Learn Session Training

Watson Discovery Service

Lab Exercise

Topic: Expert Services Learn – Watson Discovery Service

Version: **1.6**

Updated: February 2, 2018

For questions about the Watson workshop and materials, contact:
Giulio Soliani *E-mail: gsolian@us.ibm.com*

Lab Overview

This lab guides you through the configuration steps required to use Watson Discovery Service to satisfy a Voice of the Customer user scenario. Watson Discovery Service will be configured to help end-users visualize and analyze product feedback obtained via customer product reviews.

Change Log

Revision History: Changes to this document are summarized in the following table in reverse chronological order (latest version first).

Revision	Date	Author	Short Description of Changes
1.6	02/02/18	Raffi K.	Edited Section 5.4 to align with WDS enrichment updates
1.5	12/06/17	Raffi K.	Added Section 7: Relevancy Training
1.4	11/22/17	Raffi K.	Updated WDS photos and query collection steps
1.3	05/20/17	Giulio Soliani	Published Lab v1.3 with updated use case description
1.2	05/03/17	Chethan R	Published Lab v1.2: incorporated review comments
1.1	04/24/17	Chethan R	Published Lab v1.1
1.0	04/24/17	Chethan R.	Published Lab v1.0
0.3	04/21/17	Giulio S.	Reviewed Data Crawler Section
0.2	04/20/17	Giulio S.	Added Section 1
0.1	04/17/17	Chethan R.	Created Section 2-6

Contents

Course overview

Credits

This project is as a collaboration between many experts within and outside of Watson in the hopes of broadening skills within IBM, our clients, partners and the developer community.

If you have suggestions for how to improve this guide, please send a note to gsolian@us.ibm.com.

Objectives

During these lab exercises, you will configure IBM Watson Discovery Service to fit a sample discovery use case scenario. You will learn how to use the new IBM Watson Discovery service tooling to walk through the main configuration steps required to setup, build, and refine a Watson Discovery Service instance.

The tasks and tools that are covered within the Lab allow you to perform and complete all necessary steps to efficiently and effectively ingest Amazon Product Reviews data, configure your Watson Discovery Service Collection to convert, enrich, and normalize the data, and Query the collection to discover insights.

Lab Structure

When completing this lab, we included explanations for reference to support the general understanding of the subject matter and to get a better sense of what you are actually doing.

To complete the exercises, you only need to perform numbered steps.

This lab exercise is designed for every student to work independently using a Mac or PC. Each student requires their own IBM Cloud account to create an IBM Discovery Service instance in Cloud.

Exercise Environment

The lab exercise will be performed on the IBM Watson Discovery Service provided as part of the Watson Developer Cloud.

Course Focus

The lab consists of 5 main sections.

1. Getting started with Watson Discovery Service and creating your own instance and environment.
2. How to configure you through ingesting content within IBM Watson Discovery Service.
3. Ingesting data using the Watson Discovery Service Data Crawler.
4. Presenting an alternative to data ingestion: using the Watson Discovery Service to ingest your data.
5. Extracting relevant insight from your data set by querying the collection.

Preparation

To prepare for the lab exercises, locate the Lab folder provided to you by your instructor. Make sure to unzip the Lab Folder to a directory on your local hard drive.

The lab folder contains the following files:

File name	Description
WDS Lab Exercise v1.0	Lab manual
Amazon_Product_Reviews Folder	Folder containing 999 sample Amazon Product Reviews

0 Getting Started

While this Lab can be performed using Windows, the current version of this guide assumes the use of a Linux or iOS computer to complete all tasks.

To perform these lab exercises, follow these instructions to set up your computer.

0.1 Overview

This lab guides you through the configuration steps required to use Watson Discovery Service to satisfy a Voice of the Customer user scenario. Watson Discovery Service will be configured to help end-users visualize and analyze product feedback obtained via customer product reviews. The data set is a collection of reviews for Amazon Kitchen and Dining products. The application end-user is Henry, an Amazon Business Analyst, responsible for reading and analyzing consumer reviews to:

- Identify all relevant feedback on product functionality;
- Address product pain points and upsell product success point; and
- Make informed suggestions and provide actionable insight to the product development team.

The lab walks through the development activities required to configure IBM Watson Discovery Service to satisfy this use case. What this lab does not cover are the broader best practices and considerations that you should follow when creating an enterprise-grade solution.

0.2 Working with Watson Developer Cloud

0.2.1 IBM Cloud

IBM Cloud is an implementation of IBM's Open Cloud Architecture, leveraging Cloud Foundry to enable developers to rapidly build, deploy, and manage their cloud applications, while tapping a growing ecosystem of available services and runtime frameworks. You can view a short introductory video here: http://www.ibm.com/developerworks/cloud/library/cl-bluemix-dbarnes-ny/index.html

The purpose of this guide is not to introduce you to IBM Cloud, which you should already be familiar with, at least on a high level.

0.2.2 Watson Developer Cloud (WDC)

IBM Watson Developer Cloud (**WDC**) is a platform of cognitive services, designed to help developers build solutions to help users extract insight from Big Data. Cognitive computing systems learn and interact naturally with humans to augment their ability to make better decisions from data.
As such, the Watson Developer Cloud services offer a variety of services that cover various aspects of natural interaction including text (**Natural Language Understanding**, **Natural Language Classifier**, **Conversation**), images (**Visual Recognition**), and speech (**Speech To Text** and **Text To Speech**).

Furthermore, WDC offers services to understand a user's personality (**Personality Insights**) and emotional/social tone (**Tone Analyzer**) in a scalable manner.

0.2.3 Watson Discovery Service (WDS)

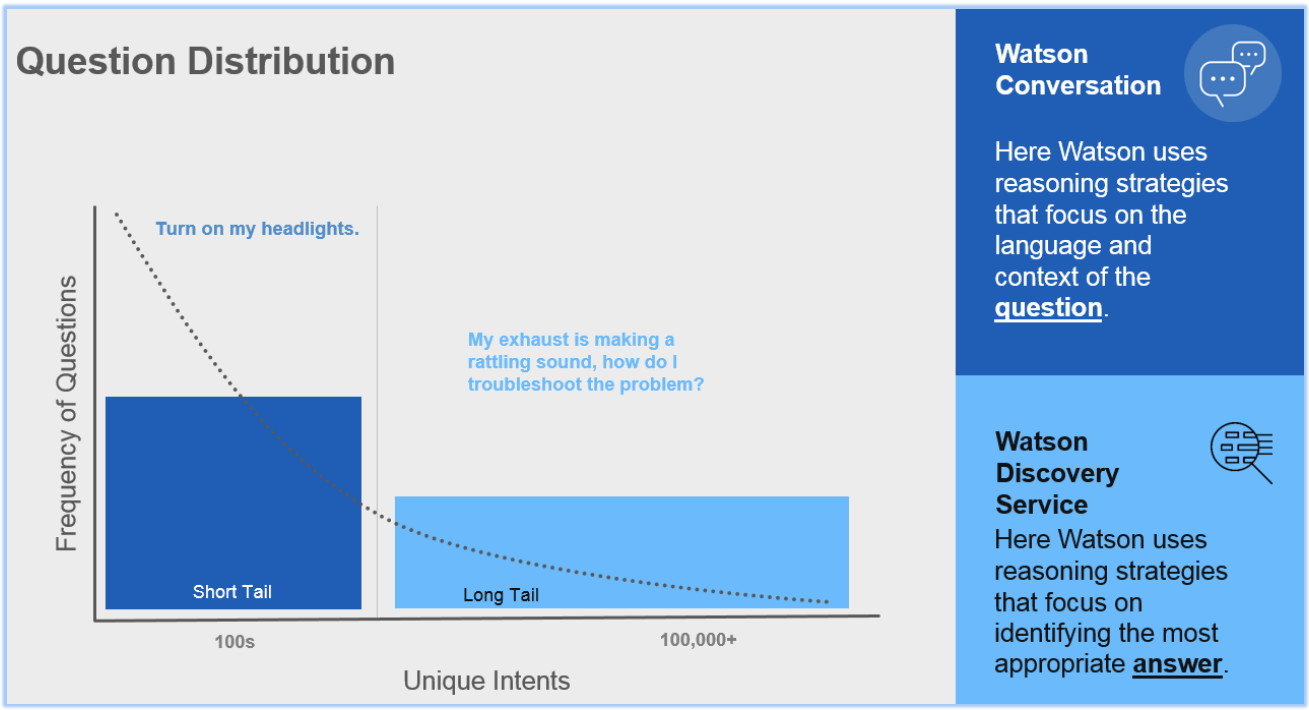
This guide takes an instructional approach to working with the IBM Watson™ Discovery Service, a cognitive search and content analytics engine that allows applications to identify patterns, trends and actionable insights that drive better decision-making.

Watson Discovery Service comes with an easy-to-use graphical interface to ingest, configure, and query content indexed into collections. Creating and querying your first collection using the IBM Watson™ Discovery service (WDS) entails the following steps:

- Creating your Watson Discovery Service environment;
- Configure the WDS pipeline to convert, enrich, and normalize your content (involves using a sample data file);
- Ingest your content in a Collection as a searchable text index; and
- Creating queries to extract relevant insight from your Collection.

IBM Watson Discovery Service is designed to securely unify structured and unstructured data with pre-enriched content, and uses a simplified query language to eliminate the need for manual filtering of results. The input data can be a HTML, DOC, PDF, or JSON document. WDS returns the output in the form of JSON which can be easily queried to gain insights into the data.

IBM Watson Discovery service can also be used to augment Watson Conversation Service and/or address use cases where "long tail" less frequently asked questions require answers obtained from documentation. This use case will not be covered in this Lab.



0.3 Prerequisites

This section provides instructions to help you get started quickly with the IBM Watson™ Developer Cloud services using Node.js as your programming runtime environment. To make it easy to get up and running with a functional application that uses the REST Application Programming Interface (API) for any Watson service, IBM provides a Node.js package with wrappers that simplify application development. The package includes simple command-line example applications to let you experiment with any of the available services.

1. To get set up, complete the following steps:

0.3.1 Obtaining a IBM Cloud account

IBM Cloud is a cloud PaaS (Platform as a Service), which allows you to host your application on-line and bind it to a variety of SaaS service offerings from IBM including Watson analytic services.

Learn more at http://www.ibm.com/bluemix and if you are new to IBM Cloud, you can create a trial account at http://www.bluemix.net/.

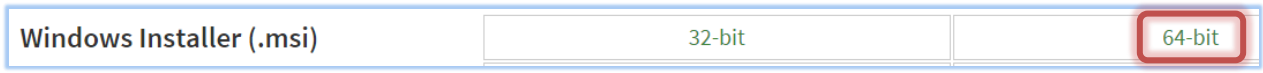
2. Direct your browser to the Cloud home page: https://console.ng.bluemix.net/home/ to access your dashboard
3. If you do not yet have a Cloud account, click **Sign Up** on the top right
4. Enter requested information and click **Create Account**

If you use your personal e-mail address, you have 30 days to evaluate Cloud. Some services, such as Watson Discovery Service, are free for limited use during the trial period.

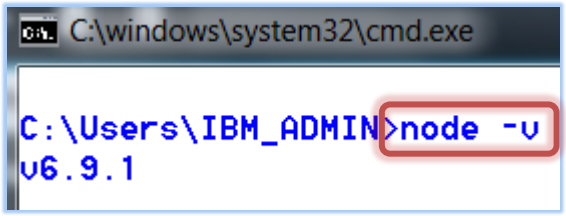
0.3.2 Installing the node.js runtime

The default installation includes both the runtime and package manager. Make sure to include the installed binaries on your PATH environment variable after installation (typically, the default installation locations that the installer selects does the inclusion).

5. Direct your browser to the nodejs.org web site: https://nodejs.org/en/
6. Click **DOWNLOADS** in the menu bar
7. Select and install the installer (not the binary), appropriate for your operating system (i.e. Windows Installer.msi (64 bit)

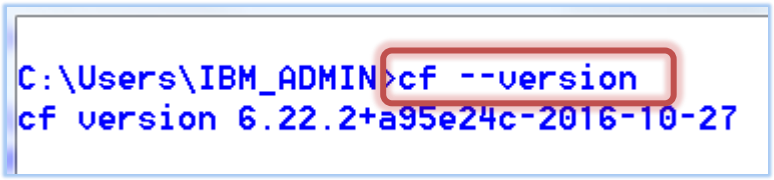


8. Complete the installation
9. Verify that node.js has been installed by opening a command line and issuing `node -v`



0.3.3 Installing cf (Cloud Foundry) CLI

10. Direct your browser to a GitHub repository: https://github.com/cloudfoundry/cli/releases
11. Download and install the most recent installer appropriate for your operating system
12. You may need to open Preferences > Security and Privacy > General tab (in Mac)
13. Unlock and change the Allow applications downloaded from Anywhere
14. Verify that installation was successful by issuing `cf --version` from a command line



0.3.4 Verifying Prerequisites

Before you start the following section, please run the following commands to verify you have the required packages installed on your machine.

1. Open a Terminal window
2. Check that git is installed by typing `git --version`

   If this returns an error, then git is not installed. Please download and install git at https://git-scm.com/download/win
3. Check that node.js is installed by typing `node -v`

   If this returns an error, then node is not installed. Please download and install Node from nodejs.org
4. npm should be installed with nodejs, verify it is installed on your machine by typing `npm -v`
5. Check cf (cloud foundry command line interface) is installed by typing `cf -v`

   If this returns an error, then cf is not installed.

0.3.5 Downloading OS-specific coding-friendly editing tool

6. If you are using a PC, you can use **Notepad ++** (https://notepad-plus-plus.org/download)
7. For Mac, use **Sublime Text** (http://www.sublimetext.com/3) → also works for Windows

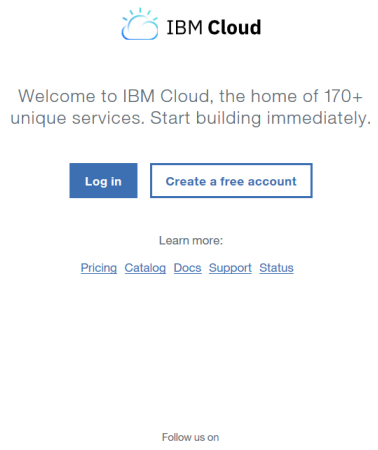
0.3.6 Installing the Watson Discovery Service Data Crawler

Watson Discovery Service provides a Data Crawler tool that will help users take their documents from the repositories where they reside and push them to the cloud, to be used by the Discovery Service. The Data Crawler is recommended when users are managing an upload of a significant number of files from a remote system, or trying to extract content from a supported repository (such as a DB2 database).

8. Installation instructions for the Data Crawler tool can be found here.
9. Please review the Data Crawler supported platforms available here.
10. Supported repositories can be found here.

