the AutoML training after performing preprocessing.
- Optuna_MODEL_NAME_feature_importance_rankings.csv - This file provides a list of top-ranked features considered important by the best-performing model (MODEL_NAME is placeholder for the best performing model's name) post training.

Note: The performance of the AutoML model is correlated to the usability of the datasets that are being fed for training. Datasets involving complex datatypes or unclean data might require tailor-made preprocessing functions to better handle the data and achieve higher accuracy.

Any model trained as part of AutoML will also be available for retraining in the Ongoing Learning module.

Once the model is trained, you will be able to use it through the web interface, or the API. To access the model via API, follow these steps:

1. Query /studio/get_local_enrichment_methods/ to get a list of custom models, locate the one you're looking for, and copy its id.

Ongoing learning

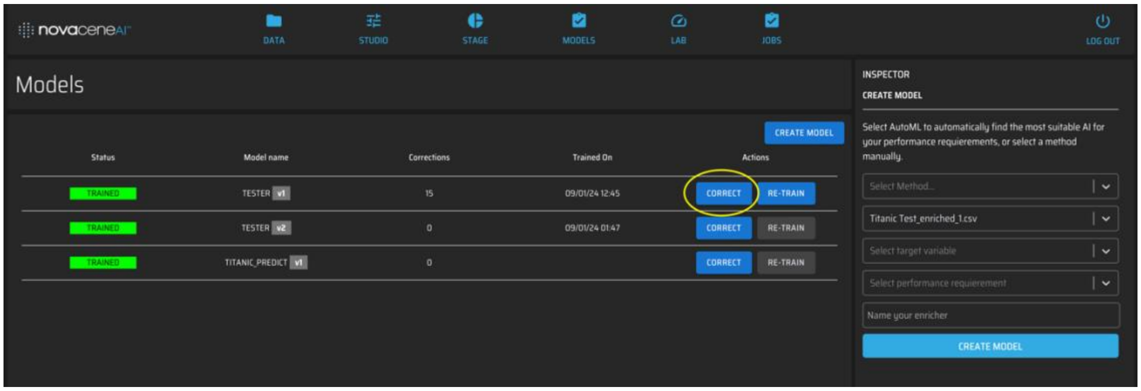
The platform enables users to provide feedback on the performance of custom enrichers and uses this feedback to improve the enrichers over time. This functionality is done by providing corrections to misclassifications and using those corrections to retrain new versions of a custom model.

Correcting misclassifications

The platform allows analysts to "correct" the mispredictions that the platform might make. The platform will then learn from these corrections and keep its AI trained with the latest data.

> **NOTE:** You must first generate predictions using the model before you are able to correct predictions. If you try to access the corrections screen before generating predictions, you will see an empty list, since no predictions have yet been generated.

To make corrections, go to the Models screen, locate the model you'd like to make corrections for, click the "Correct" button, and follow the instructions below.



Models

Status	Model name	Corrections	Trained On	Actions
TRAINED	TESTER v1	15	09/01/24 12:45	CORRECT RE-TRAIN
TRAINED	TESTER v2	0	09/01/24 01:47	CORRECT RE-TRAIN
TRAINED	TITANIC_PREDICT v1	0		CORRECT RE-TRAIN

INSPECTOR

CREATE MODEL

Select AutoML to automatically find the most suitable AI for your performance requierements, or select a method manually.

Select Method...

Titanic Test_enriched_1.csv

Select target variable

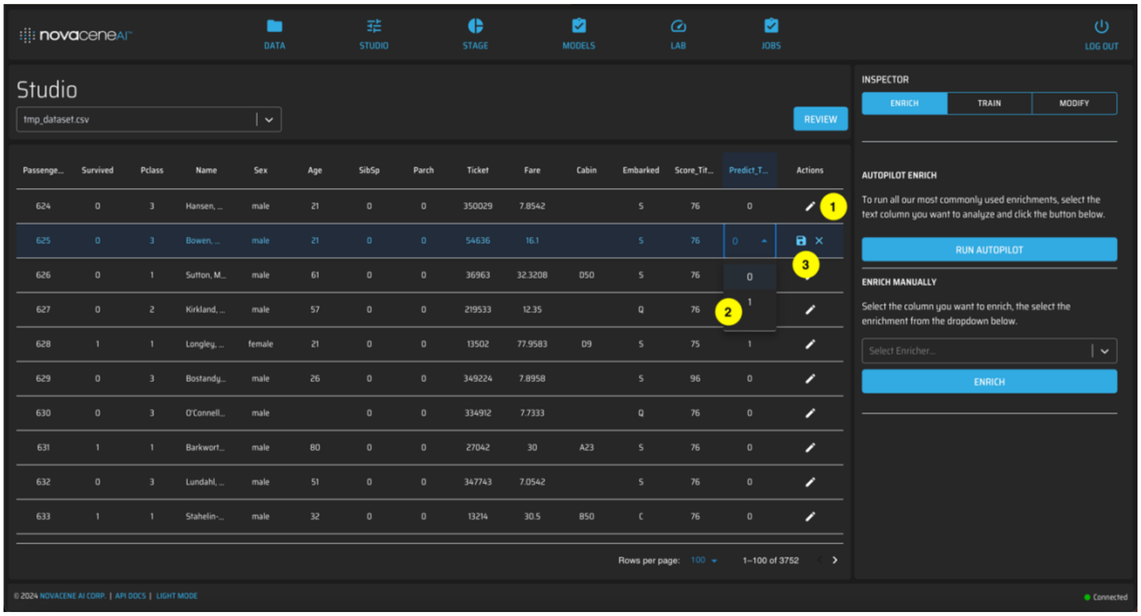
Select performance requierement

Name your enricher

Figure 9: Accessing the corrections interface

1. **Making corrections:** You will be taken to the Studio, where you will be able to correct mispredictions. To do so, locate the row you'd like to correct, (1) click the pencil icon, (2) select the correct value from the drop-down menu, and (3) click the save icon. Repeat these steps for each correction you'd like to make.

> **NOTE:** You must first generate predictions using the model in order to be able to make corrections. If you try to access the corrections screen before generating predictions, you will see an empty list, since no predictions have yet been generated.



Studio

tmp_dataset.csv

Passenge...	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked	Score_Tit...	Predict_T...	Actions
624	0	3	Hansen, ...	male	21	0	0	350029	7.8542		S	76	0	✎
625	0	3	Bowen, ...	male	21	0	0	54636	16.1		S	76	0	🖫 ×
626	0	1	Sutton, M...	male	61	0	0	36963	32.3208	D50	S	76	0	
627	0	2	Kirkland, ...	male	57	0	0	219533	12.35		Q	76	1	✎
628	1	1	Longley, ...	female	21	0	0	13502	77.9583	D9	S	75	1	✎
629	0	3	Bostandy...	male	26	0	0	349224	7.8958		S	96	0	✎
630	0	3	O'Connell...	male		0	0	334912	7.7333		Q	76	0	✎
631	1	1	Barkwort...	male	80	0	0	27042	30	A23	S	76	0	✎
632	0	3	Lundahl, ...	male	51	0	0	347743	7.0542		S	76	0	✎
633	1	1	Stahelin-...	male	32	0	0	13214	30.5	B50	C	76	0	✎

Rows per page: 100 1–100 of 3752

AUTOPILOT ENRICH

To run all our most commonly used enrichments, select the text column you want to analyze and click the button below.

ENRICH MANUALLY

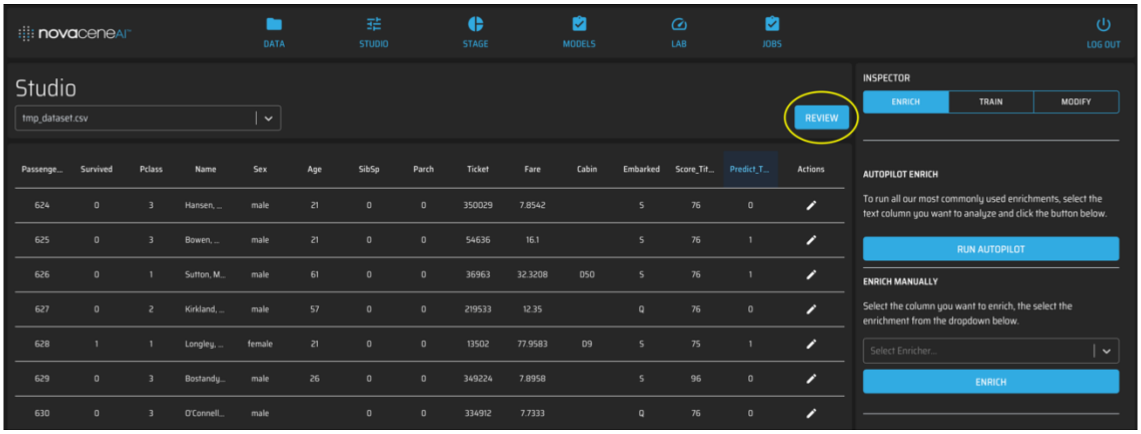
Select the column you want to enrich, the select the enrichment from the dropdown below.

Select Enricher...

© 2024 NOVACENE AI CORP. | API DOCS | LIGHT MODE

Figure 10: Correcting mispredictions

2. **Reviewing corrections:** Once you're done making corrections, you will need to merge those corrections back into the datasets that will be used to train a new version of the model. To do so, click the "Review" button to review your changes, and resolve any conflicts that may have been introduced during the correction process.



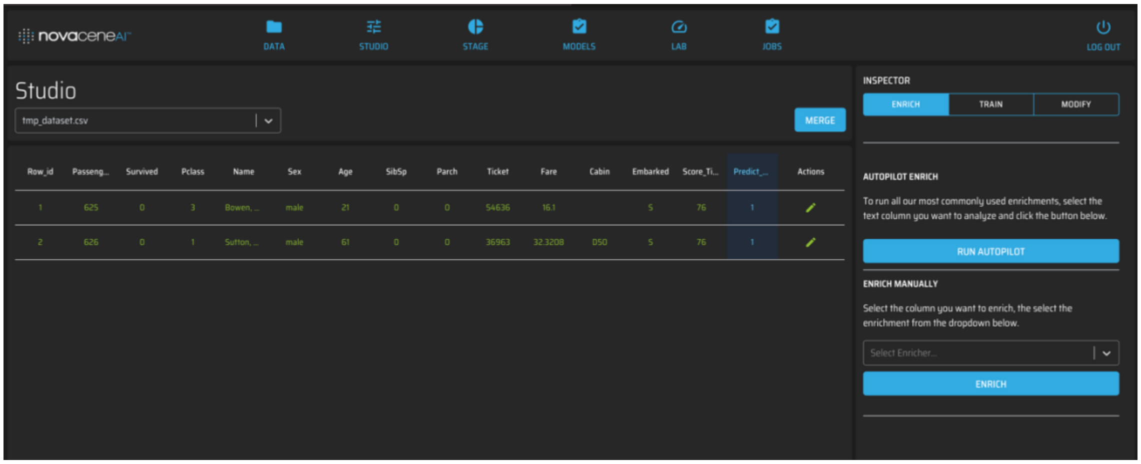
Passenge...	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked	Score_Tit...	Predict_T...	Actions
624	0	3	Hansen, ...	male	21	0	0	350029	7.8542		S	76	0	✎
625	0	3	Bowen, ...	male	21	0	0	54636	16.1		S	76	1	✎
626	0	1	Sutton, M...	male	61	0	0	36963	32.3208	D50	S	76	1	✎
627	0	2	Kirkland, ...	male	57	0	0	219533	12.35		Q	76	0	✎
628	1	1	Longley, ...	female	21	0	0	13502	77.9583	D9	S	75	1	✎
629	0	3	Bostandy...	male	26	0	0	349224	7.8958		S	96	0	✎
630	0	3	O'Connell...	male		0	0	334912	7.7333		Q	76	0	✎

To run all our most commonly used enrichments, select the text column you want to analyze and click the button below.

Select the column you want to enrich, the select the enrichment from the dropdown below.

Figure 11: Accessing the Review interface

3. **Resolving conflicts:** The Review screen will show you the rows that you have corrected. The rows with green text indicate changes that can be merged without issues. The rows with red text show the rows where the corrections contradict existing identical cases. Change any values you would like to update and proceed to merge your changes.



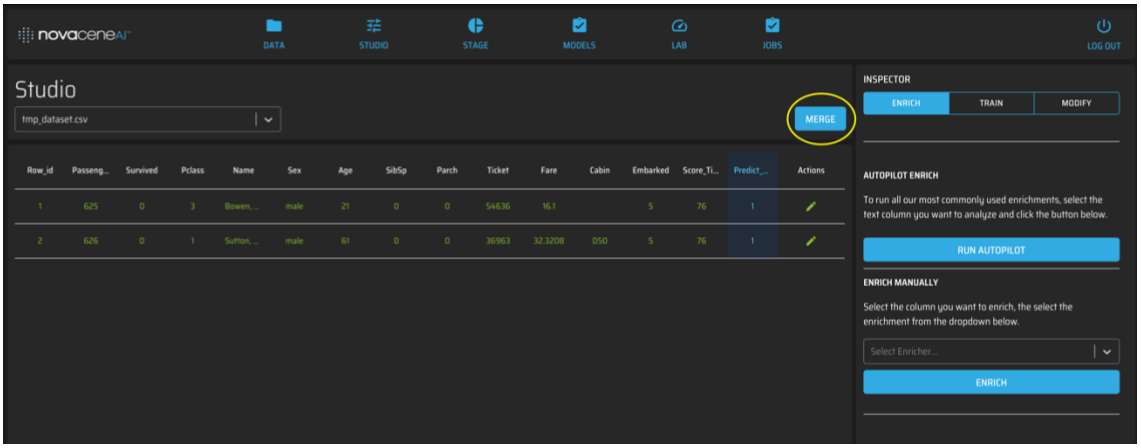
Row_id	Passeng...	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked	Score_Ti...	Predict_...	Actions
1	625	0	3	Bowen, ...	male	21	0	0	54636	16.1		S	76	1	✎
2	626	0	1	Sutton, ...	male	61	0	0	36963	32.3208	D50	S	76	1	✎

To run all our most commonly used enrichments, select the text column you want to analyze and click the button below.

Select the column you want to enrich, the select the enrichment from the dropdown below.

Figure 12: Reviewing changes

4. **Merging changes:** Once you're happy with all the changes, click the "Merge" button to merge those changes into the datasets that will be used by the retraining process.



Studio

tmp_dataset.csv

Row_id	Passeng...	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked	Score_Ti...	Predict_...	Actions
1	625	0	3	Bowen, ...	male	21	0	0	54636	16.1		S	76	1	✎
2	626	0	1	Sutton, ...	male	61	0	0	36963	32.3208	D50	S	76	1	✎

AUTOPILOT ENRICH

To run all our most commonly used enrichments, select the text column you want to analyze and click the button below.

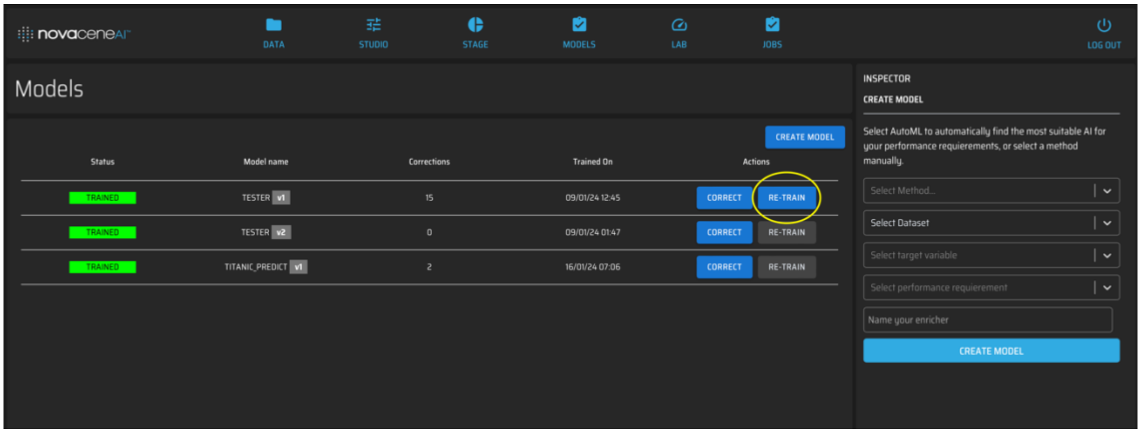
ENRICH MANUALLY

Select the column you want to enrich, the select the enrichment from the dropdown below.

Figure 13: Merging changes

Retraining enrichers

Once enough corrections have been accumulated, the "Re-Train" button will become active. Clicking the button will trigger the training of a new version of the model. Once the new version is trained, you will be able to compare the performance of the new model against that of the old one and choose whether to start using the new version or the previous one.



Models

Status	Model name	Corrections	Trained On	Actions
TRAINED	TESTER v1	15	09/01/24 12:45	CORRECT RE-TRAIN
TRAINED	TESTER v2	0	09/01/24 01:47	CORRECT RE-TRAIN
TRAINED	TITANIC_PREDICT v1	2	16/01/24 07:06	CORRECT RE-TRAIN

CREATE MODEL

Select AutoML to automatically find the most suitable AI for your performance requierements, or select a method manually.

Figure 14: Triggering a retraining job

Deleting custom enrichers

Depending on where you created the custom enricher, you'll need to follow different set of instructions to delete it.