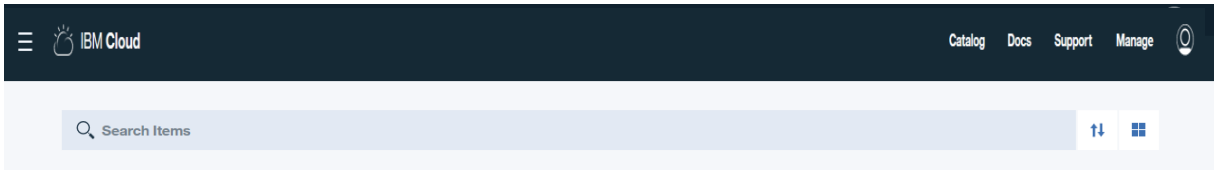
1 Creating Your Watson Discovery Instance

1. Navigate to IBM Cloud - https://console.ng.bluemix.net/ and click on Log In

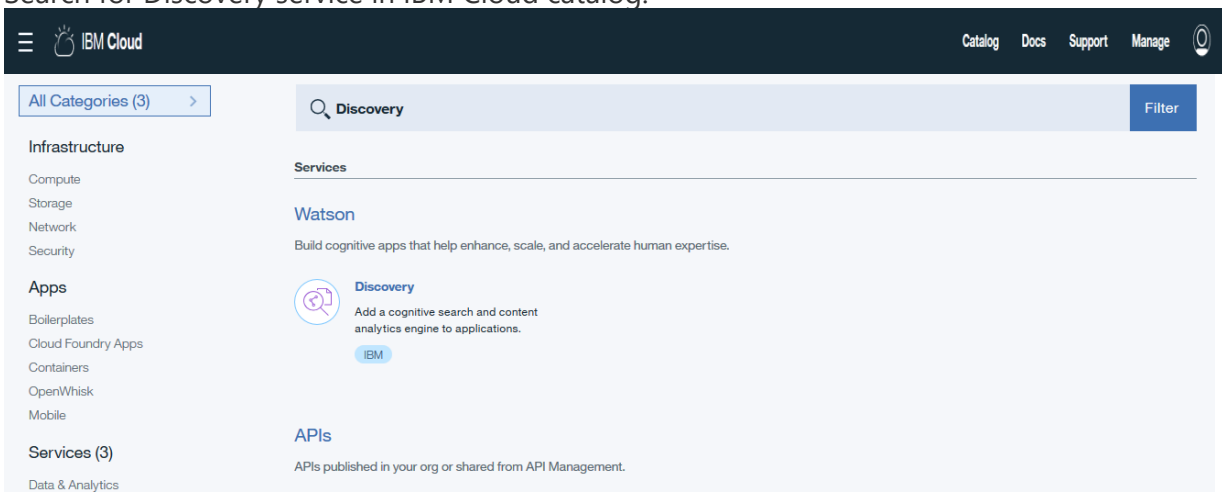


2. Use your organization credentials obtained in Step 0.3.1 to Log in to IBM Cloud.

3. Once logged in, navigate to the Cloud Catalog by clicking on Catalog in the top left-hand corner. You can alternatively follow this link - https://console.ng.bluemix.net/catalog/



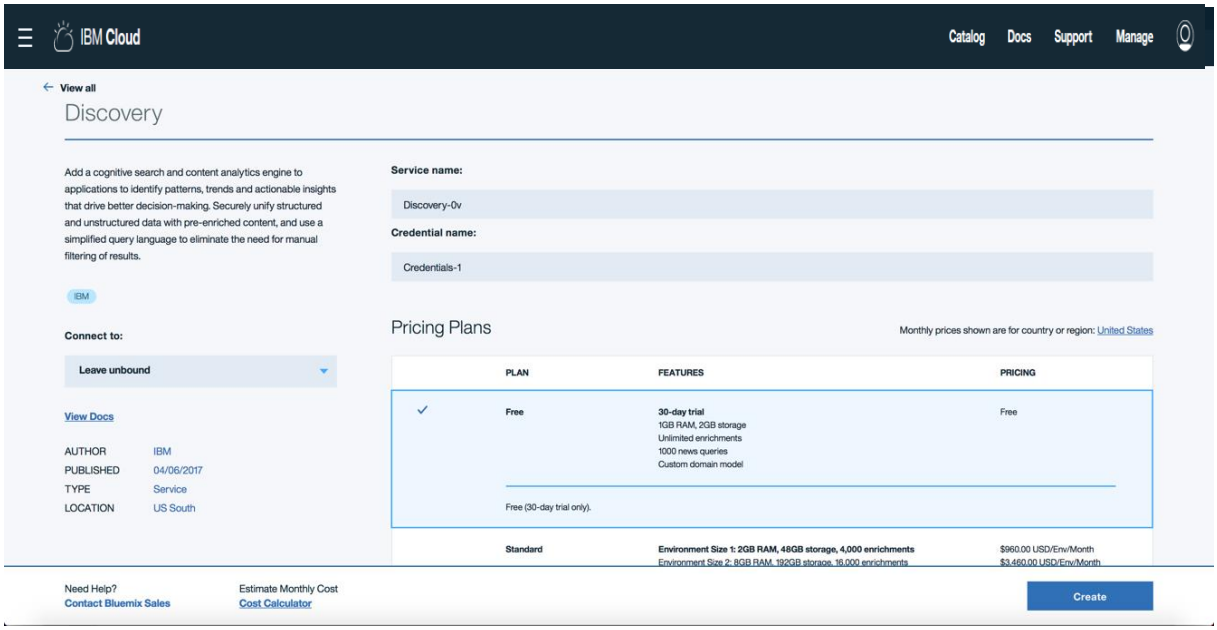
4. Search for Discovery service in IBM Cloud catalog:



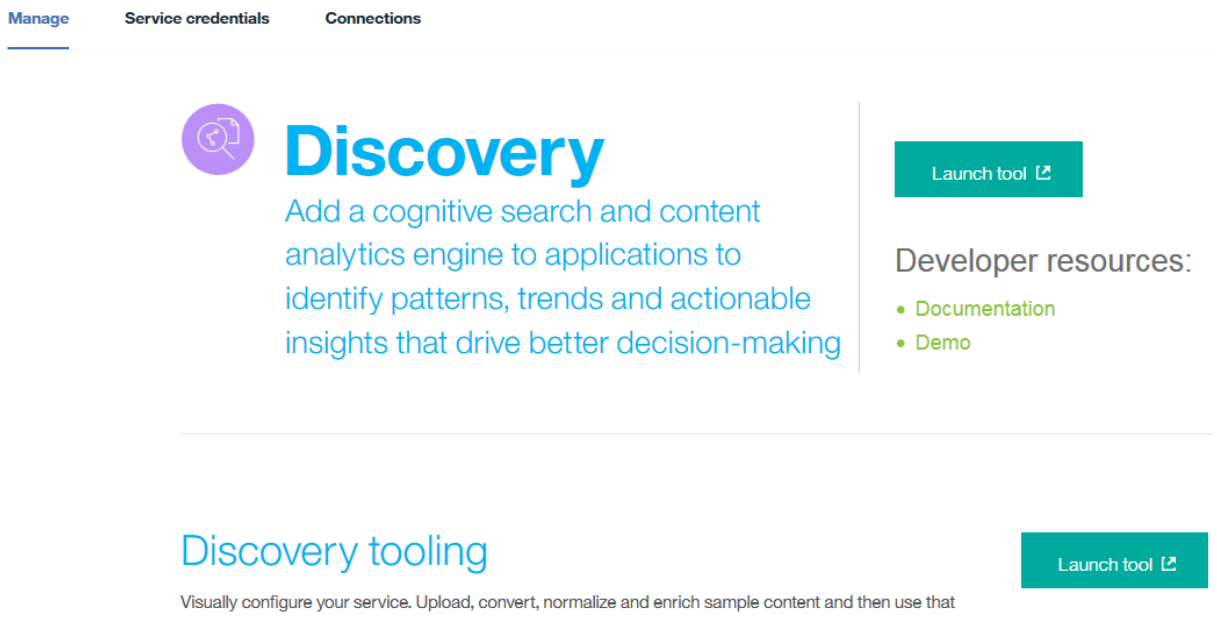
5. Click on the Discovery tile

6. Within this window, you can provide a unique Service name (e.g. Discovery_TeamName). Click on the blue box under Service name to provide a unique name to your instance. Do not make any changes to the Credential name

7. Create a trial service instance by selecting the Free Plan option (default setting) and by clicking Create.



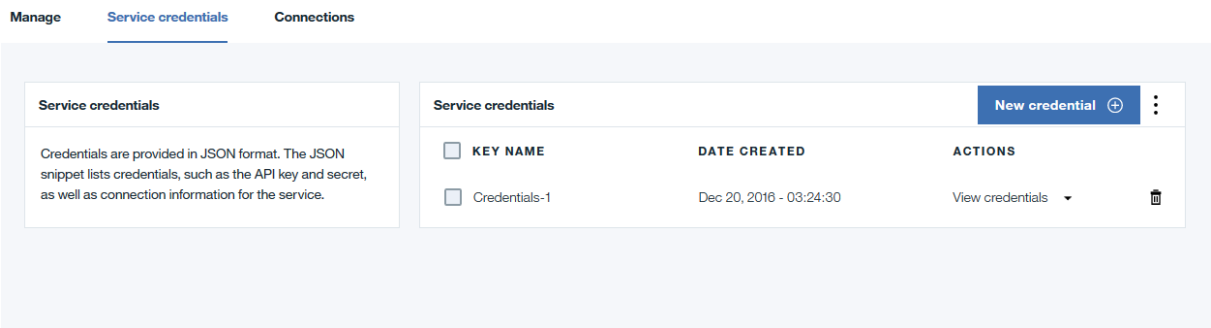
8. IBM Cloud will now instantiate your personal Watson Discovery Service instance. It will automatically guide you to the service dashboard.



If you close the tab and would want to re-access it, you can navigate to the service dashboard directly by selecting the created discovery service from the list of services created in your account. The list of the services you have created can be found at https://console.ng.bluemix.net/dashboard/apps

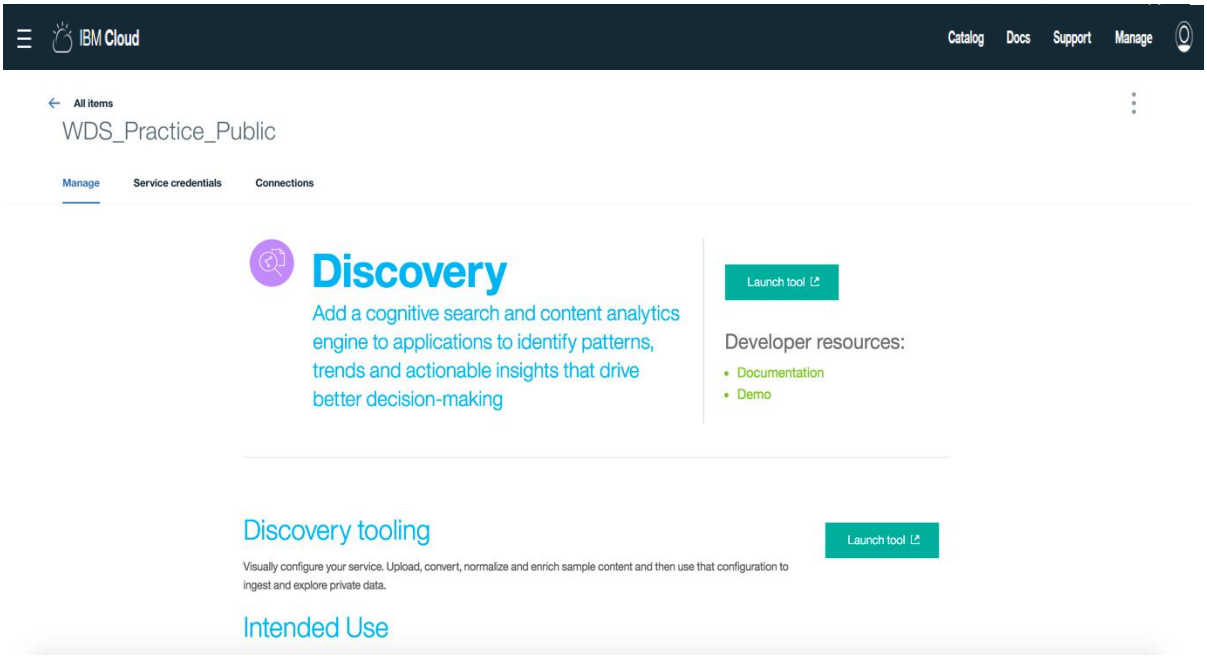
9. In the service dashboard click on the “Service Credentials” tab and go to service credentials section. Click on “View Credentials” dropdown to view Username, password and URL

10. Select and Copy the service credentials to a local file and save it for future reference

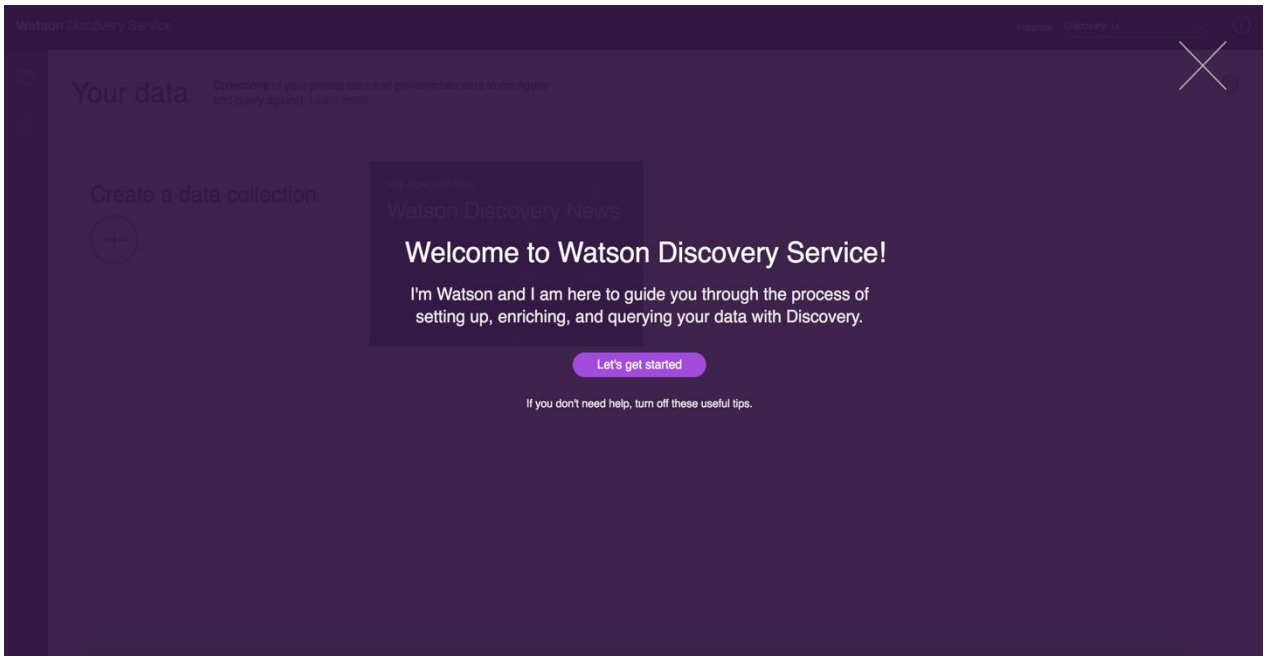


11. Once you’ve recorded Username and Password. Click on the Manage tab to return to the service dashboard.

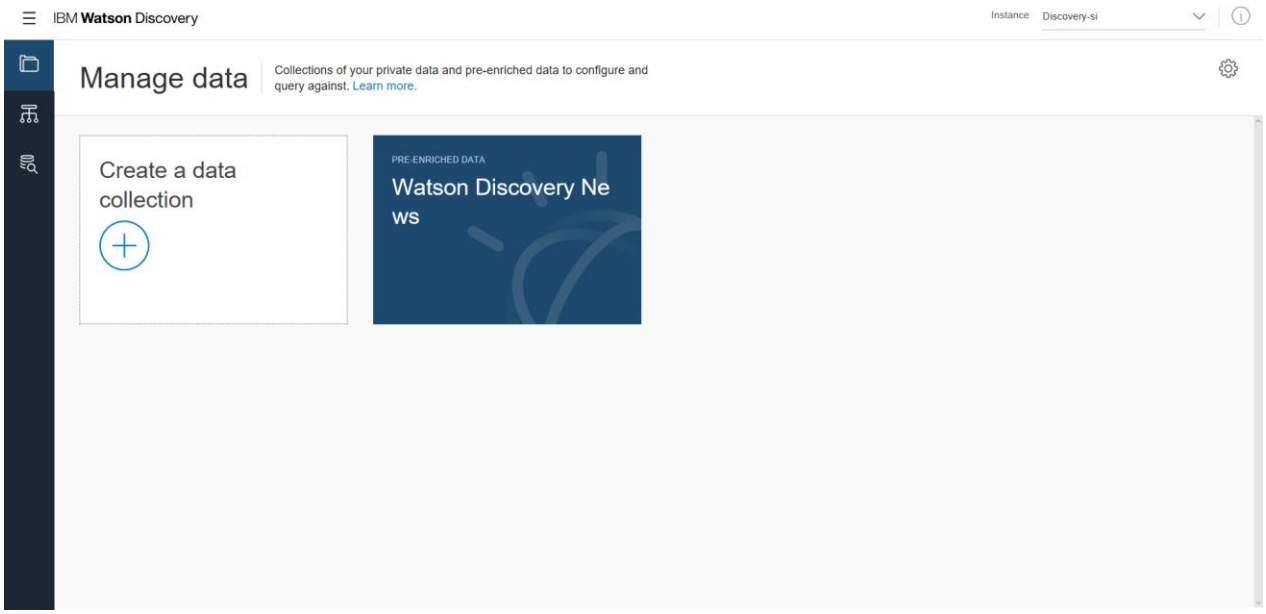
12. Launch the discovery tool by clicking “Launch Tool” button from the discovery service instance



13. Launching the tool takes you to the dashboard with welcome message overlay. Click on the cross button on the right top to close the overlay window.



14. In the dashboard, you can see the collections present in your Watson Discovery Service environment. An option is provided to create a new collection.