1. **Enrichers created via the Studio screen:**

    a. Open your browser and go to your Novacene instance, /admin/ (e.g., my.novacene.ai/admin/)
    b. Go to DATASETS > **Enrichment methods**
    c. Select the model(s) you want to delete
    d. Select the option to "Delete..." from the dropdown and click the "Go" button
    e. Log in to the server where you deleted the model
    f. Cd to /opt/web_app/{env}/data/MLmodels/
    g. Remove the corresponding .sav files

2. **Enrichers created via the Models screen:**
    a. Go to https://{subdomain}.novacene.ai/admin/models/trainingmodelmanager/
    b. Select the model(s) and in the “Action” menu, select “Delete selected model(s) and associated master datasets”, and click the “Go” button.
    c. Verify that that deleted models no longer appear on the Models list.

Deleting custom enrichers using the API

You can delete custom models via the API as well.

Measuring enricher accuracy

The NovaceneAI Platform allows users to test the accuracy of their models using a confusion matrix. Confusion matrices are available for binary and multiclass classification models.

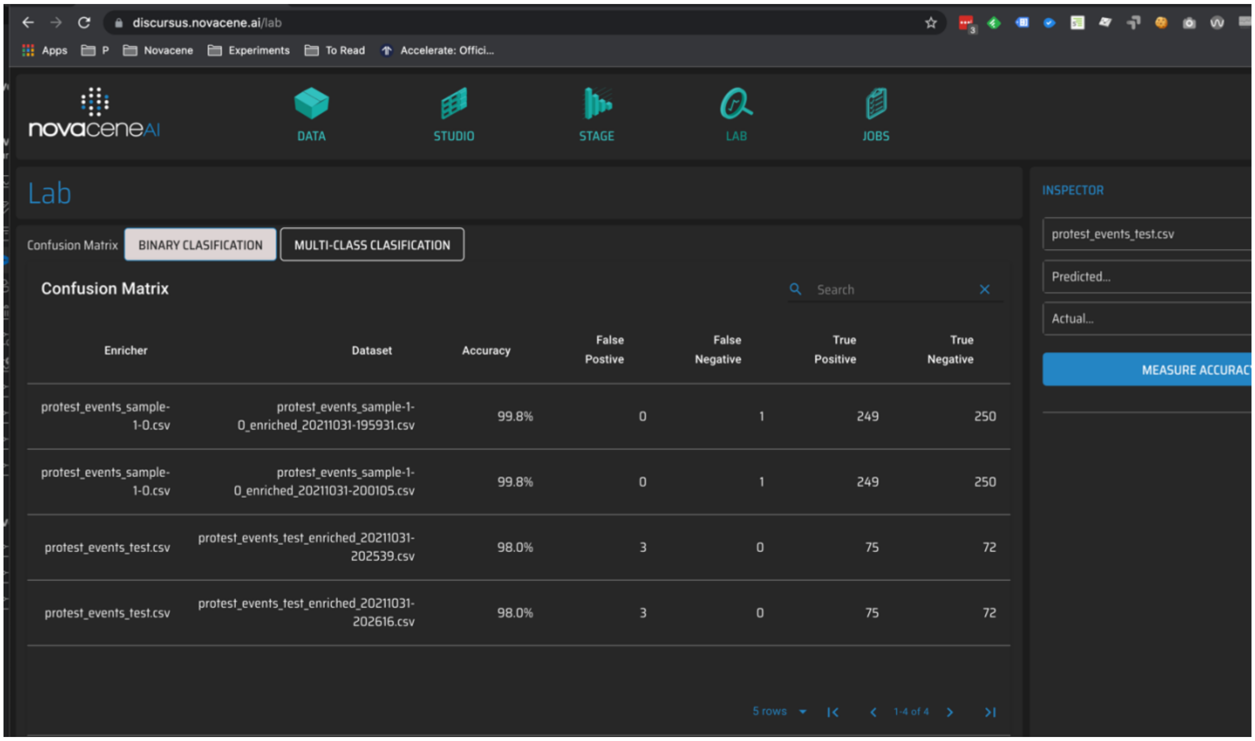
Generating a confusion matrix

Please contact Novacene for instructions.

Deleting a confusion matrix

To delete all confusion matrices (both binary and multiclass):

1. Go to /admin/datasets/dataset/
2. Select the datasets:
    a. *models_confusion_matrix.csv*
    b. models_classification_report.csv
3. In the dropdown next to “Action:”, select “Delete selected data sets” and click the “Go” button
4. Confirm that you wish to delete the datasets
5. All the confusion matrices will be deleted



Lab

Confusion Matrix BINARY CLASIFICATION MULTI-CLASS CLASIFICATION

Confusion Matrix

Enricher	Dataset	Accuracy	False Postive	False Negative	True Positive	True Negative
protest_events_sample-1-0.csv	protest_events_sample-1-0_enriched_20211031-195931.csv	99.8%	0	1	249	250
protest_events_sample-1-0.csv	protest_events_sample-1-0_enriched_20211031-200105.csv	99.8%	0	1	249	250
protest_events_test.csv	protest_events_test_enriched_20211031-202539.csv	98.0%	3	0	75	72
protest_events_test.csv	protest_events_test_enriched_20211031-202616.csv	98.0%	3	0	75	72

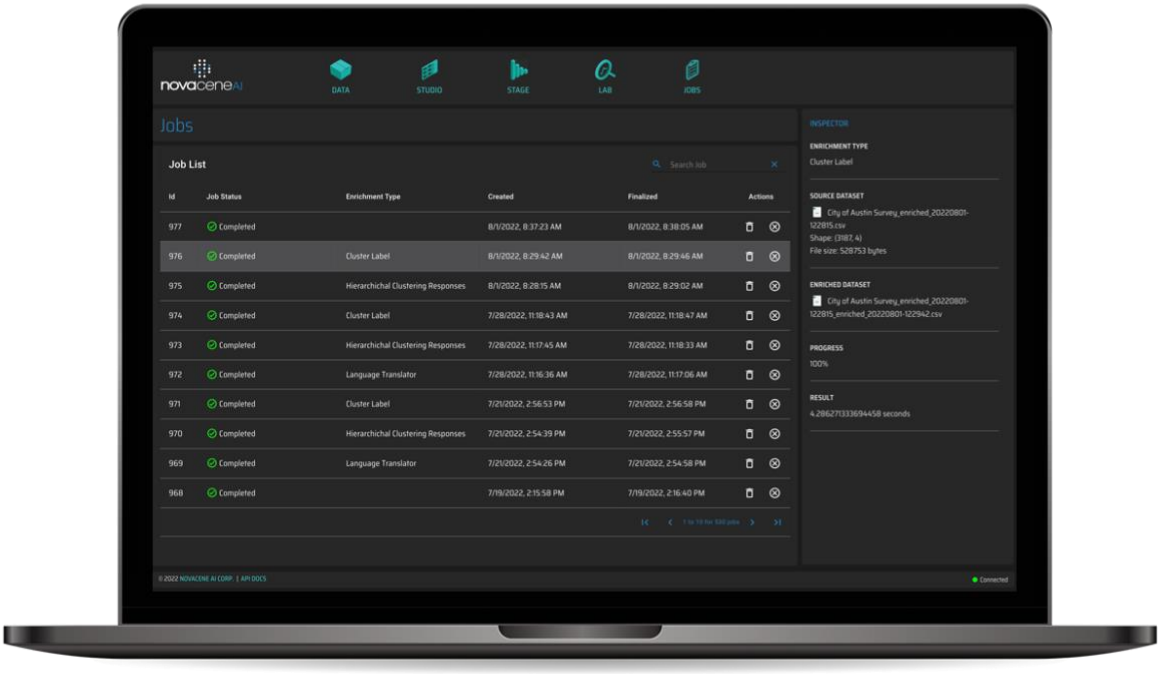
protest_events_test.csv

Predicted...

Actual...

Monitoring enrichments

The application allows you to monitor the progress of both enrichments and training jobs. You can monitor both types of jobs on the Jobs screen.



Jobs

Job List

Id	Job Status	Enrichment Type	Created	Finalized	Actions
977	✓ Completed		8/1/2022, 8:37:23 AM	8/1/2022, 8:38:05 AM	🗑 ⊗
976	✓ Completed	Cluster Label	8/1/2022, 8:29:42 AM	8/1/2022, 8:29:46 AM	🗑 ⊗
975	✓ Completed	Hierarchichal Clustering Responses	8/1/2022, 8:28:15 AM	8/1/2022, 8:29:02 AM	🗑 ⊗
974	✓ Completed	Cluster Label	7/28/2022, 11:18:43 AM	7/28/2022, 11:18:47 AM	🗑 ⊗
973	✓ Completed	Hierarchichal Clustering Responses	7/28/2022, 11:17:45 AM	7/28/2022, 11:18:33 AM	🗑 ⊗
972	✓ Completed	Language Translator	7/28/2022, 11:16:36 AM	7/28/2022, 11:17:06 AM	🗑 ⊗
971	✓ Completed	Cluster Label	7/21/2022, 2:56:53 PM	7/21/2022, 2:56:58 PM	🗑 ⊗
970	✓ Completed	Hierarchichal Clustering Responses	7/21/2022, 2:54:39 PM	7/21/2022, 2:55:57 PM	🗑 ⊗
969	✓ Completed	Language Translator	7/21/2022, 2:54:26 PM	7/21/2022, 2:54:58 PM	🗑 ⊗
968	✓ Completed		7/19/2022, 2:15:58 PM	7/19/2022, 2:16:40 PM	🗑 ⊗

INSPECTOR

ENRICHMENT TYPE
Cluster Label

SOURCE DATASET
City of Austin Survey_enriched_20220801-122815.csv
Shape: (3187, 4)
File size: 528753 bytes

ENRICHED DATASET
City of Austin Survey_enriched_20220801-122815_enriched_20220801-122942.csv

PROGRESS
100%

RESULT
4.286271333694458 seconds

Monitoring progress of enrichment jobs

In addition to being able to monitor enrichment jobs on the Jobs screen, you can also see the progress on the Datasets screen. To see the progress, locate the dataset being enriched. You will see a progress wheel overlaid on the dataset. This progress wheel will show progress from 0 to 100, denoting the percentage of completion of enrichment.