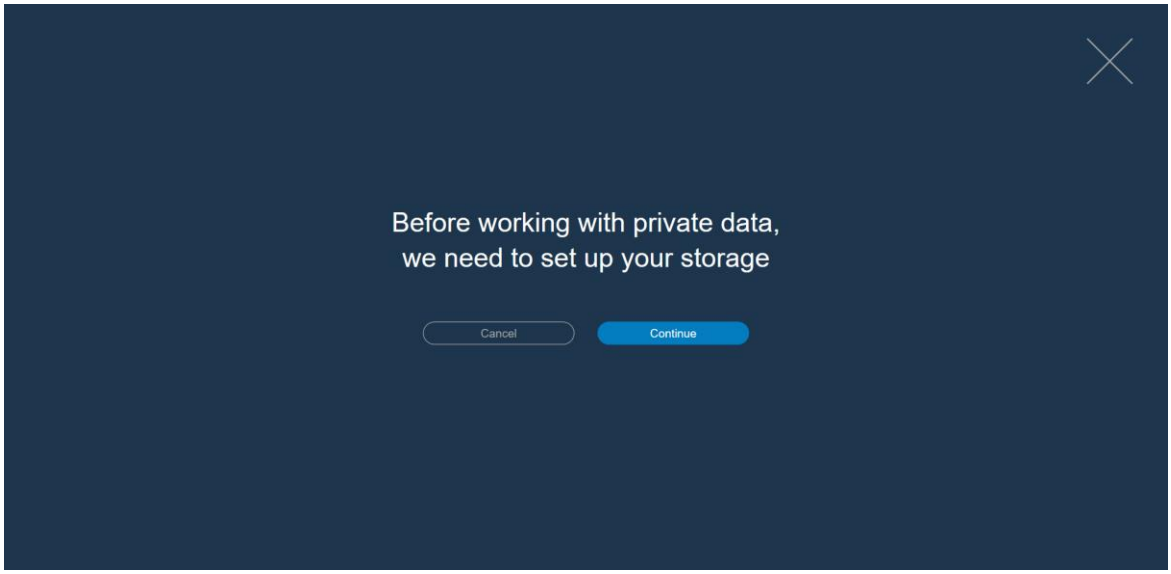


A new Watson Discovery Instance will come with a default environment, configuration and a pre-enriched collection, Watson News. To upload Amazon review data, you will have to create a new collection.

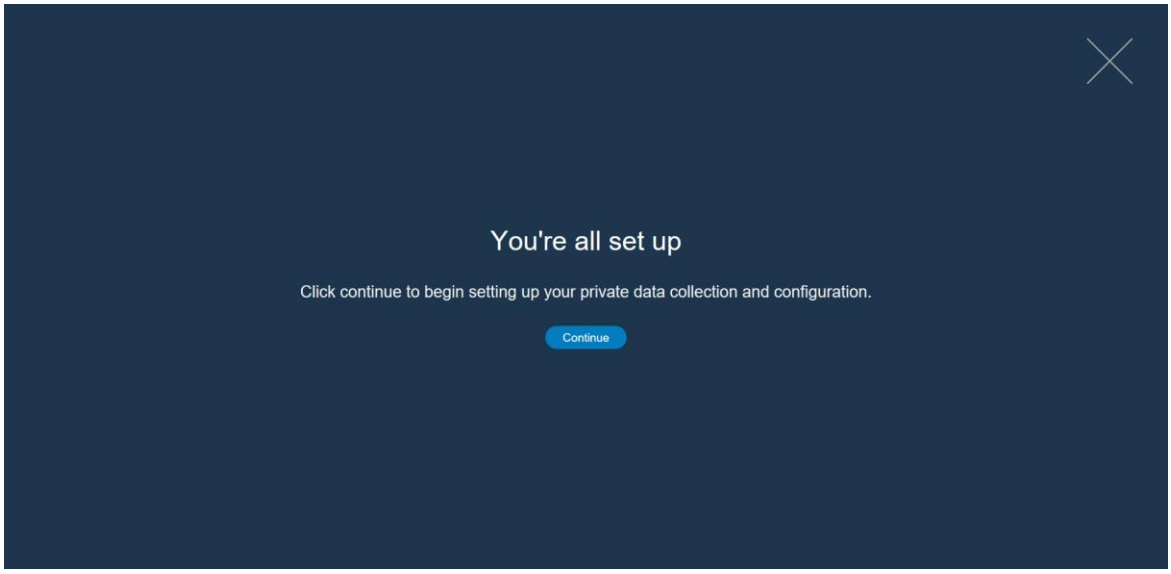
2 Configuring Your Watson Discovery Service Instance

1. Create "Amazon Product Reviews" collection by clicking on Create a data collection button

2. When you create your first collection, the tooling will indicate that the storage needs to be setup before working with private data. Click "Continue" button. This process can take 1-2 minutes

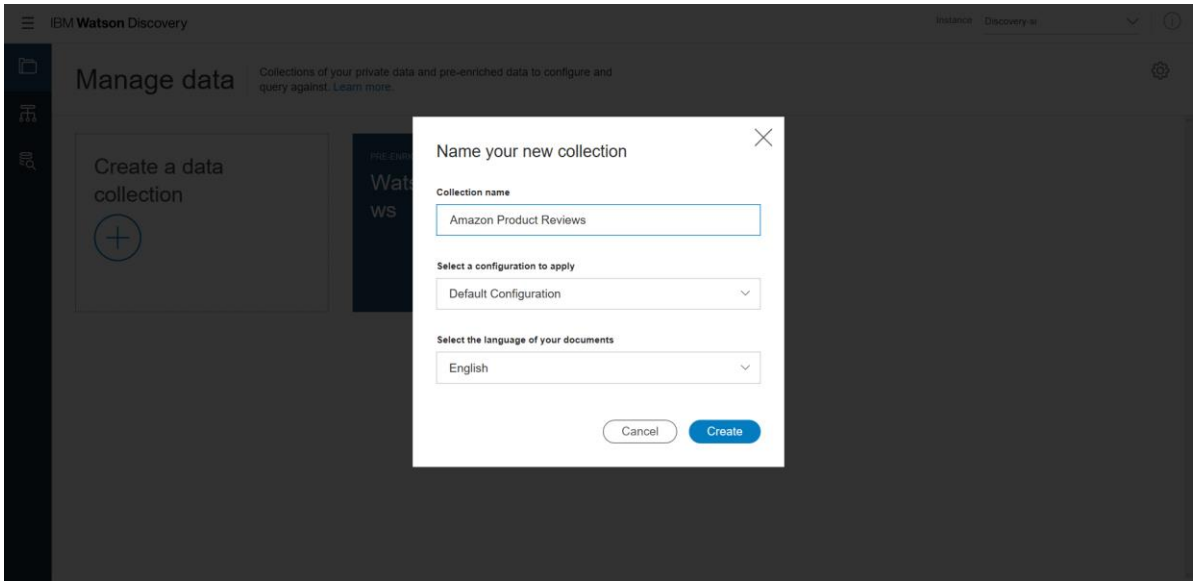


3. You get a message saying, "You're all set up" once the storage is provisioned and the collection is created. Click "Continue" button



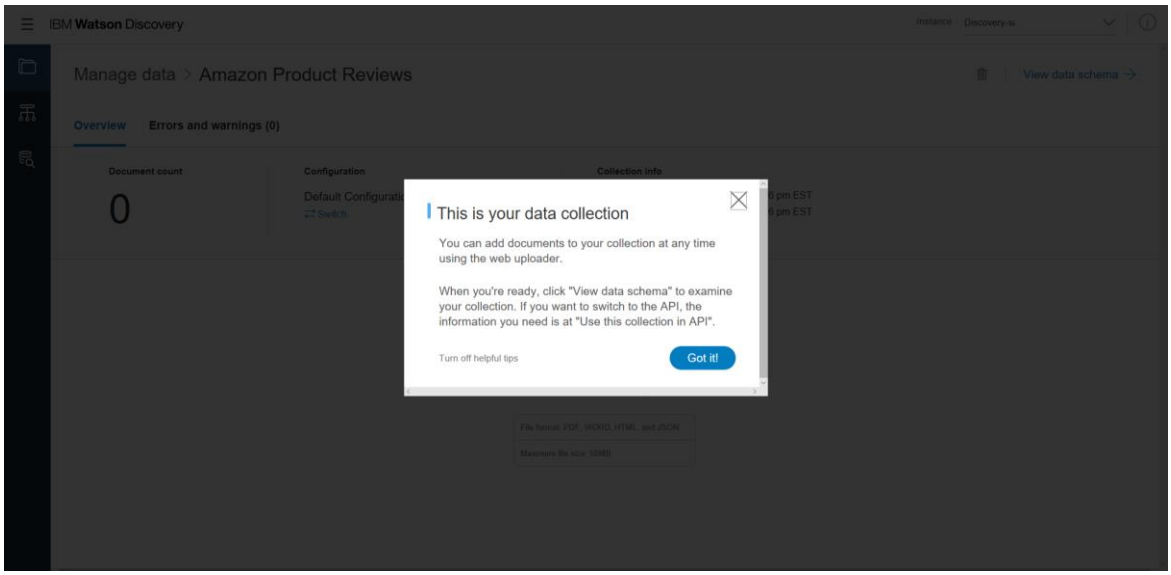
4. Again, click on Create a new collection and specify the name for your Collection. In this Lab, you will create a collection that you will want to name "Amazon Product Reviews".

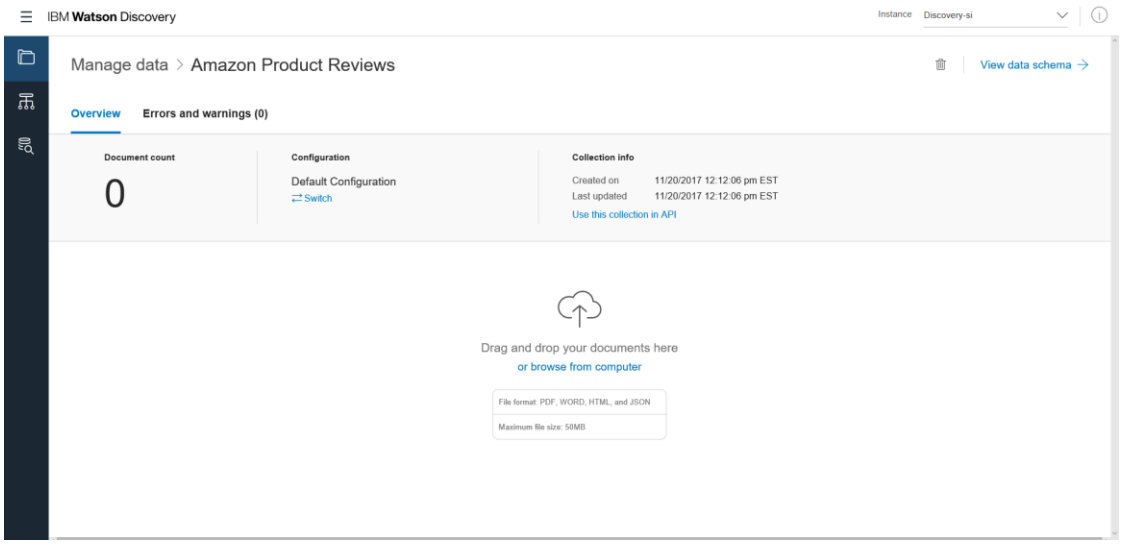
5. Apply the Default Configuration to the collection (which is already selected for you) and click the Create button.



6. If you are creating a collection for the first time, it automatically creates a hosting Environment which may take up to 2 minutes to provision the storage for environment

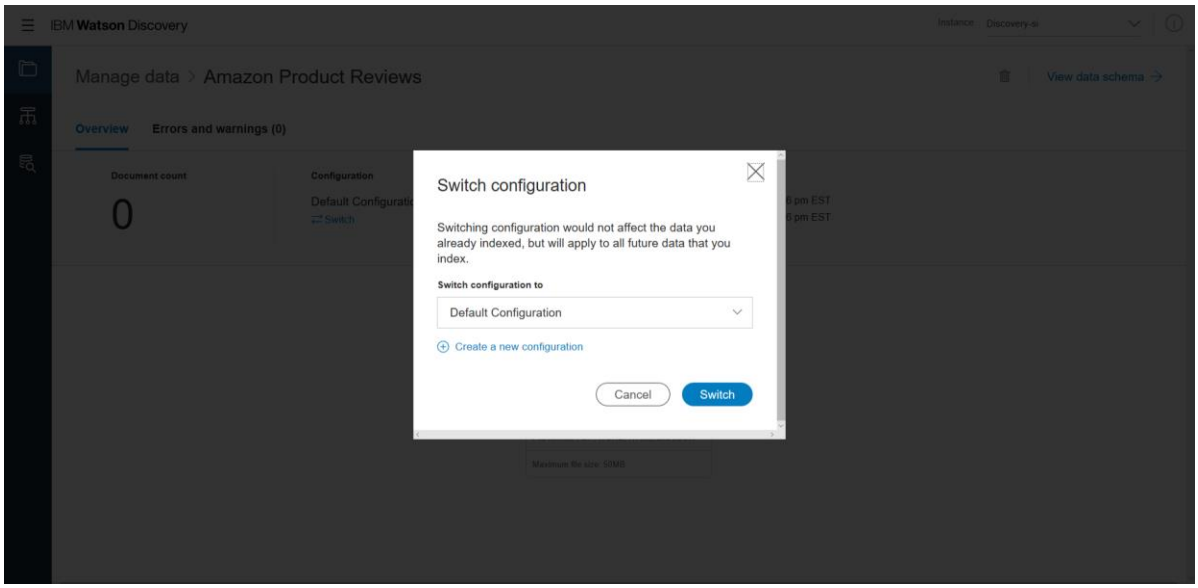
7. The collection landing page opens up automatically. The collection landing page displays information like configuration id, environment id, collection id and the current configuration that was selected for your collection.



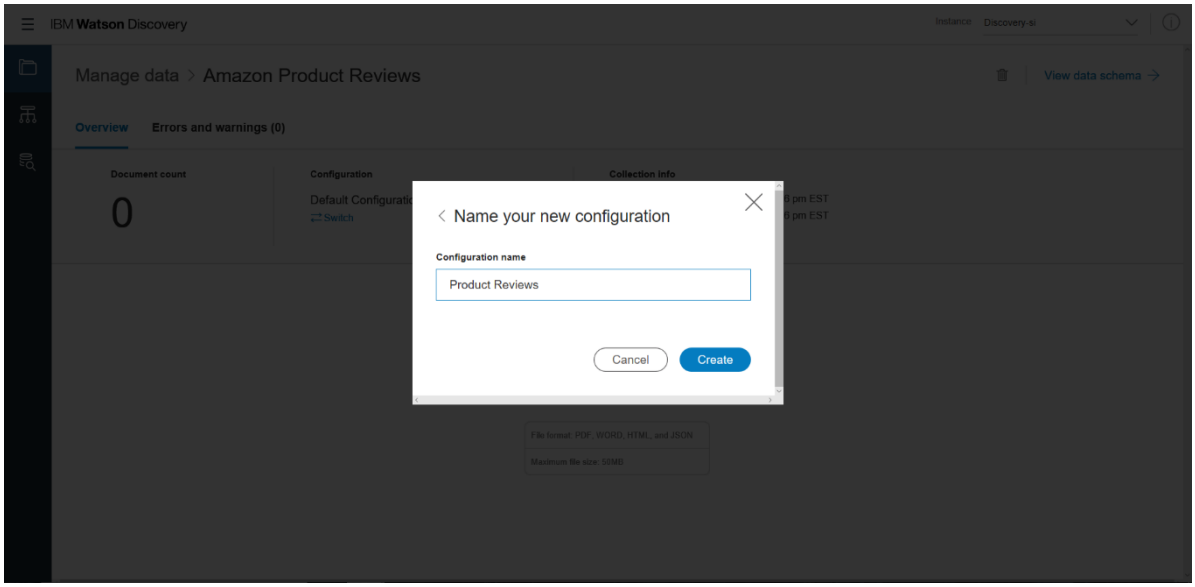


Before ingesting Amazon Product Review data, it is important that you create a configuration tailored to your content format and structure.

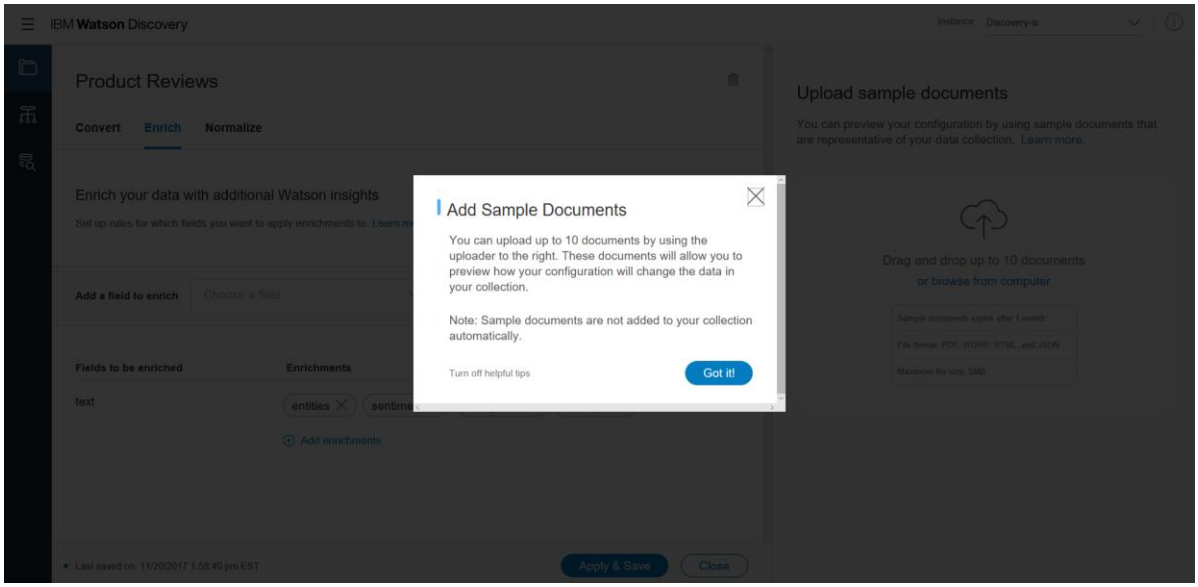
8. To create a new configuration, click on "Switch" button next to Default Configuration
9. Select Create a new configuration.



10. Name the configuration: "Product Reviews"



Note: If there is already a preexisting configuration which you had created before, you can select it from the dropdown when you are switching the configuration



11. You are now ready to create a Configuration that is specific to Amazon Product Reviews. Before moving to the next Section, you will need to confirm that the Amazon data set is present within your Lab folder directory. Make sure you locate the Amazon Data files on your machine.

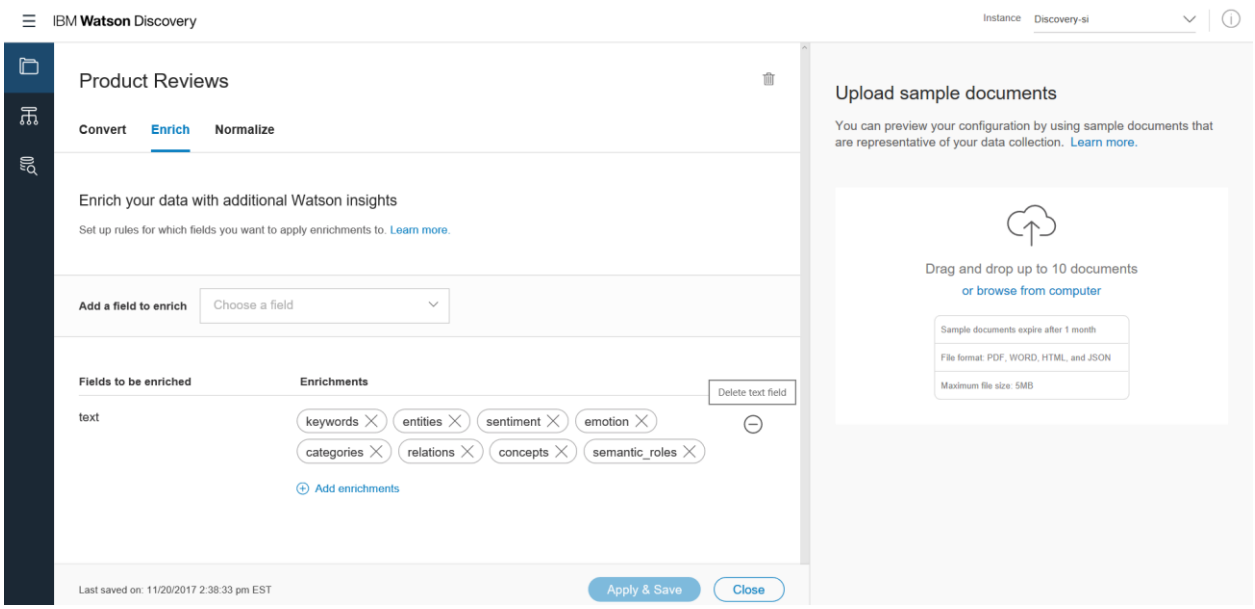
2.1 Selecting attributes for conversion, enrichment, and normalization in the configuration

User should identify the data field from the Amazon Product Review data which needs to be converted, enriched, and normalized by the Watson Discovery Service. NLU and NLP is embedded in WDS through the enrichments so you can receive seamless data insights with only a few clicks. The identified field is

then selected for enrichment during creating/modifying configuration. In general, the data will be analyzed to extract features out of it against which the analysis will be made. But, the Amazon review data is available in a structured format(JSON) and data is arranged in the form of attribute:value pair. Users will need to identify which field to be attributed. For this use case, we are enriching review_text from the review documents. As mentioned earlier, it is essential to set the configuration and the attributes to be enriched before uploading all your data files.

2.1.1 Configuring Enrichment Attributes

1. Remove the existing field "text" from the fields to be enriched section since the review does not have any attribute called "text". You can remove the field by clicking the "minus" button next to the field which appears when you scroll over on the field – the delete button will be to the far right.

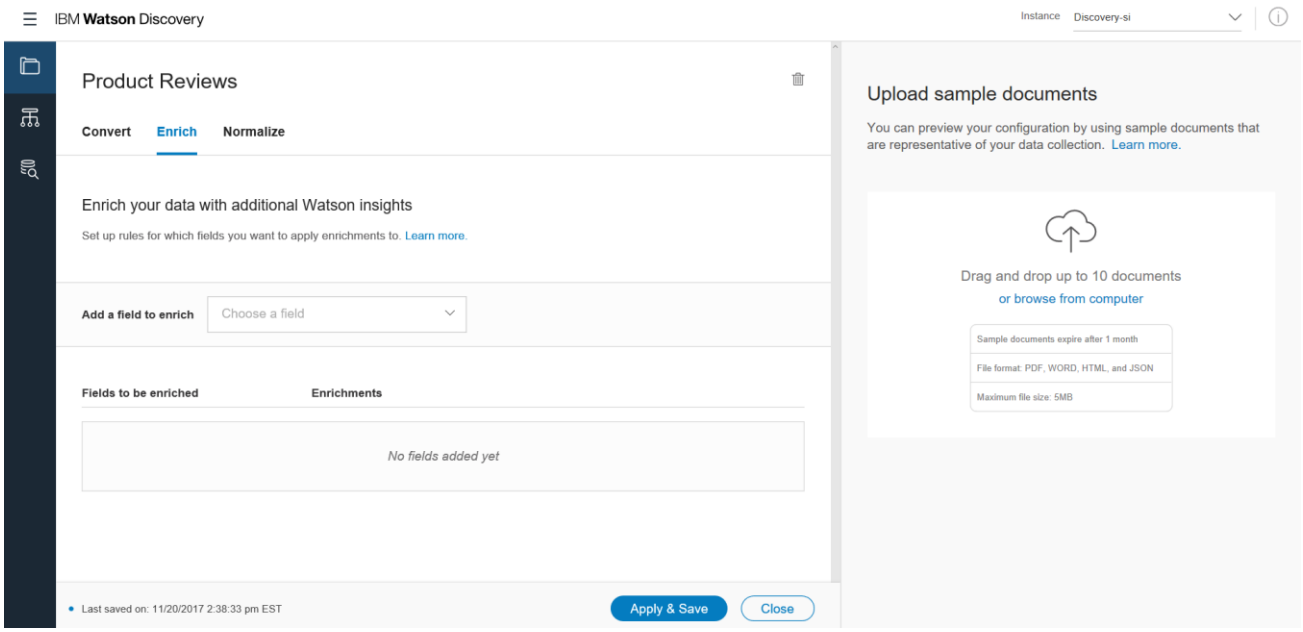


2. Download the amazon review files and store it in a new folder and name it as "contents". You can find the review files at https://ibm.ent.box.com/folder/23913798838

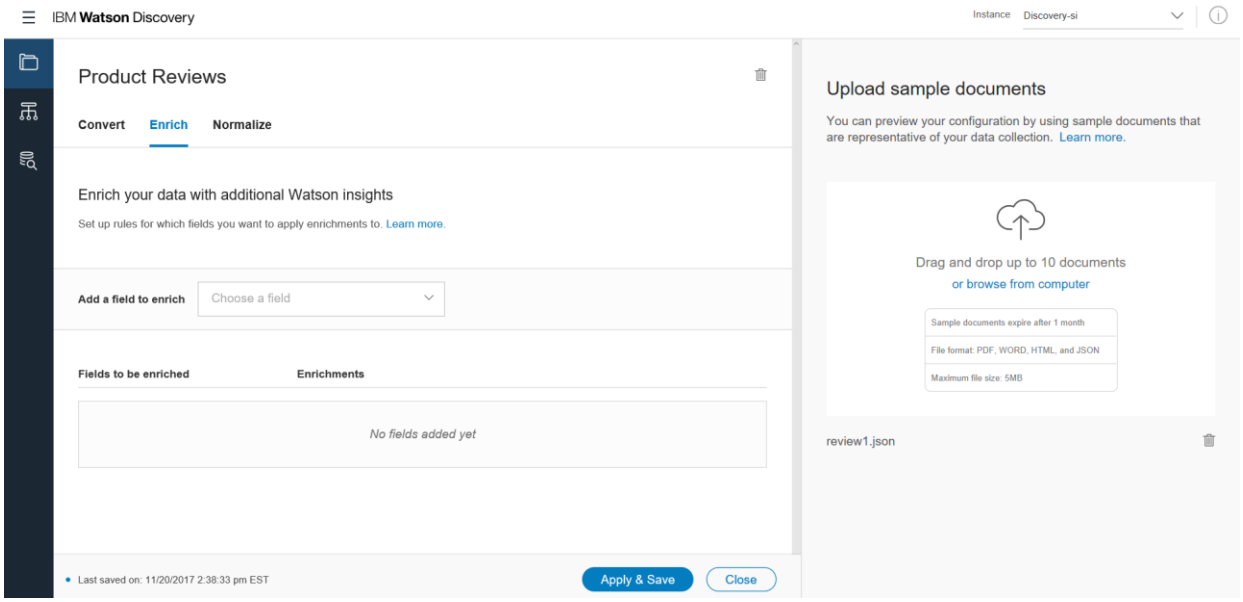
3. Identify a sample Amazon Product Review from the Content folder.

4. Upload Review1.json from the Content Folder by browsing to the Content Folder using the "browse from computer" button.

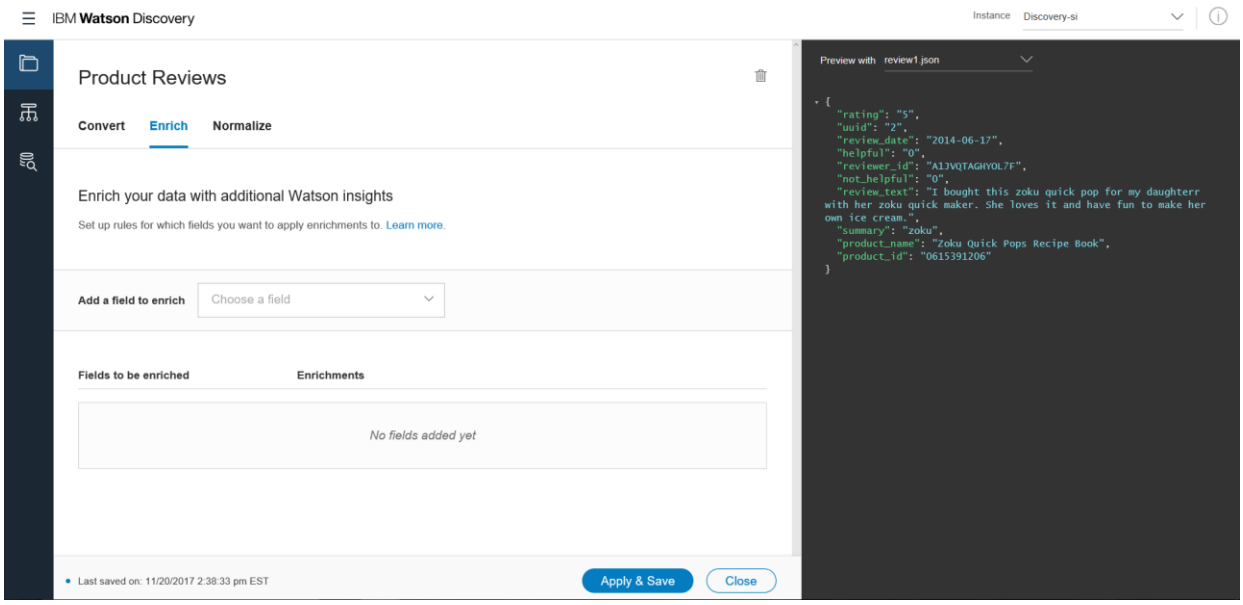
5. Select Review1.json and click Upload to complete the ingestion of the sample document.



6. Once it has been uploaded, the Review1.json is shown under Upload sample documents section which can be clicked to open in preview mode

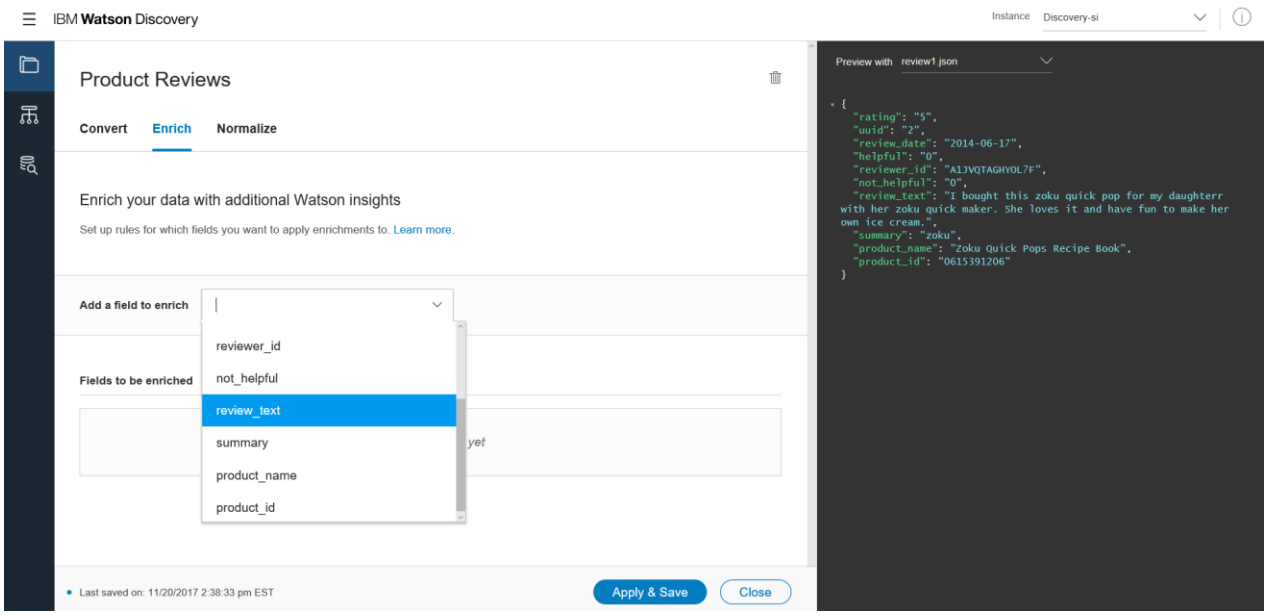


7. Click on Review1.json to open up a preview of the document. The Amazon Product Review data fields are displayed within the window.

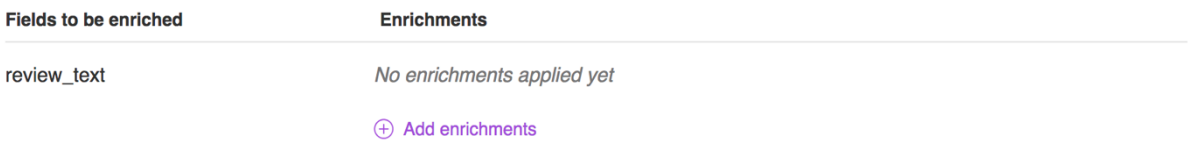


8. The automatically identified fields from the sample review document are available to be selected for enrichment from the dropdown menu next to Add a field to enrich.