FAQs about monitoring

Why some enrichment progress indicators jump from 0 to 100 in one step? There are two cases in which the progress of an enricher cannot be quantified and therefore the progress remains at 0% and will jump to 100% as soon the job is completed. The cases are:

1) Enrichments that require the processing of the entire dataset at once (such as Clustering enrichers). In this case, the enrichment job must be done all at once.
2) Enrichers that add rows to a dataset (such as Sentence Segmentation or Clause Extraction). In this case, the added rows affect the progress shown because the progress is based on the dataset’s size.

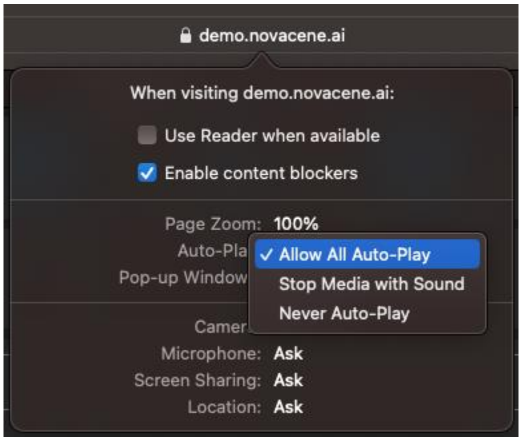
Why do progress indicators sometime update frequently and sometimes in larger chunks? Because datasets are divided in chunks of 100 records to be processed at a time. In the case of this dataset of 400 rows, the update frequency will be every 25% (400 rows / 100 chunks = 4 intervals = 25%)

What do the different job states mean? The table below provides descriptions for each job state:

State name	Description
UPLOADED	Dataset uploaded
INITIALIZED	Job received by server
ENQUEUED	Job is in queue to be run
RUNNING	Job is currently being processed
COMPLETED	Job completed successfully
FAILED	Job failed
CANCELLED	Job cancelled by user

Notification of completed jobs

When a job finish running, you will hear a notification. This is intended to draw your attention on the finished job in case you are away from the NovaceneAI Platform when this happens. Note that to hear the notification some browsers may require that you allow for this to happen.



demo.novacene.ai

When visiting demo.novacene.ai:

□ Use Reader when available

☑ Enable content blockers

Page Zoom: 100%

Auto-Pla ✓ Allow All Auto-Play

Pop-up Window Stop Media with Sound

Never Auto-Play

Camer

Microphone: Ask

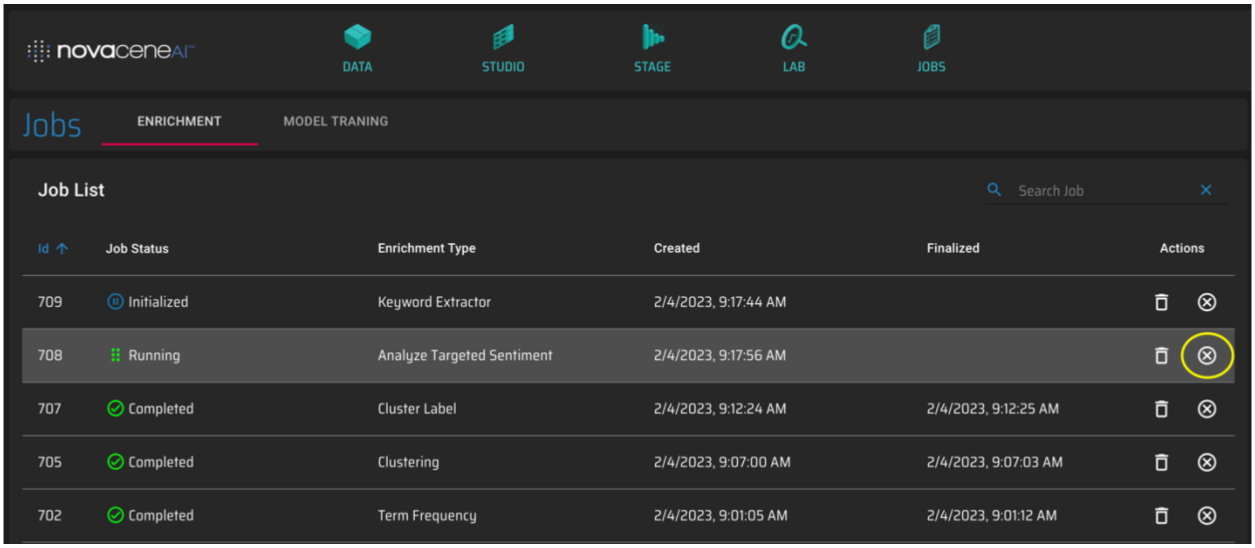
Screen Sharing: Ask

Location: Ask

Figure 15: Some browsers require that you set permissions for notifications to be heard

Cancelling jobs

To cancel a job, click the Cancel icon corresponding to the running job.



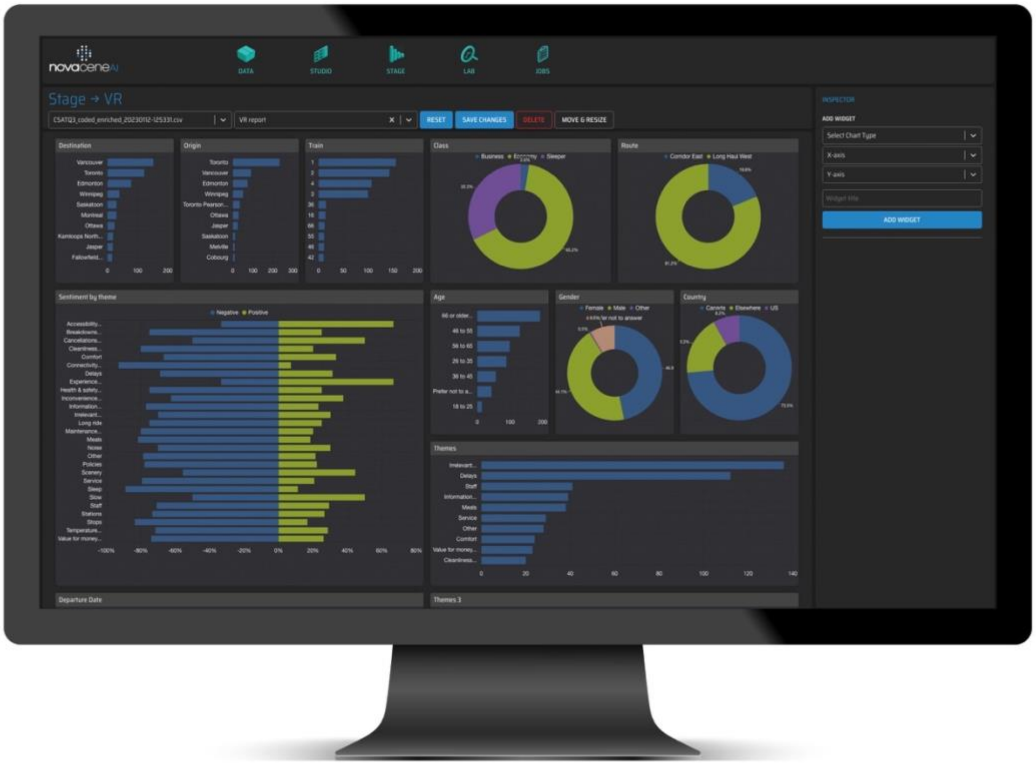
Job List

Id ↑	Job Status	Enrichment Type	Created	Finalized	Actions
709	Initialized	Keyword Extractor	2/4/2023, 9:17:44 AM		🗑 ⊗
708	Running	Analyze Targeted Sentiment	2/4/2023, 9:17:56 AM		🗑 ⊗
707	Completed	Cluster Label	2/4/2023, 9:12:24 AM	2/4/2023, 9:12:25 AM	🗑 ⊗
705	Completed	Clustering	2/4/2023, 9:07:00 AM	2/4/2023, 9:07:03 AM	🗑 ⊗
702	Completed	Term Frequency	2/4/2023, 9:01:05 AM	2/4/2023, 9:01:12 AM	🗑 ⊗

Why does a cancelled job continue to run, or re-appears as Running on the Jobs screen? Some enrichers cannot be cancelled. You can either let the enricher finish, or contact us and we will cancel the job for you.

Data visualization

Use the *Stage* screen to visualize your data.



In this section:

1. Stage overview
2. Auto-visualization
3. Creating reports
4. Widgets

Stage overview

The Stage screen enables you to create reports comprised of data visualization widgets. You can add as many widgets as you want, size them, and lay them out in whichever arrangement makes the most sense to the analysis you are performing.

Auto-visualization

When you load a dataset in the Stage screen, the platform will automatically show widgets for each of the columns from your dataset that are likely to benefit from a visualization. The platform will automatically determine the type of visualization widget to use based on each column's data.

The platform will not produce automated visualizations for a given column when the column contains:

- a single identical value across all rows
- distinct values in every row
- only blank values

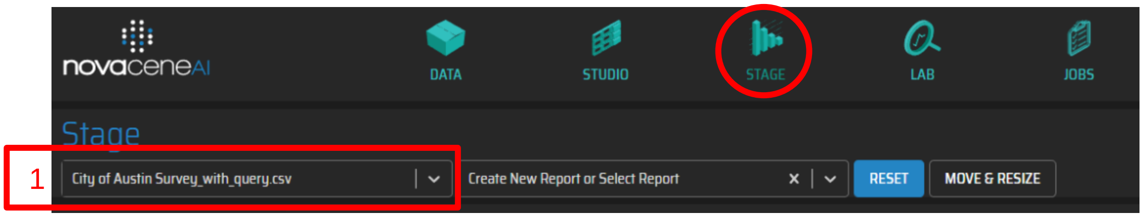
Creating reports

Reports are collections of widgets laid out in a specific arrangement. You can create as many reports as you want, and you can save them to retrieve them later. Saved reports will remember the layout, any applied filters, and the zoom and pan settings for each widget.

When you create a report, the report is associated with the dataset you used to create it.

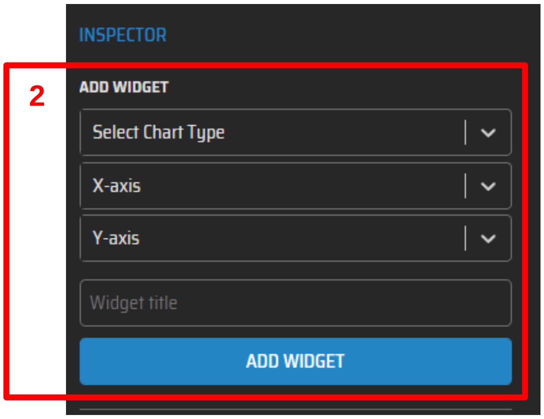
To create a report:

1. Load the dataset you would like to visualize by clicking the drop-down arrow.

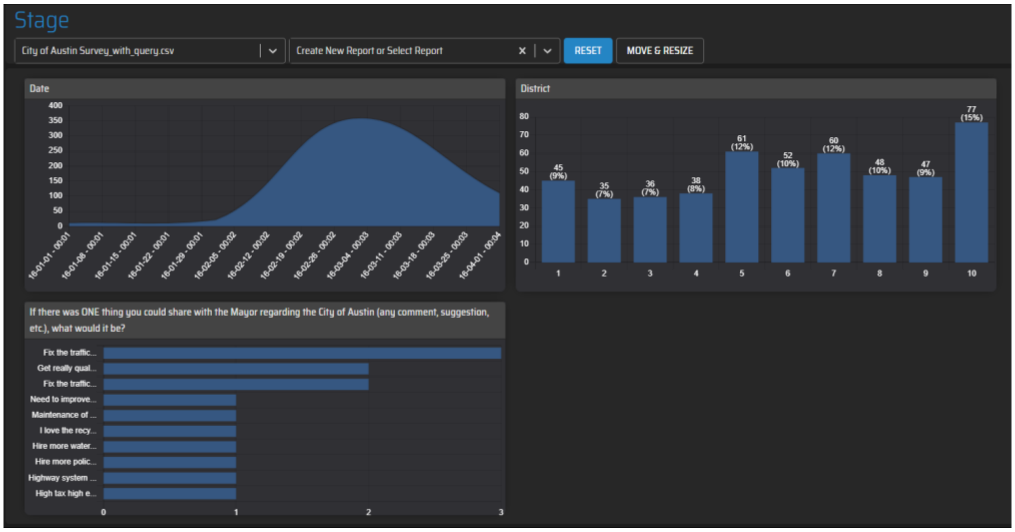


City of Austin Survey_with_query.csv

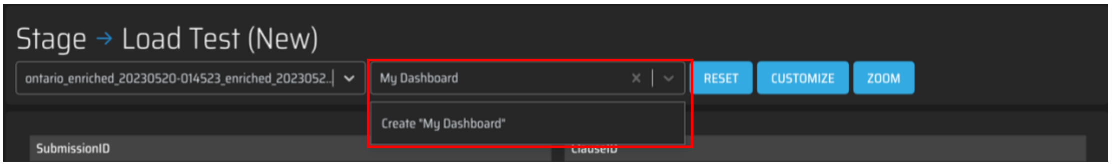
2. The application will automatically generate visualization widgets for your data. For more information, consult the Auto-visualization section.
3. Remove any widgets you don't need.
4. Add your own widgets as necessary. To add a widget:
    a. From the Inspector window on the right-hand side, use the *Select Chart Type* drop-down to choose a chart type for to use.



b. Subsequently, select the desired *X-axis Y-axis* and *Widget title*. Once the *Add Widget* button is clicked, the chart will be displayed in the main *Stage* area as shown below:



5. Customize the report to resize and re-arrange widgets in whatever layout makes the most sense for your analysis.
6. Use the “Create New Report or Select Report” drop down menu to name your report, and click the *“Create ‘My report name’”* option.



Widgets

Data visualization widgets are charts that help you visualize your data, so it is easy to understand insights and spot trends. Each widget provides a set of options, including Chart Type, X-axis and Y-axis, Group-by, and Widget title.

Chart types

Bar (Vertical)

A vertical bar chart.

Bar (Horizontal)

A horizontal bar chart.

Bar (Diverging Stacked)

A horizontal bar chart where the X axis represents a value range which neutral point is centered, and the Y axis represents a breakdown by which values are reported. For example, you can visualize the distribution of sentiment along the X-axis and provide a breakdown on the Y-axis (such as by topic). This chart allows you to gain insights into the distribution of sentiment across different topics. Positive opinions are represented on the right side of the X-axis and negative opinions on left. The center of the chart indicates neutral or balanced sentiment.

Note that the Group-by option cannot be applied to this chart.

Here are sample settings to create a chart:

x-axis: sentiment
y-axis: cluster label
grouped by: [leave empty]

Doughnut

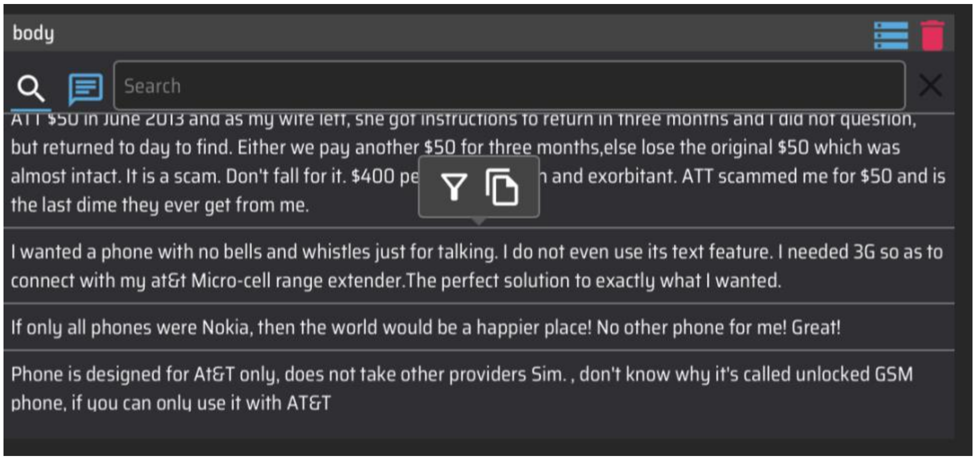
A circular chart showing the percentage of share of values in a range. Similar to a pie chart.

Line

A continuous line chart.

List

A list of items. This widget is a good choice for displaying text, or information you'd prefer to read. Items displayed in the list are searchable. The buttons that appear on hover shown in the image below allow you to copy text in the list widget, and filter the stage by a particular item in the list widget.