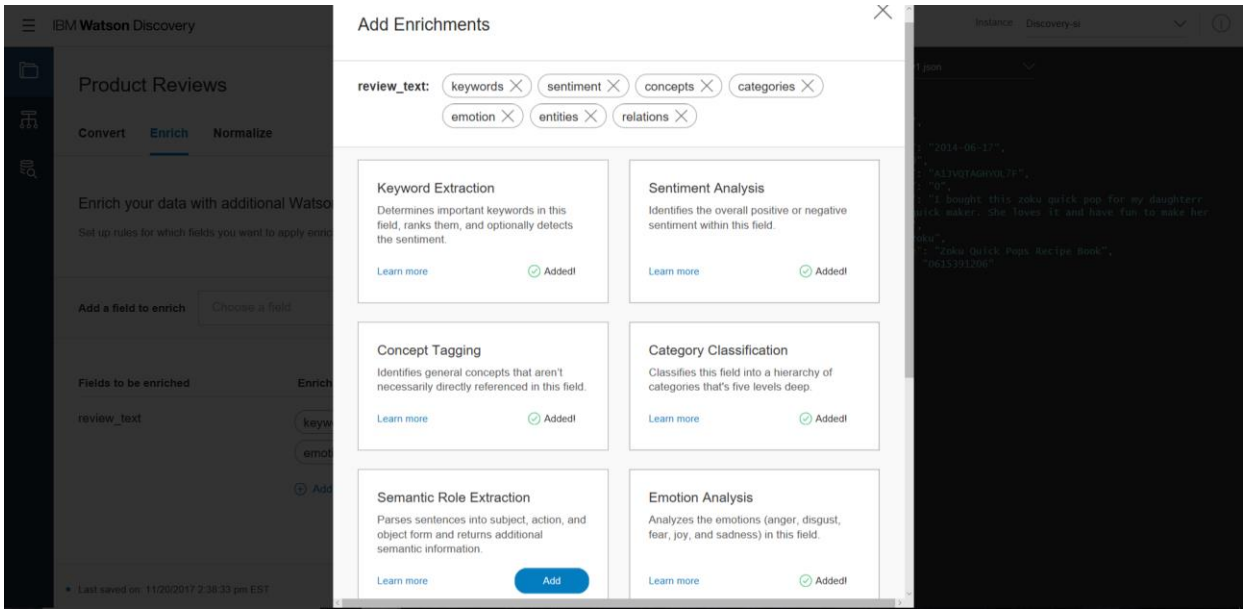
9. Click on the drop-down menu and select “review_text” for enrichment



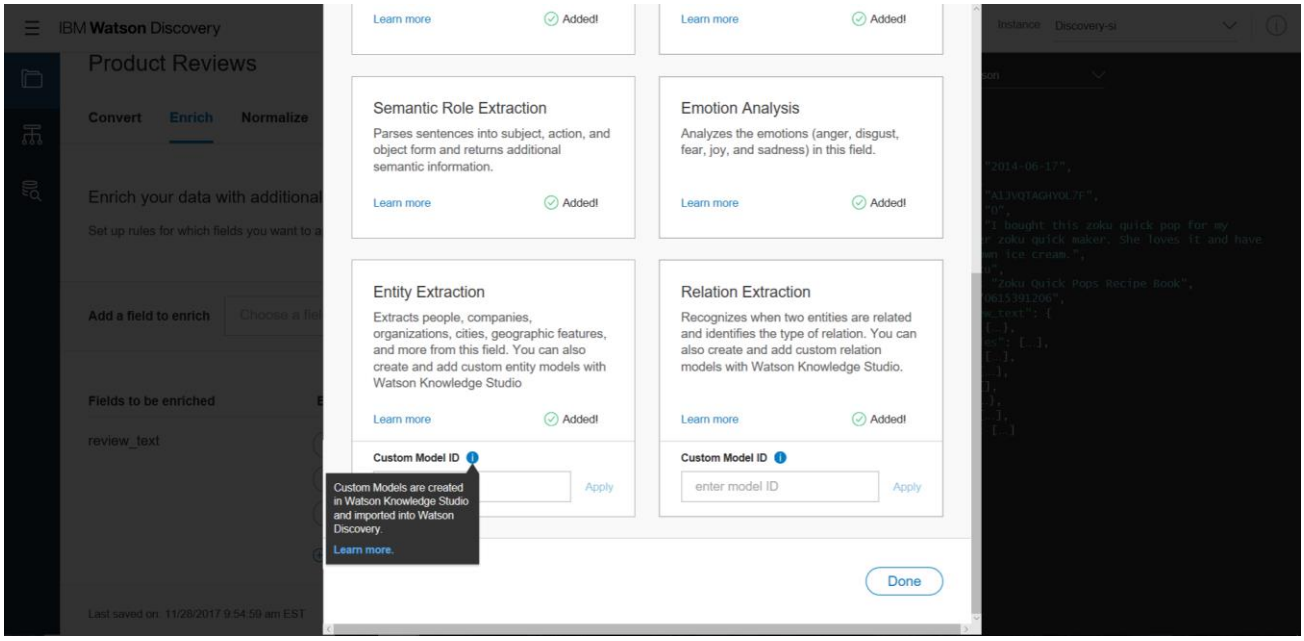
10. After selecting the field to be enriched, you need to select the type of enrichments you want applied to your data. Click on “Add Enrichments” next to the field to be enriched



11. Select all enrichment types by clicking Add on all tiles. The "added" message will appear once the enrichment is selected. Once you've selected all enrichments, click Done.

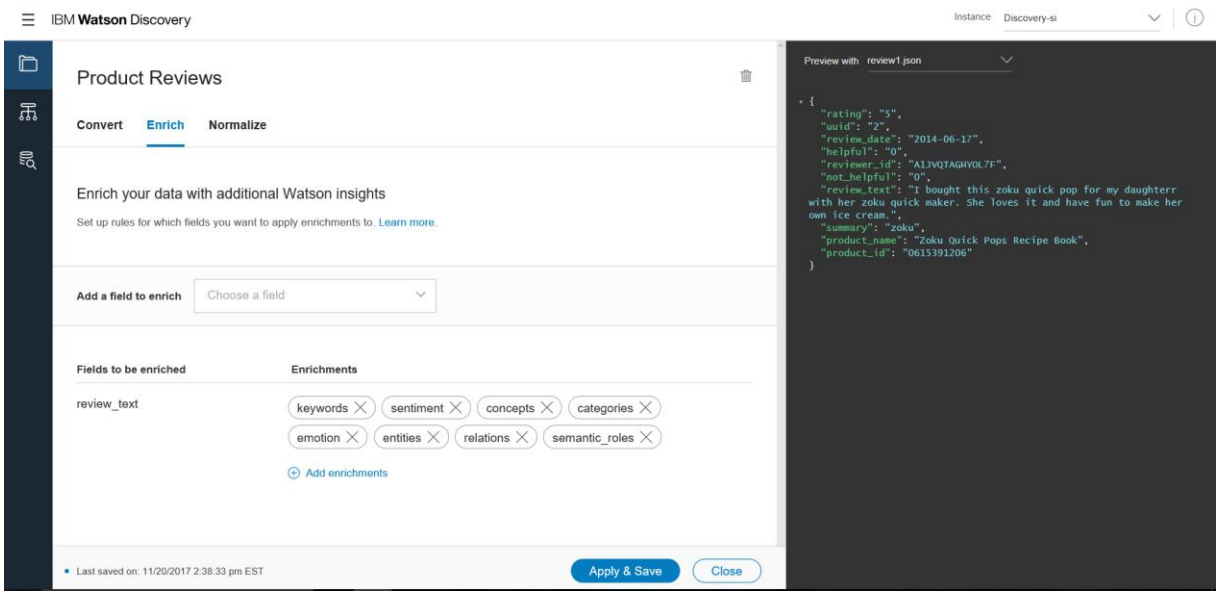


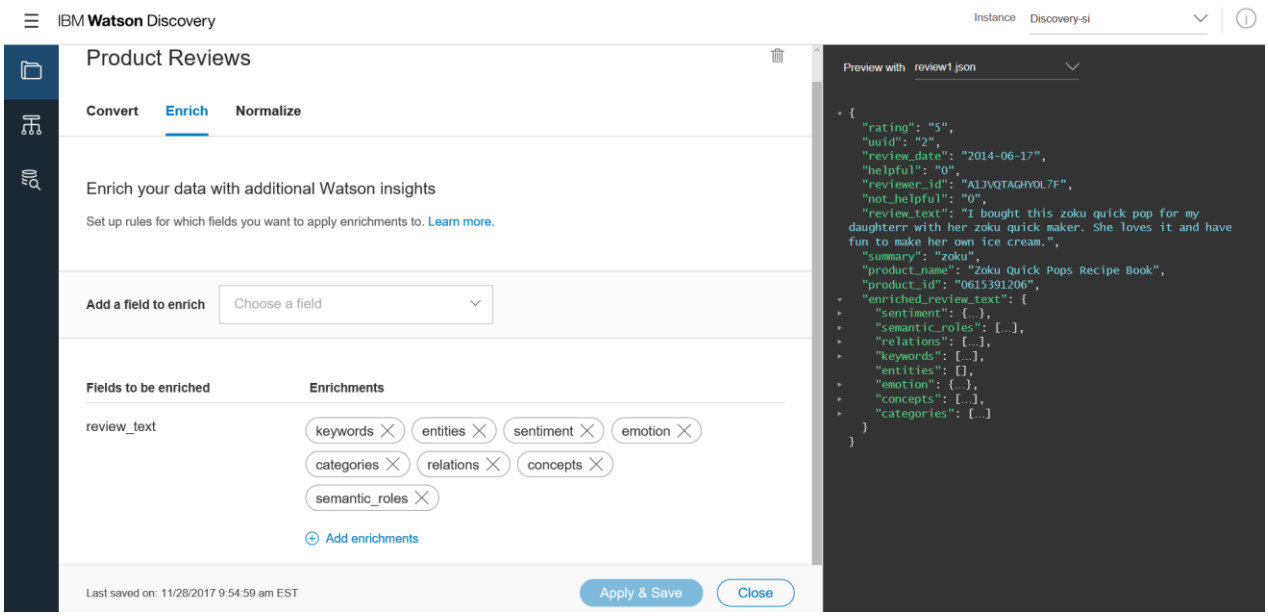
Note: If you have a WKS model, you can deploy it in enrichments by adding the model ID



12. The selected type of enrichments is now shown in the Enrich page. Now click "Apply and Save" button to save the configuration. This configuration will reflect against any documents added to the

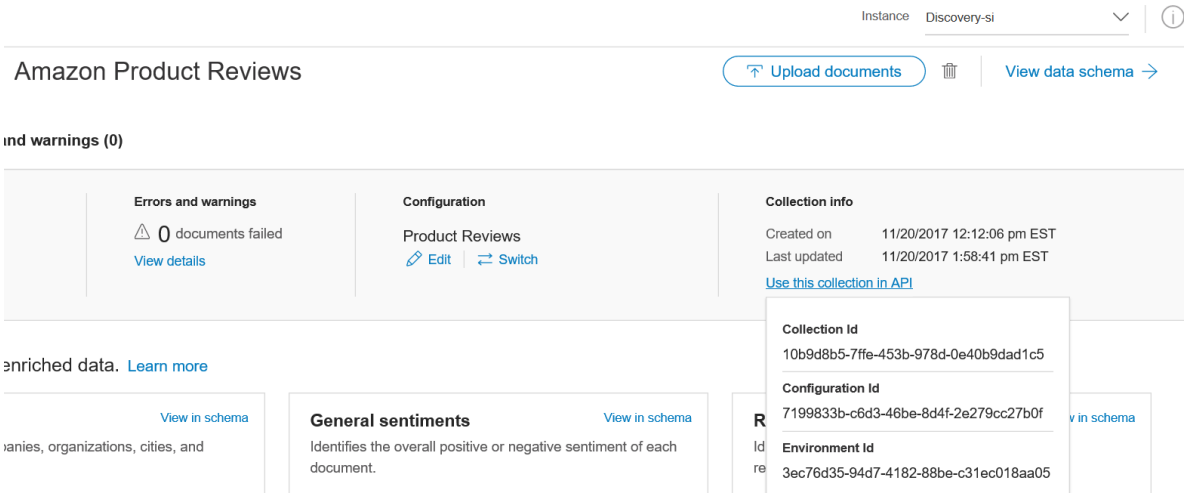
collection from now on.





13. Click on the Close button to return to the Your Data dashboard.

14. Click on the Amazon Product Review blue collection tile to open the collection dashboard. From the collection info, click on 'Use this collection in API' to receive your Collection ID, Configuration ID, and your Enviroment ID. Make sure you record in a separate text file the Environment ID, Configuration ID, and Collection ID. This will be used later for data crawling.



15. Now that you have defined the Enrichment configuration, you are ready to ingest data within the collection. For more complex use cases, you will also have to set Conversion and Normalization configuration attributes. For this lab, we are working with clean Amazon Product Review data so you are not required to set Conversion and Normalization attributes. However, we do cover those tasks as Optional in Section 2.1.2 and Section 2.1.3. Feel free to explore these optional Tasks.

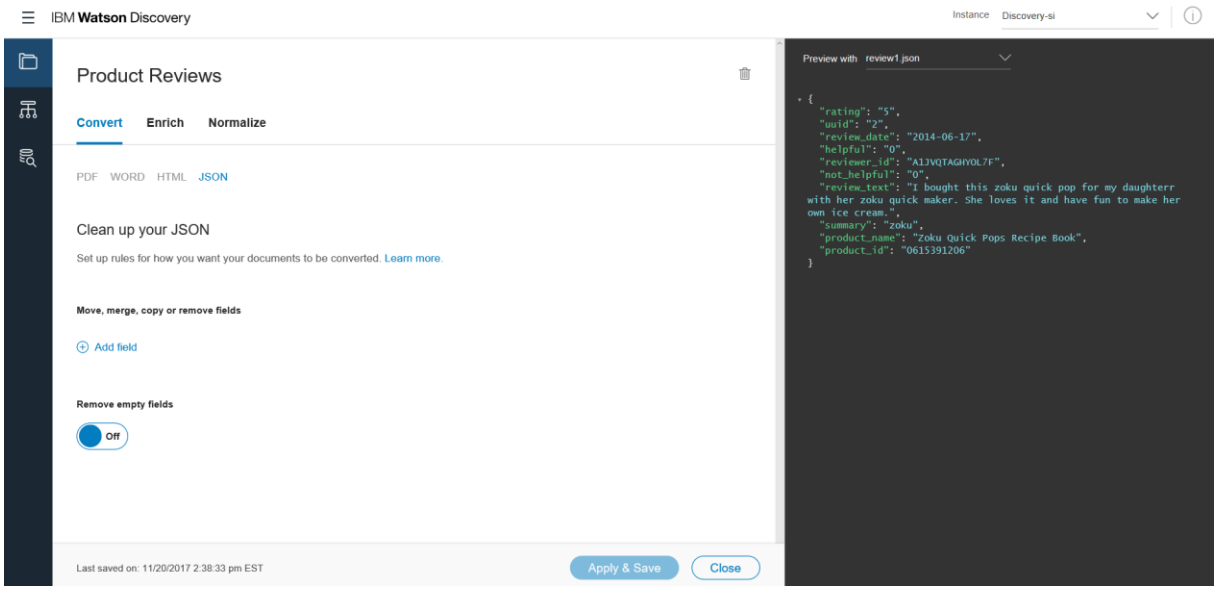
16. Alternatively, if you are a developer, please refer to section 3 to learn how to use data crawler to ingest files.

17. If you are a general user, please refer to section 4 to learn how to use tooling interface to upload files.

18. Users can either use crawler tool or tooling interface but not both.

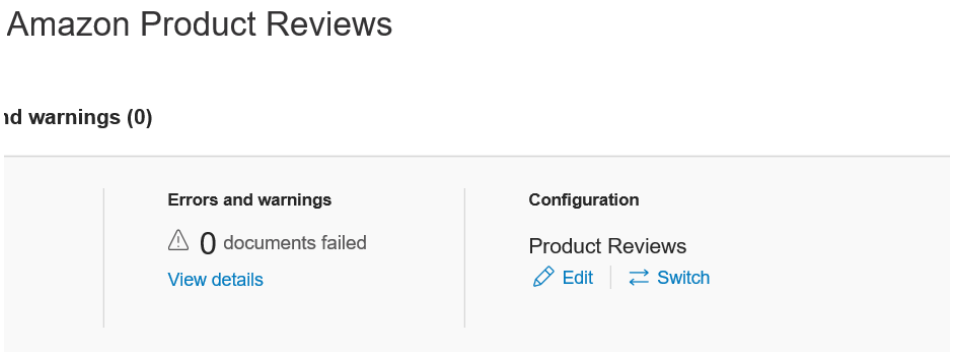
2.1.2 Configuring Conversion Attributes (Optional for this lab)

You can perform the following conversion steps after completing Section 2.1.1. These steps are optional and meant to illustrate the Conversion capability that is provided with Watson Discovery Service. Keep in mind that users should configure the conversion attributes prior to selecting the enrichment attributes. The configuration pipeline can be visualized as going left to right: from Convert to Enrich and then to Normalize.

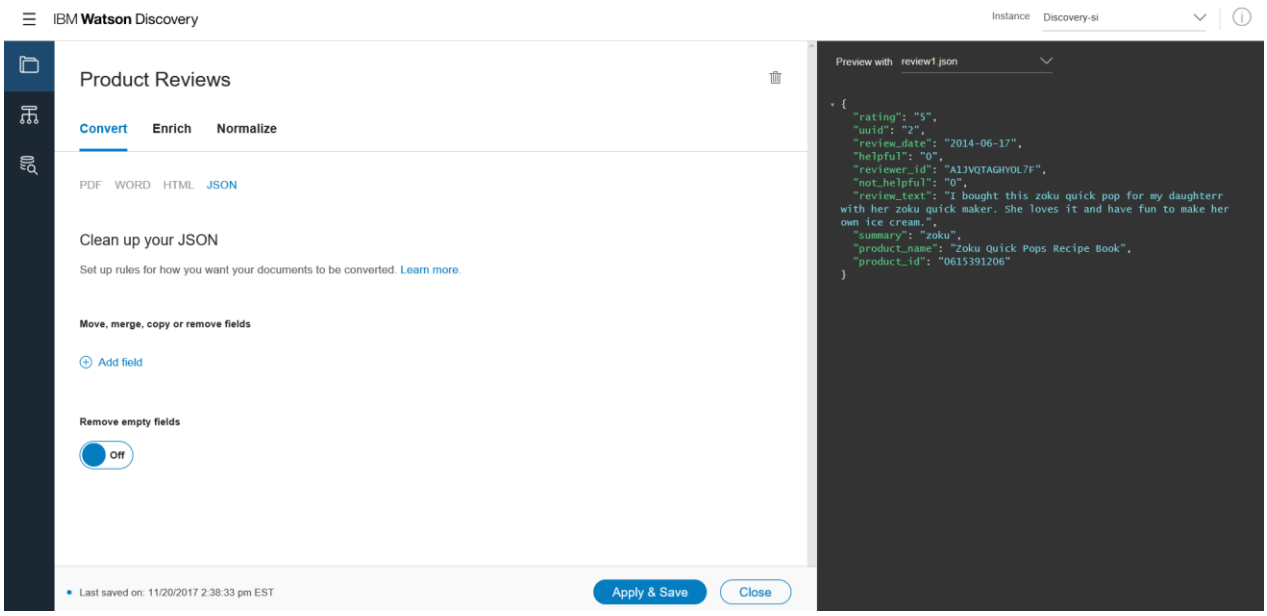


The ingested content (PDF, Word, HTML, JSON) data can be converted to JSON to suit your requirements by removing, copying, moving attributes.