ATT $50 in June 2013 and as my wife left, she got instructions to return in three months and I did not question, but returned to day to find. Either we pay another $50 for three months,else lose the original $50 which was almost intact. It is a scam. Don't fall for it. $400 pe h and exorbitant. ATT scammed me for $50 and is the last dime they ever get from me.

I wanted a phone with no bells and whistles just for talking. I do not even use its text feature. I needed 3G so as to connect with my at&t Micro-cell range extender.The perfect solution to exactly what I wanted.

If only all phones were Nokia, then the world would be a happier place! No other phone for me! Great!

Phone is designed for At&T only, does not take other providers Sim. , don't know why it's called unlocked GSM phone, if you can only use it with AT&T

Additionally, questions can be asked against the items in the list using Generative AI. This function will use all the items displayed in the list as context for the Generative AI, along with your query.

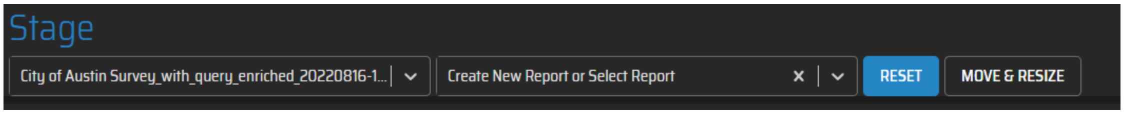
The stage will automatically create a List widget to display columns with text.

Word Cloud

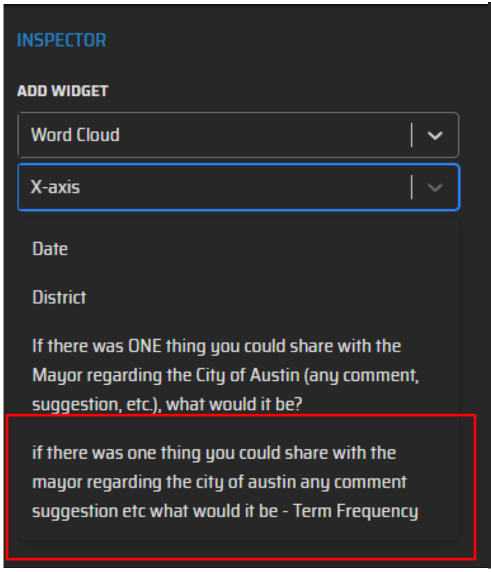
A Word Cloud is a group of words or phrases where the size of the items represents their frequency. The Word Cloud chart type requires the Term Frequency enricher to applied on the Studio page before the Word Cloud tool can be used.

The process below outlines the steps to use the *Word Cloud* visualization tool following the steps described above.

1. Go to the Stage screen
2. Select the dataset


Stage
City of Austin Survey_with_query_enriched_20220816-1...
Create New Report or Select Report
RESET
MOVE & RESIZE

3. Select the Chart type called Word Cloud. Please note that the selected column name includes *Term Frequency.*



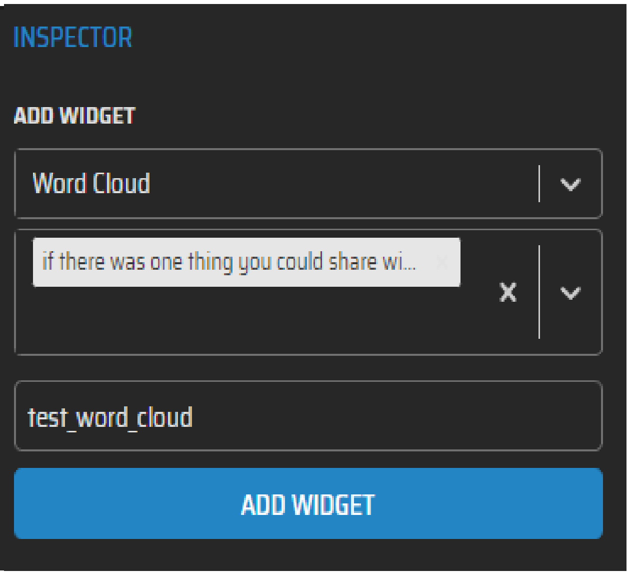
INSPECTOR

ADD WIDGET

Word Cloud

X-axis

Date
District
If there was ONE thing you could share with the Mayor regarding the City of Austin (any comment, suggestion, etc.), what would it be?
if there was one thing you could share with the mayor regarding the city of austin any comment suggestion etc what would it be - Term Frequency



INSPECTOR

ADD WIDGET

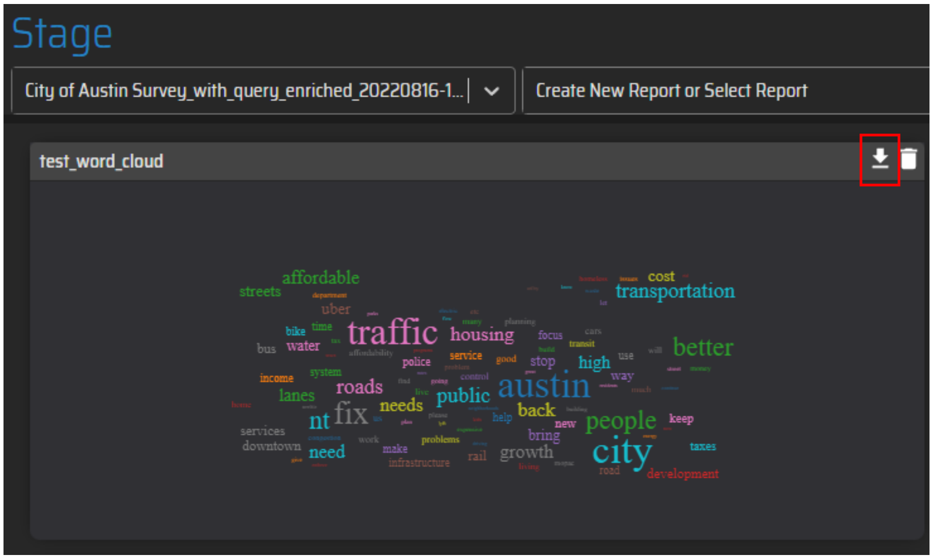
Word Cloud

if there was one thing you could share wi...

test_word_cloud

4. Press the *Add Widget* button.

5. The *Word Cloud* is displayed on the main *Stage* page.



Stage

City of Austin Survey_with_query_enriched_20220816-1...

Create New Report or Select Report

test_word_cloud

6. The image of the *Word Cloud* can be downloaded using the arrow button as shown in the picture above.

X-axis and Y-axis

Net Sentiment

Net Sentiment shows the overall sentiment score recorded over a period of time, broken down by theme.

To create a Net Sentiment visualization:

1. Select the “Line” chart type
2. For the X axis, select a column containing date or time data
3. For the Y axis, select the “Net Sentiment” option, and then the column that contains the Themes
4. Name your widget and click the “Add Widget” button

Group-by

Use the group-by option when you want values to be grouped by values of another column. For example, let’s say you have a dataset with 1,000 reviews. When you apply clause extraction, the platform will add rows your dataset to accommodate each clause it its own row. To preserve the relationship between a clause and the original submission, the platform adds a column called *SubmissionID.* Now let’s say you want to create a widget of type vertical bar where you can see the number of unique submissions by age range. You would then select your **Bar (Vertical)** chart type, ***Age Range*** for the X-axis, ***count*** for the Y-axis, and ***SubmissionID*** as the group-by option.

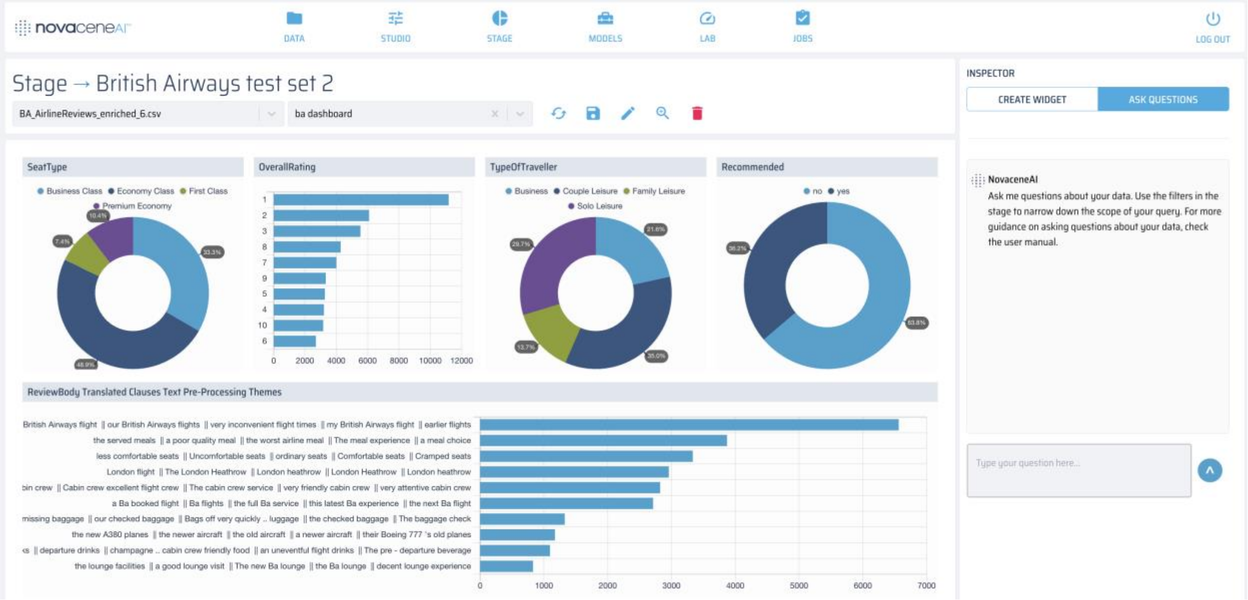
Widget title

The widget title option allows you to give your widget a name.

Chat with your Data

Using the “Ask Questions” Feature

We’ve added a new and improved feature that allows you to chat with your data. In the stage, you can now collaborate with generative AI to analyze your data.


novaceneAI
DATA
STUDIO
STAGE
MODELS
LAB
JOBS
LOG OUT
Stage → British Airways test set 2
BA_AirlineReviews_enriched_6.csv
ba dashboard
INSPECTOR
CREATE WIDGET
ASK QUESTIONS
SeatType
Business Class
Economy Class
First Class
Premium Economy
10.4%
7.4%
33.3%
48.9%
OverallRating
1
2
3
8
7
9
5
4
10
6
0
2000
4000
6000
8000
10000
12000
TypeOfTraveller
Business
Couple Leisure
Family Leisure
Solo Leisure
21.6%
29.7%
13.7%
35.0%
Recommended
no
yes
36.2%
63.8%
ReviewBody Translated Clauses Text Pre-Processing Themes
British Airways flight || our British Airways flights || very inconvenient flight times || my British Airways flight || earlier flights
the served meals || a poor quality meal || the worst airline meal || The meal experience || a meal choice
less comfortable seats || Uncomfortable seats || ordinary seats || Comfortable seats || Cramped seats
London flight || The London Heathrow || London heathrow || London Heathrow || London heathrow
bin crew || Cabin crew excellent flight crew || The cabin crew service || very friendly cabin crew || very attentive cabin crew
a Ba booked flight || Ba flights || the full Ba service || this latest Ba experience || the next Ba flight
missing baggage || our checked baggage || Bags off very quickly .. luggage || the checked baggage || The baggage check
the new A380 planes || the newer aircraft || the old aircraft || a newer aircraft || their Boeing 777 's old planes
ks || departure drinks || champagne .. cabin crew friendly food || an uneventful flight drinks || The pre - departure beverage
the lounge facilities || a good lounge visit || The new Ba lounge || the Ba lounge || decent lounge experience
0
1000
2000
3000
4000
5000
6000
7000
Type your question here...

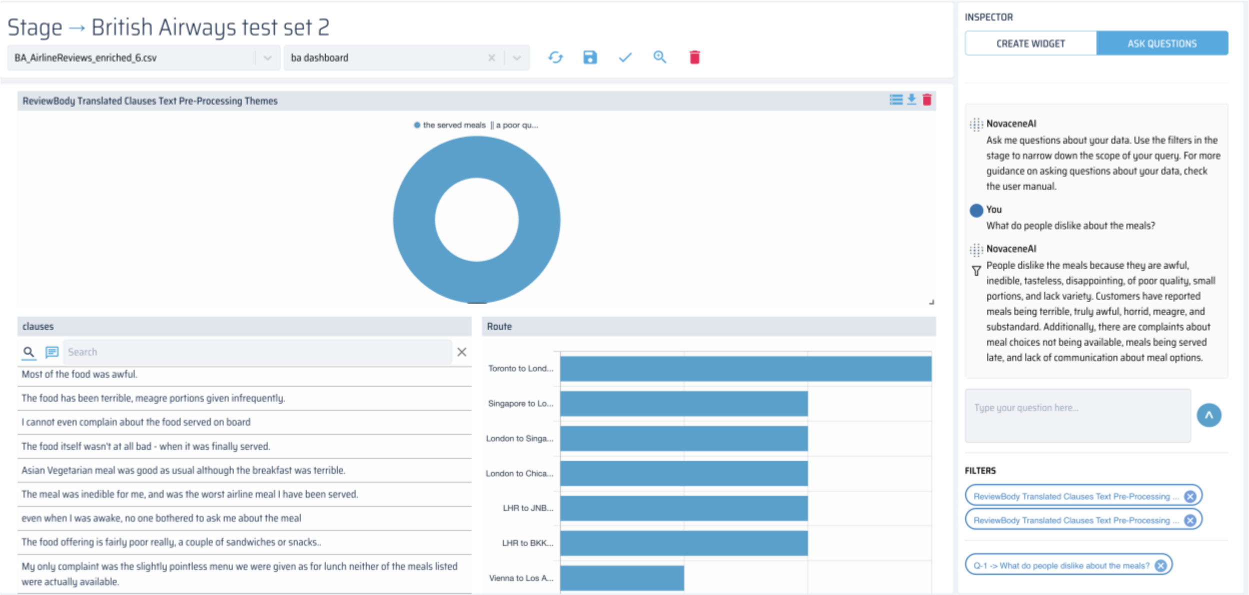
NovaceneAI

Ask me questions about your data. Use the filters in the stage to narrow down the scope of your query. For more guidance on asking questions about your data, check the user manual.

Now, when you ask a question about your data, you will be able to view exactly what pieces of text in your dataset were used to generate the answer. These pieces of your dataset are also visualized in the dashboard.

For example: You have a dataset of a survey about experiences on an airline. You're interested understanding how the meal experience can be improved. So, you click on the “Meal Experience” bar in the “Themes” chart, and then in the Sentiment doughnut chart, you click on “Negative” to filter your dashboard.

Now, we can ask in chat-- “What do reviewers dislike about the meals?”



Stage → British Airways test set 2

BA_AirlineReviews_enriched_6.csv
ba dashboard
INSPECTOR
CREATE WIDGET
ASK QUESTIONS
ReviewBody Translated Clauses Text Pre-Processing Themes
the served meals || a poor qu...
clauses
Search
Route
Toronto to Lond...
Singapore to Lo...
London to Singa...
London to Chica...
LHR to JNB...
LHR to BKK...
Vienna to Los A...
Type your question here...
FILTERS
ReviewBody Translated Clauses Text Pre-Processing ...
ReviewBody Translated Clauses Text Pre-Processing ...
Q-1 -> What do people dislike about the meals?

Most of the food was awful.

The food has been terrible, meagre portions given infrequently.

I cannot even complain about the food served on board

The food itself wasn't at all bad - when it was finally served.

Asian Vegetarian meal was good as usual although the breakfast was terrible.

The meal was inedible for me, and was the worst airline meal I have been served.

even when I was awake, no one bothered to ask me about the meal

The food offering is fairly poor really, a couple of sandwiches or snacks..

My only complaint was the slightly pointless menu we were given as for lunch neither of the meals listed were actually available.

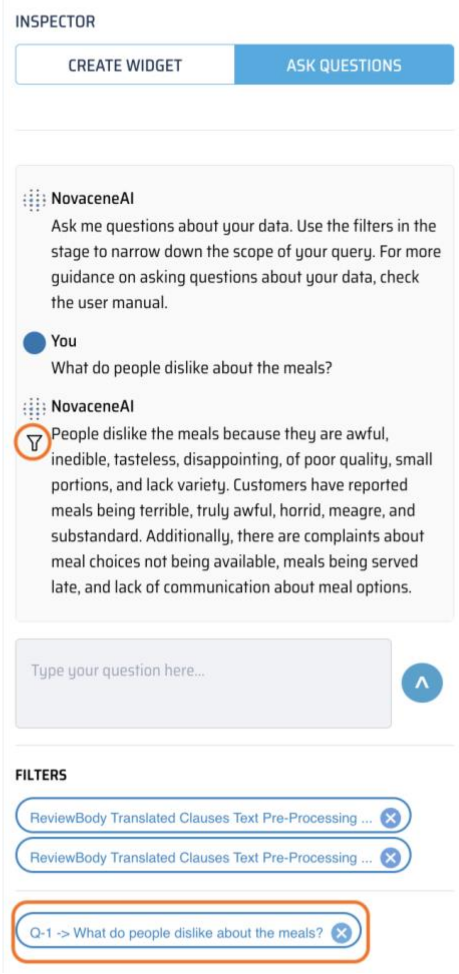
NovaceneAI
Ask me questions about your data. Use the filters in the stage to narrow down the scope of your query. For more guidance on asking questions about your data, check the user manual.

You
What do people dislike about the meals?

NovaceneAI
People dislike the meals because they are awful, inedible, tasteless, disappointing, of poor quality, small portions, and lack variety. Customers have reported meals being terrible, truly awful, horrid, meagre, and substandard. Additionally, there are complaints about meal choices not being available, meals being served late, and lack of communication about meal options.