1. Click on the Edit button next to "Product Reviews" in the collection dashboard.



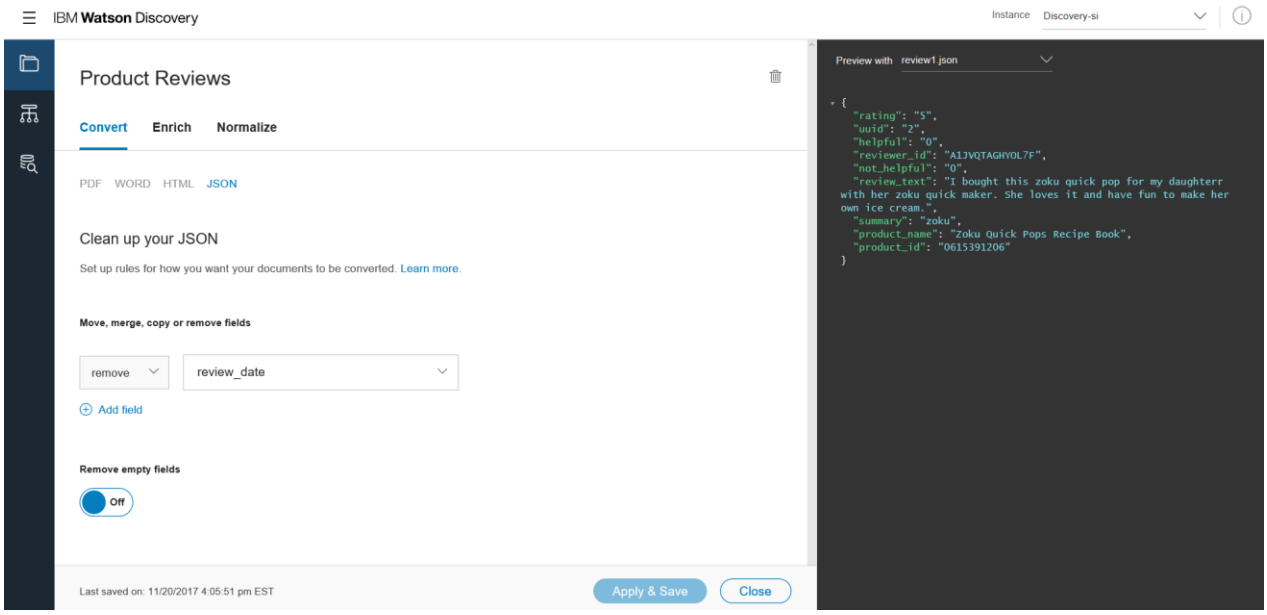
2. In this lab scenario, we are already ingesting JSON data and there are no attributes that need to be modified within the Amazon Product Review data. We are therefore going to walk through some sample steps to illustrate how the functionality works. Start by selecting 'Convert" tab from "Product Reviews" configuration page



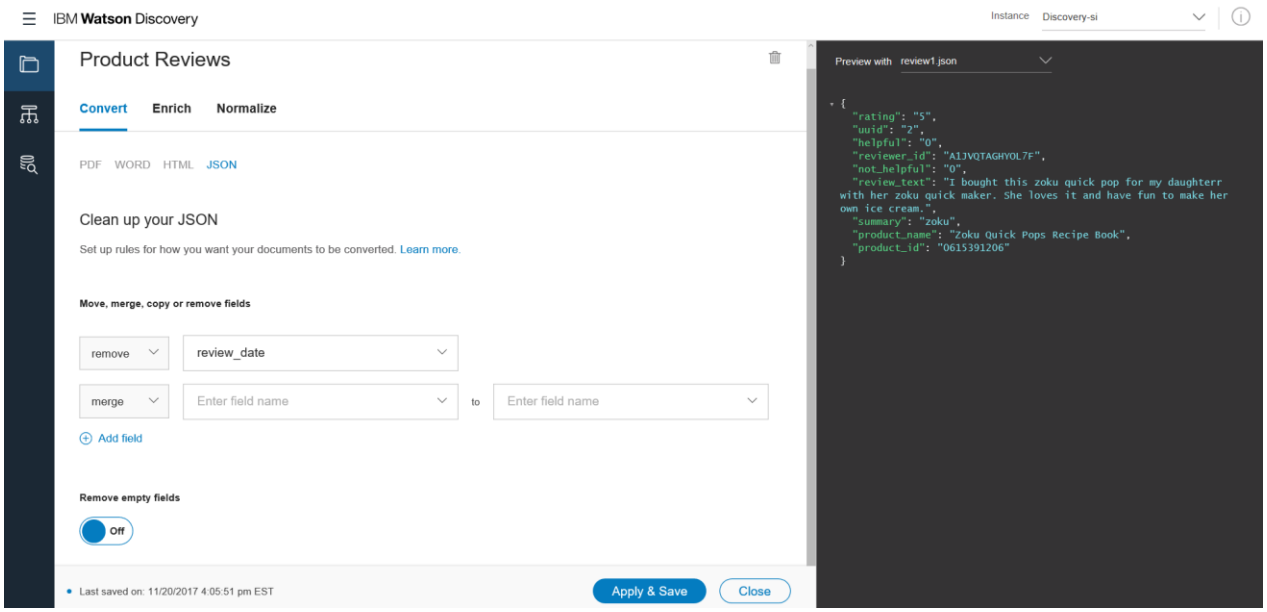
3. Click on "Add field" link under "Move, merge, copy or remove fields" section. You will get a drop-down menu with the following options: remove, copy, move or merge the attributes.

4. Select "remove" from the dropdown and select "review_date" field from the dropdown next to it (for this example we are assuming that review_date field is not needed to satisfy the use cases)

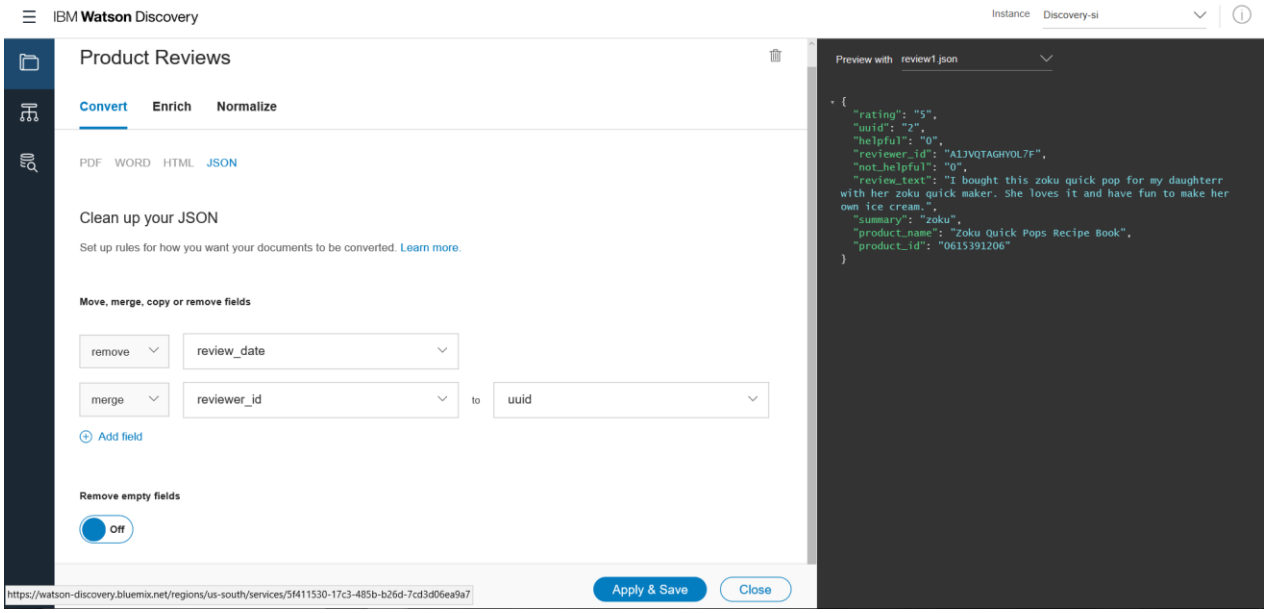
5. Click "Apply & Save" button. You could see in the right window that the attribute "review_date" is removed from review1.json



6. To see an example of how the merge functionality works, click the "Add field" link (underneath Move, merge, copy or remove fields section) and select "merge" from the dropdown menu.

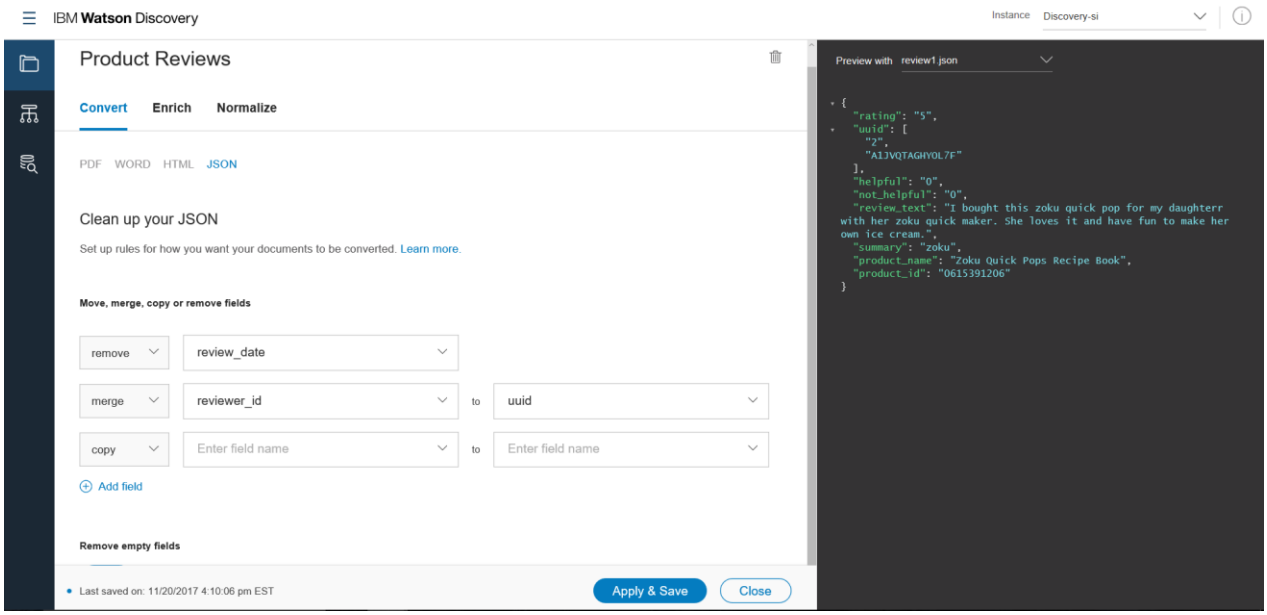


7. Now you will have the option to select two fields: each representing the fields that you want to merge with one another. Select “reviewer_id” from the first dropdown menu and “uuid” from the second dropdown menu.

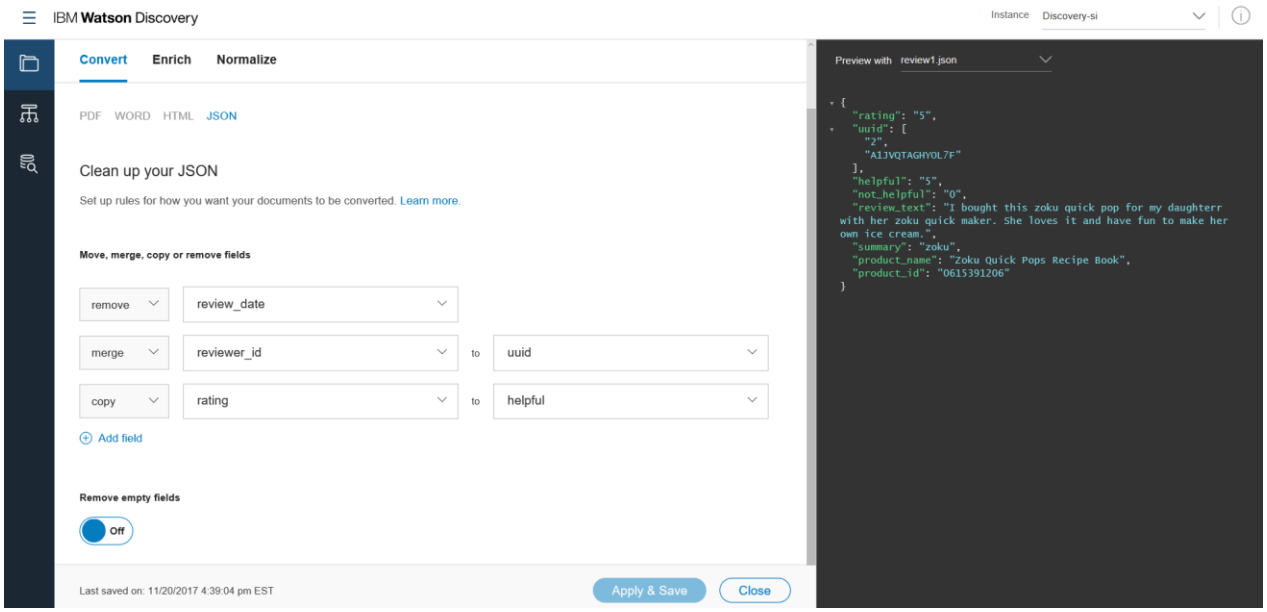


8. Click “Apply & Save” button. In the right window, you will see that the “reviewer_id” field is removed and the data field is merged into “uuid” field.

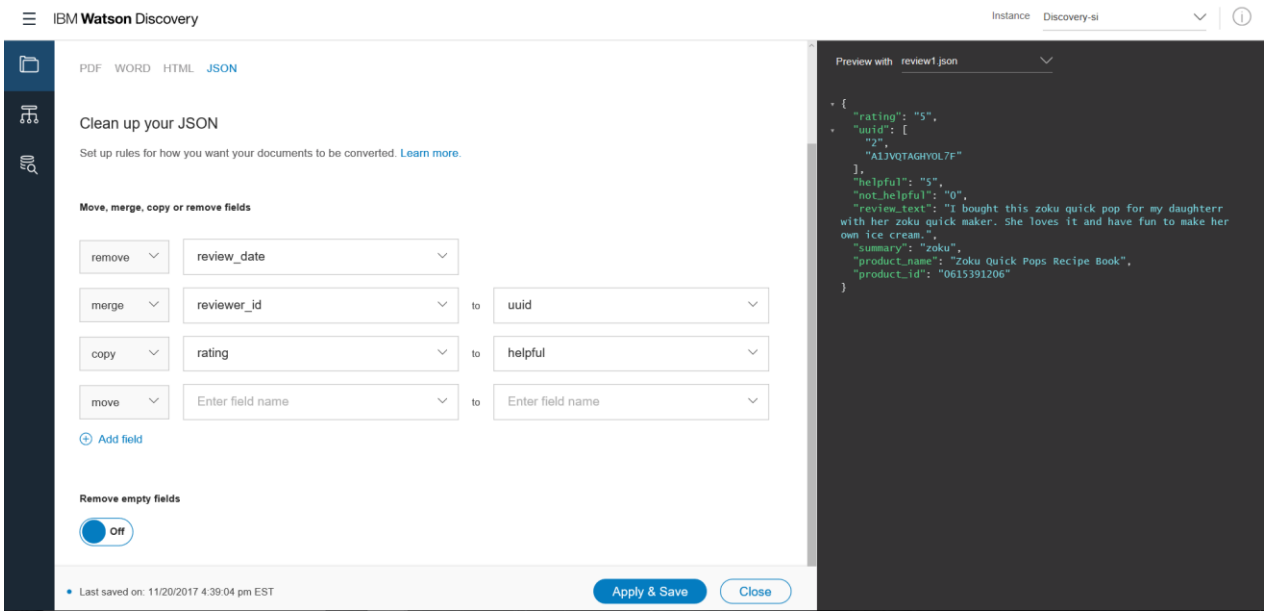
9. You can also copy one data field to another field. To do so click the “Add field” button and select “copy” from the dropdown menu.