Next, an answer is generated in the chat, and the dashboard is **filtered** to show you the exact text that was used to generate the answer in the text box. This allows you to ensure that generated answers are accurate, and to potentially find patterns in the visualization widgets. Take a look at the generated answer above, and the List widget with the clauses used to generate this answer.

You can add or remove these filters as shown in the image below.



NovaceneAI
Ask me questions about your data. Use the filters in the stage to narrow down the scope of your query. For more guidance on asking questions about your data, check the user manual.

You
What do people dislike about the meals?

NovaceneAI
People dislike the meals because they are awful, inedible, tasteless, disappointing, of poor quality, small portions, and lack variety. Customers have reported meals being terrible, truly awful, horrid, meagre, and substandard. Additionally, there are complaints about meal choices not being available, meals being served late, and lack of communication about meal options.

FILTERS

ReviewBody Translated Clauses Text Pre-Processing ...

ReviewBody Translated Clauses Text Pre-Processing ...

Q-1 -> What do people dislike about the meals?

You can add or remove Question filters using the filter toggle next to the generated answer, or remove them by hitting the “X” on the filter with your question text.

Tips for writing good questions:

- **Collaborate** with genAI by using the filters to help narrow down the scope of your query. For example, if you are interested in negative comments, use the negative sentiment enrichment filter. Let’s say we’re interested understanding why there are so many negative comments about the meals in this dataset—we'd use the topic filter “Meals” and sentiment filter “negative” to get the most relevant data for the genAI.
  You get the best results when you work together.
- Remember that this feature **only** sends the enriched text to the genAI—no other fields. For example: You have reviews about Shop A, B, and C. If you use the genAI feature

with no filtering to ask “What do people think about Shop A?” you might get an answer saying “I don’t know!”. Instead, filter the dataset by Shop A, and ask “What do reviewers think about this shop?”

- **Verify** your generated answers quickly and easily using the list widget. All generative AI is vulnerable to hallucinations. With this feature, you can see exactly what text has been sent to the AI to answer your question in your dashboard.

Supported systems

The application is programmed to work on most modern browsers. For the best experience, please use Chrome or Safari on PC or Mac.

Contact us

To contact Novacene, send an email to support@novacene.ai

Appendix: Built-in enrichers

The following is a partial list of built-in enrichers available on the application. Additional built-in enrichers may be described in section: Available built-in enrichers.

Note: Some enrichers require a previously-clustered datasets; these enrichers have been marked with a **superscript ([1]).**

Enricher name	Column suffix	Description
Clause Extraction	Clauses	Extracts clauses from input text.
Cluster Label[1]	Themes	Extracts the most central sample from each cluster.
Cluster Sampler[1]	Cluster Sampler	Outputs a set of the topmost representative samples in a cluster.
Cluster Summary Generator[1]	Cluster Summary	Generates a summary of the contents of the cluster using the OpenAI GPT API. Note that to overcome limits imposed by the API, the summary is based on a subset of the contents of each cluster. Requires that the input data has already been clustered.
LLM Cluster Themes[1]	Themes	Creates a label for clusters using the OpenAI GPT API. Requires that the input data has already been clustered.
Cluster Theme Extraction[1]	Themes	Outputs key phrases that describe the contents of a cluster. Extractive method.
Clustering	Cluster	Clusters text by grouping similar content together into categories (Does not include an unclustered category). Increasing the threshold produces a higher number of clusters with more members, while decreasing the threshold produces a lower number of clusters with less members.
Custom Theme Analysis	Themes	Uses an LLM to sort text into user-defined themes.
Emotion Analysis	Emotion Analysis	Classifies the emotion of the text as: *Anger, Annoyance, Anticipation, Compassion, Concern, Confusion, Contempt, Disappointment, Discomfort, Disgust, Embarrassment, Fear, Frustration,*

		Gratitude, Happiness, Hatred, Joy, Neutral, Relief, Sadness, Satisfaction, Surprise
HTML Content Cleanser		Removes HTML tags from text.
Ideas and Comment Classifier	Idea and Comment Classifier	Classifies input text as an *idea* or *comment.* Useful for idea extraction and analysis.
Image Quality		Assigns a score based on image quality factors such as contrast, sharpness and more. (Lower scores equal higher image quality).
Language Translator (Online)	Translated	Detects non-English text and translates into English. (Supports French, Spanish, and Chinese).
Language Translator (Offline)	Translated	Detects non-English text and translates into English. (Supports 25 languages, see Appendix: Supported Languages).
Names Entity Recognition	NER	Detects and highlight people, places, organizations, and other known entities found in text.
Public Support Detector	Public Support Detector	Assigns one of five classes on a 5-point Likert Scale ranging from *strongly approve* to *strongly disapprove.*
Public Support Sampler	Public Support Sampler	Outputs the top 5 most disapproving samples in decreasing stance score.
Sentiment Analysis	Sentiment Analysis	Outputs a *Positive, Negative,* or *Neutral* label for each sample. Score ranges from 0 – 1.0 and represents the confidence in the prediction.
Social Media Content Cleanser	Cleansed	Removes special characters commonly used in social media updates such as hashtags, @mentions, and more.
Term Frequency	Term Frequency	Outputs the number of times a *unigram* or *noun-phrase* appears in an input text.
Thematic Analysis	Themes	Outputs a theme group number and a label. Users can select a maximum amount of themes. Works best on datasets with more than 150k rows.
Threats / Abuse	Toxicity	Detects threatening or abusive language.

Table 2: Built-in enrichers

[1] Requires a previously-clustered dataset.

[2] Order of source dataset columns matter.

[3] Number of source dataset column matter.

Deprecated enrichers

Enricher name	Description	Deprecation reason
Sentiment Sampler	Outputs the top 5 most negative samples in the dataset in decreasing order of negative sentiment score.	Ranking to be treated as a function on the visualization system and not as part of the classification algorithm.
Cluster Summary[1]	Performs a one sentence summary using extractive summarization.	Generative methods outperform this extractive method.
Analyze Targeted Sentiment		
FinTech Analysis		
FinTech Social Media Cleanser	Strips tweets from potentially distracting content such as broken URLs, special symbols, ReTweet (RT) tags, etc.	Replaced by HTML cleanser and Social Media Content Cleanser
Hierarchical Clustering Responses	Clusters text by grouping similar content together into categories. (Optimized for some types of survey responses).	Inflexible for different data formats
Quantum Classifier	Binary classification algorithm that runs on a Quantum backend.	
Hierarchical Clustering Social	Clusters text by grouping similar content together into categories. (Optimized for short texts or social media updates).	
Peer Clustering[2] [3]	Clusters records that share many similar attributes.	
Sentence Segmentation	Expands the input text by segmenting input text into separate sentences.	Replaced by Clause Extraction
Sentiment Pre-Processing		Replaced by Sentiment Analysis and Emotion Analysis.

Sentiment Sampler	[TBD]	Please contact us for instructions.
Sentiment Analysis (Retail)	Specific for retail data, classifies the tone of the text as being *positive, negative, or neutral.*	Replaced by Sentiment Analysis
Topic Modelling and Clustering M1	Clustering for text	Earlier versions of Clustering
Topic Modelling and Clustering M2	Clustering for text	Earlier versions of Clustering

Appendix: Supported Languages

Arabic	Dutch	Greek	Italian	Russian
Catalan	Esperanto	Hebrew	Japanese	Spanish
Chinese	Finnish	Hindi	Korean	Swedish
Czech	French	Hungarian	Persian	Turkish
Danish	German	Indonesian	Portuguese	Ukrainian

Appendix: Ideas and Comments Classifier

The Generation Energy Idea and Comment Submissions dataset is comprised of 1779 instances. Text from the dataset included a wide range of topics which can be observed in the table below.

Theme contained in dataset	
Forum Related Themes	**Number of Related Ideas**
Affordability/Abordabilité	64
Biomass/Biomasse	34
Communities/Commuautés	71
Energy Efficiency/Efficacité énergétique	4
Electricity/Électricité	87
Finance/La finance	183
Geothermal/Géothermie	22
Governance/Gouvernance	47
Heating /Le chauffage	20
Hydro/Hydroélectricité	36
Hydrogen/D’hydrogène	15
Information/Information	11
Innovation/Innovation	98
International/Internationale	107
Labour Markets/Marché du travail	47
Natural Gas/Gaz Naturel	71
Nuclear/Énergie Nucléaire	26
Petroleum/Pétrolier	149
Pipelines/Pipelines	66
Public Confidence/La confiance du public	3
Remote Communities/Les collectivités éloignées	11
Renewables/Énergies renouvelables	280

Security/Sûreté	10
Solar/Solaire	150
Storage/Stockage d’énergie	34
Tidal/Énergie Marémotrice	15
Transportation/Transports	166
Wind/Éolienne	65
Youth/jeunesse	5