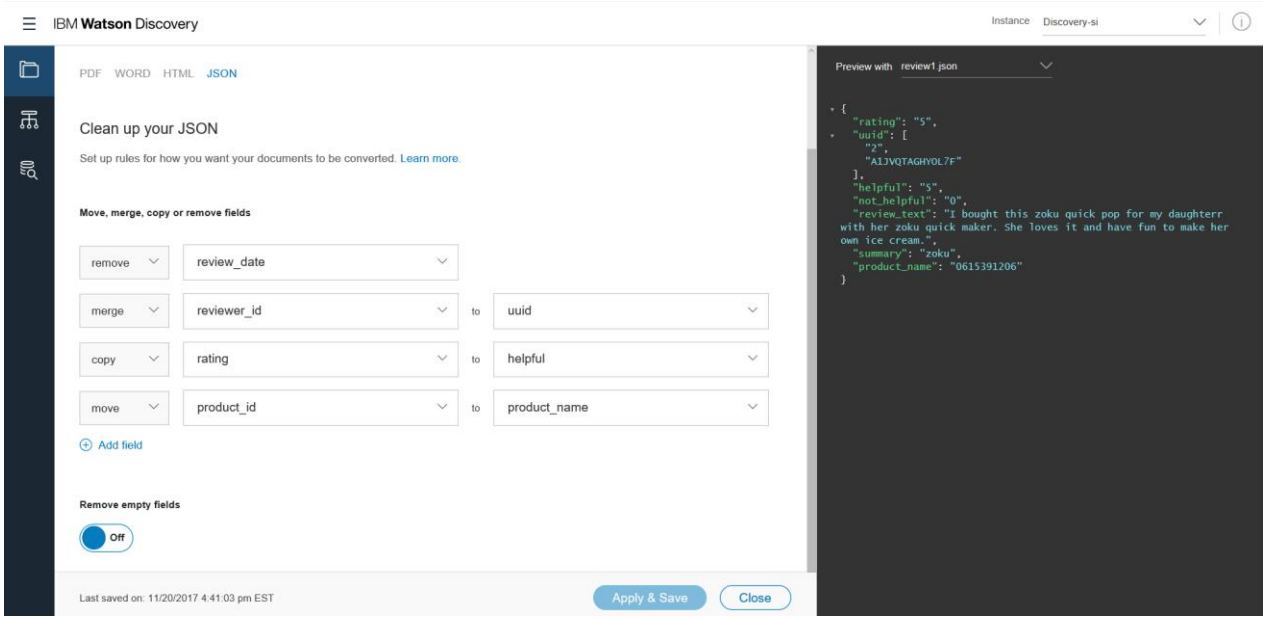
10. Select "rating" field from the first dropdown menu and the "helpful" field from the second dropdown menu. Click "Apply & Save" to see that the "rating" field is copied into the "helpful" field



11. And finally, you can try moving one data field to another field. Click the "Add field" button and select "move" from the dropdown menu.



12. Select the "product_id" field from the first dropdown menu and the "product_name" field from the second dropdown menu. Click "Apply & Save" to see that the contents of "product_id" field replaces the content within the "product_name" field and the "product_id" field content is removed from review1.json.

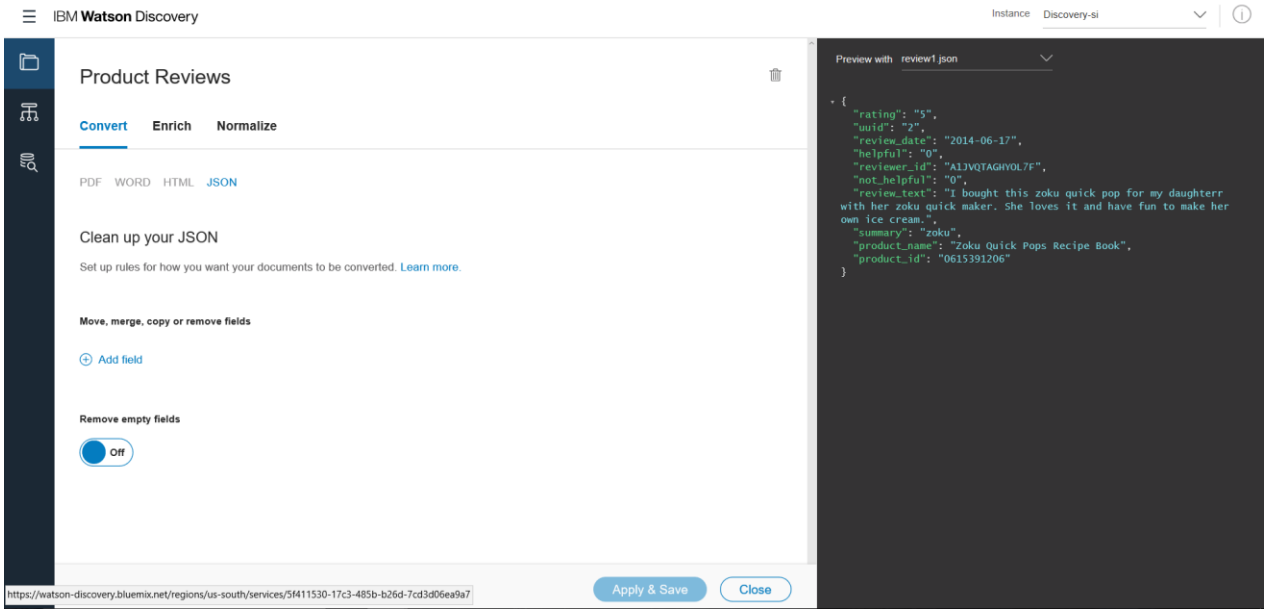


13. Any changes made in "Convert" tab attributes are reflected on all the data files ingested.

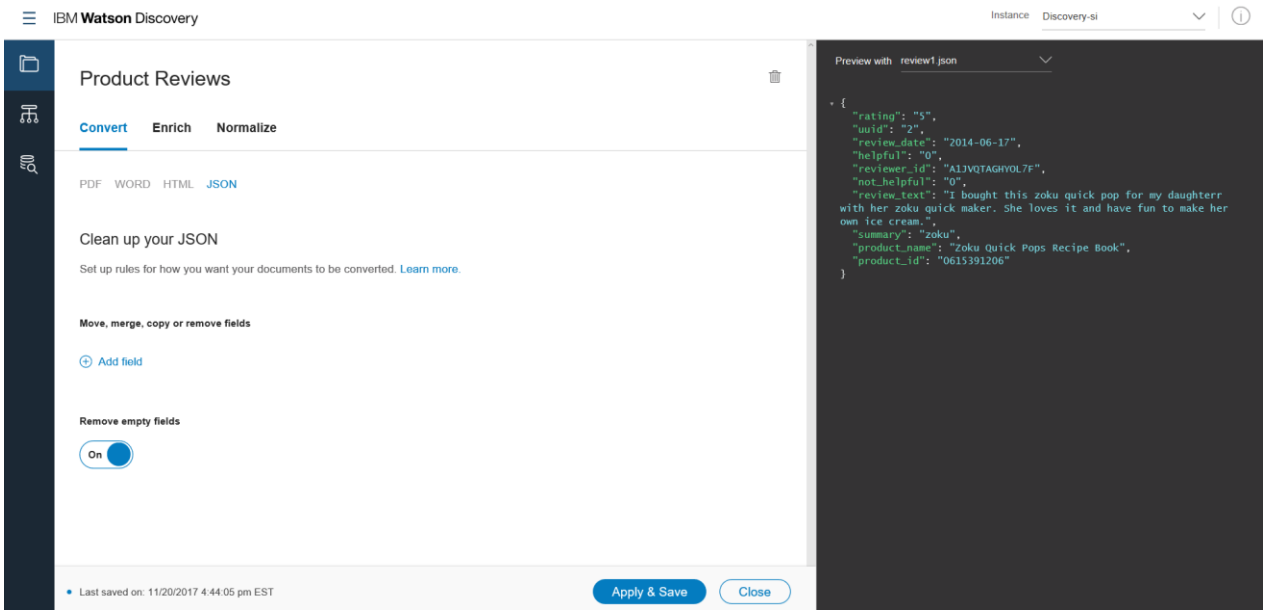
    Now that you have learned how to configure conversion attributes, you can remove all the changes. Remember, the Amazon Product Review data does not need any conversion attributes for this Lab.

14. You should revert all the changes you made in Convert section. To delete the Convert configuration, click on "minus" button on far right of each field added for conversion. The minus sign is displayed when you scroll over the field with your mouse.

15. After removing all the conversion attributes, click the "Apply & Save" button to return to the original review1.json. Make sure your preview file looks like what is shown in the following snapshot.



16. Additional item for consideration. By switching the "Remove empty fields" switch to "on", any empty fields from the JSON data will be removed during conversion of the ingested documents.

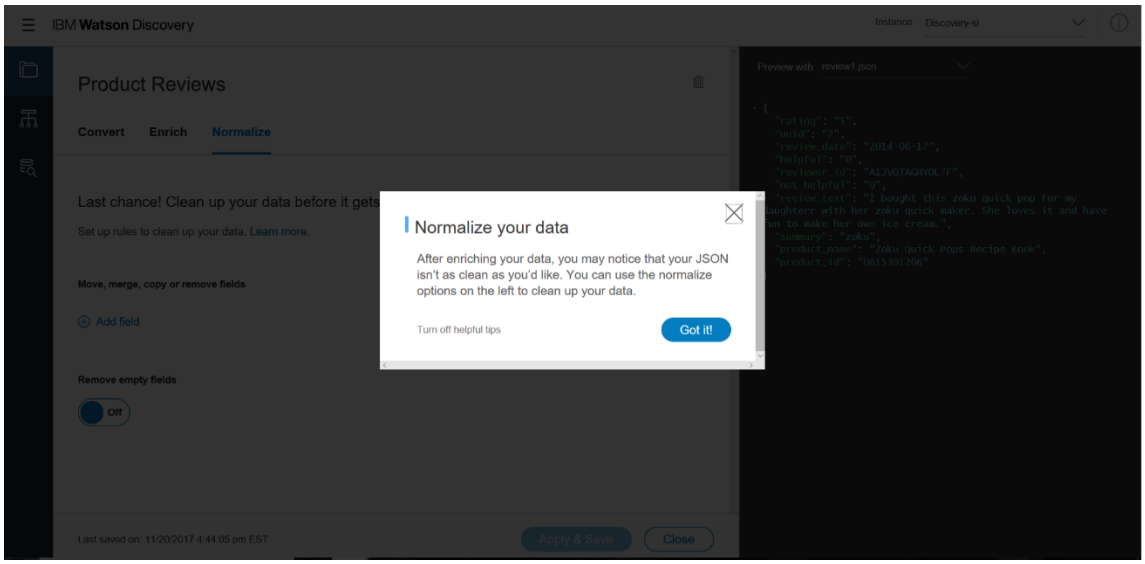


More details about converting attributes can be found by clicking "Learn more" link under "Clean up your JSON" header

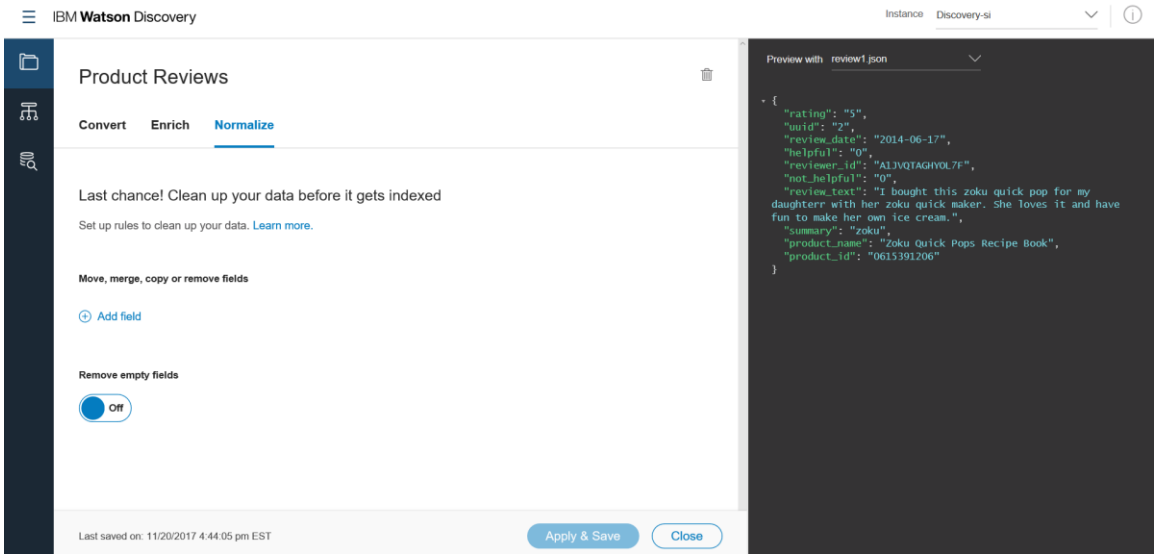
2.1.3 Configuring Normalization Attributes (Optional for this lab)

Attributes normalization follows a similar process to that of conversion However, conversion attributes are only applied to the original data field in the JSON file. Normalization is performed for both the original and enriched data field. It is useful when you want to remove certain enrichment attributes, copy them to another field etc.,

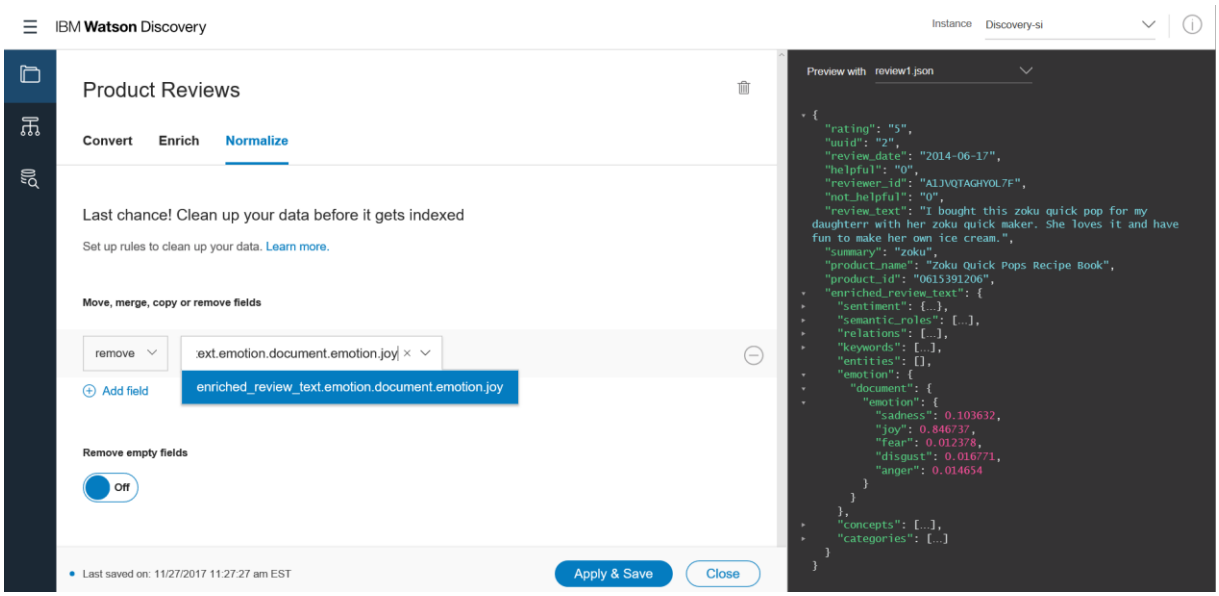
1. Start by selecting 'Normalize" tab from "Product Reviews" configuration page



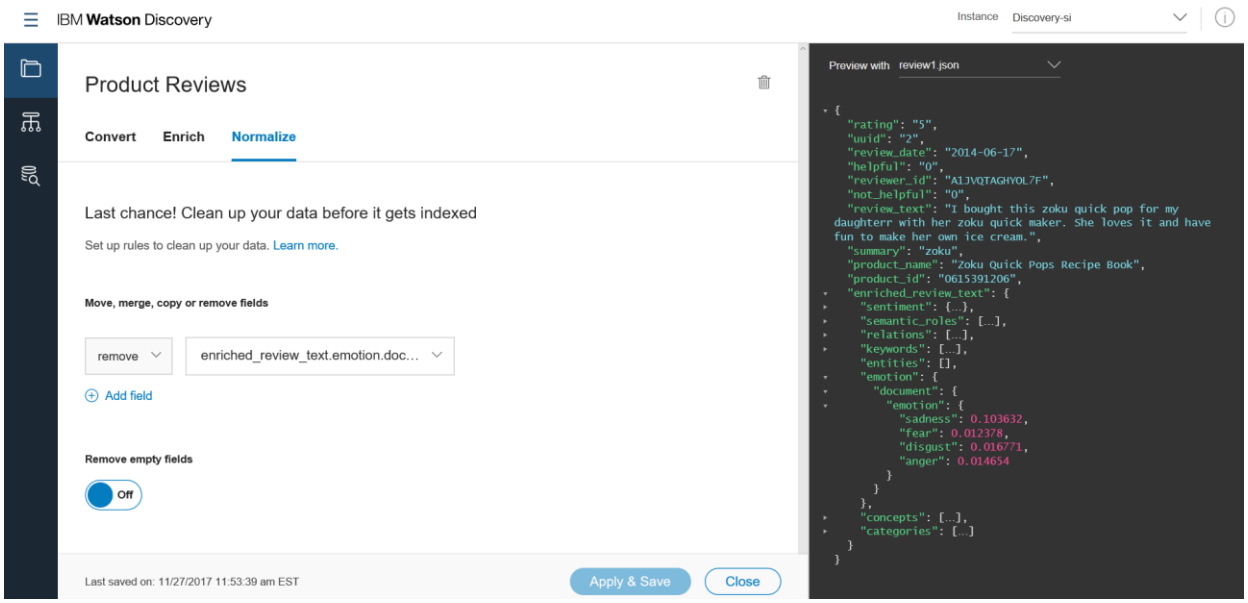
2. Click on "Add field" link under "Move, merge, copy or remove fields" section. You will get a drop-down menu with the following options: remove, copy, move or merge the attributes.



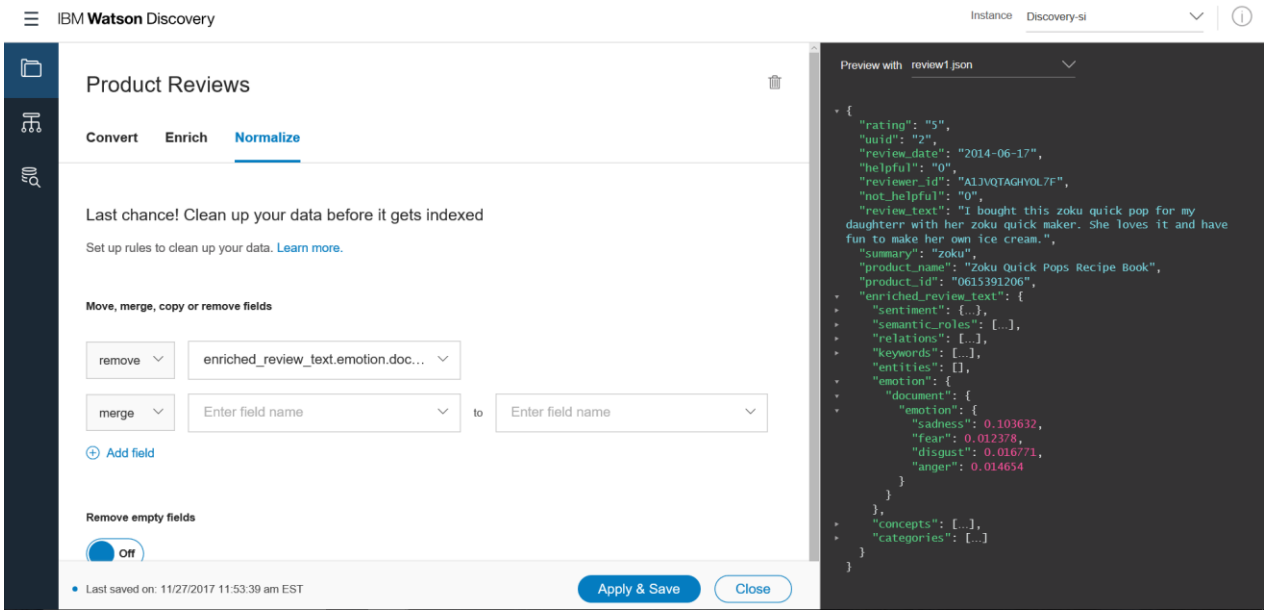
3. Select "remove" from the dropdown and select "enriched_review_text.emotion.document.joy" field from the dropdown next to it (for this example we are assuming that the joy field is not necessary for the emotional analysis)



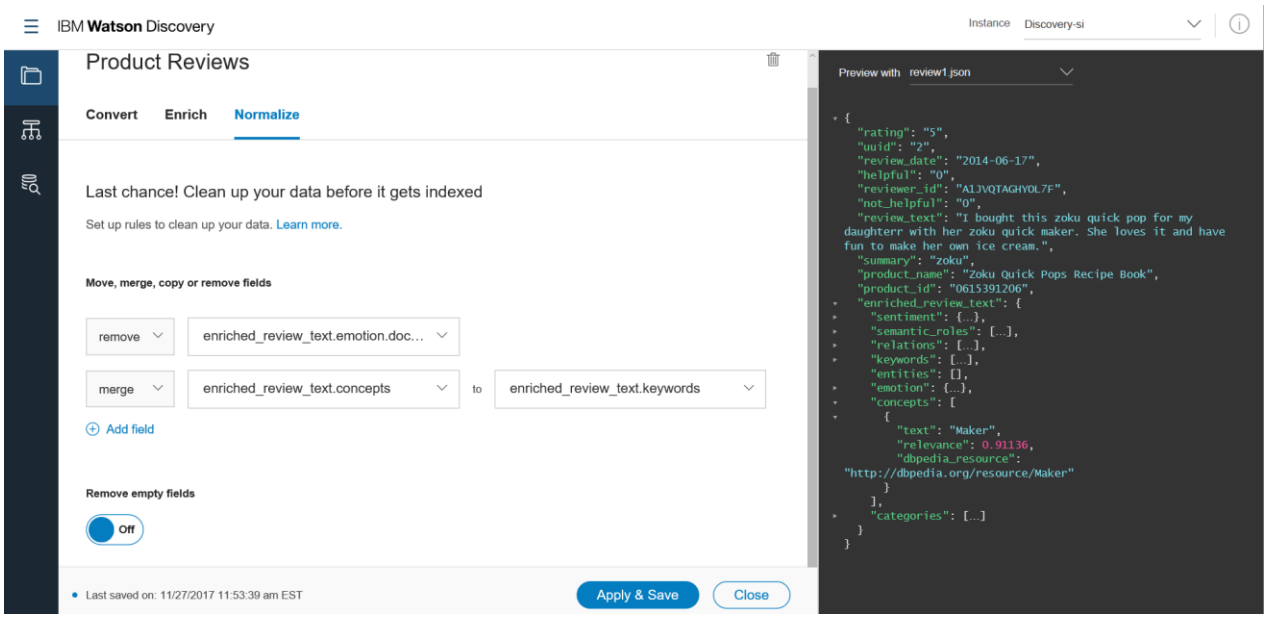
4. Click "Apply & save" button. In the right window, you can see that the "joy" field under enriched_review_text.emotion.document.emotion is removed



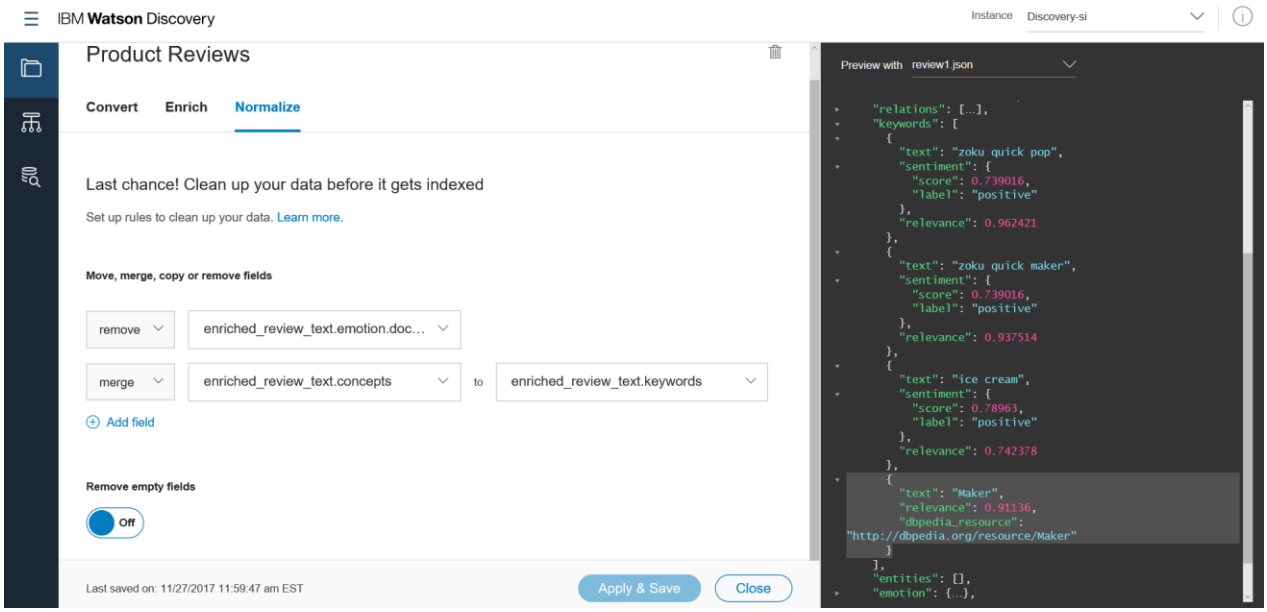
5. To see an example of how the merge functionality works, click the "Add field" link (underneath Move, merge, copy or remove fields section) and select "merge" from the dropdown menu.



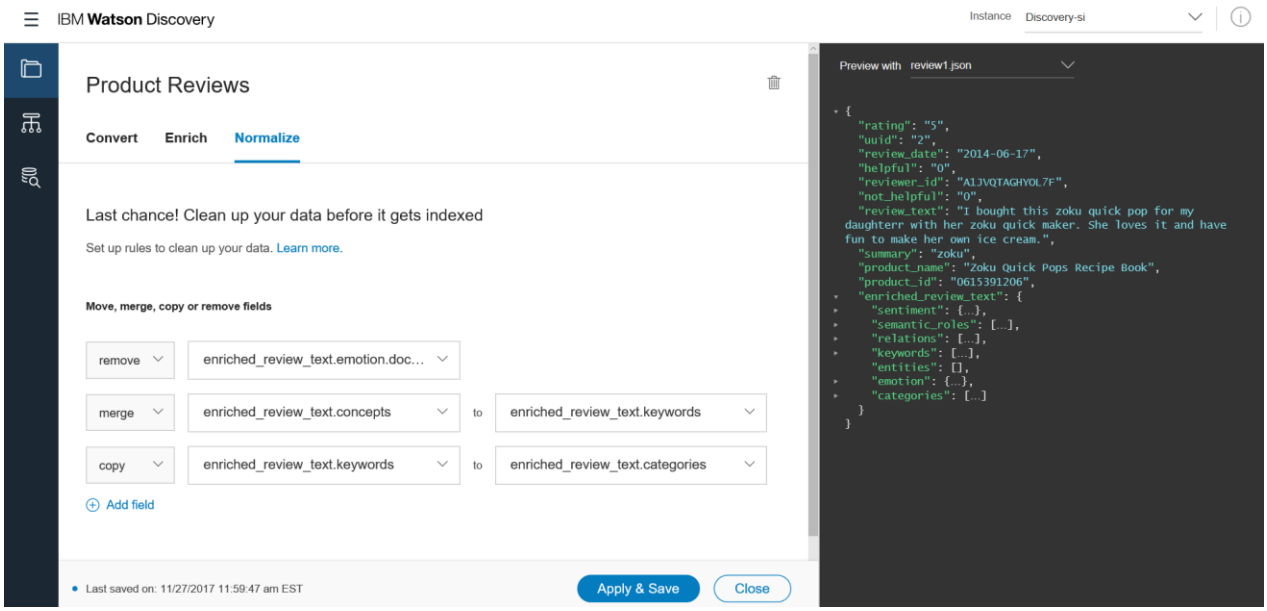
6. Now you will have the option to select two fields: each representing the fields that you want to merge with one another. Select "enriched_review_text.language" from the first dropdown menu and "enriched_review_text.status" from the second dropdown menu.



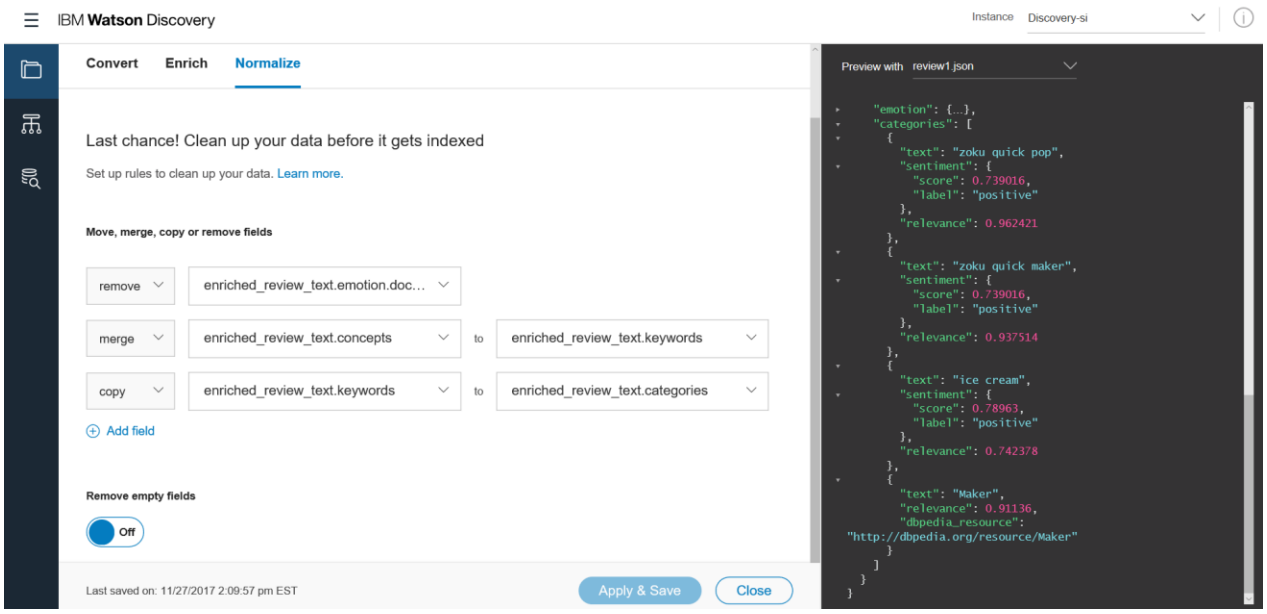
7. Click "Apply & Save" button. In the right window, you will see that the "enriched_review_text.concepts" field is removed and the data field is merged into "enriched_review_text.keywords" field.



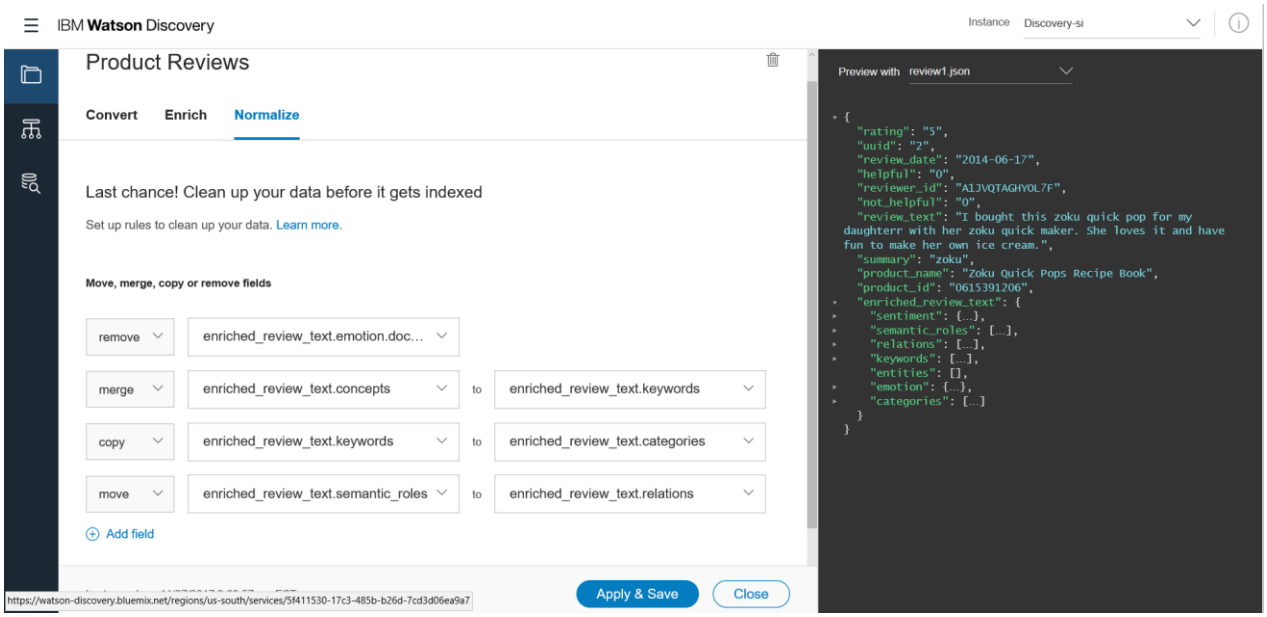
8. You can also copy one data field to another field. To do so click the “Add field” button and select “copy” from the dropdown menu. Select “enriched_review_text.keywords” field from the first dropdown menu and the “enriched_review_text.categories” field from the second dropdown menu



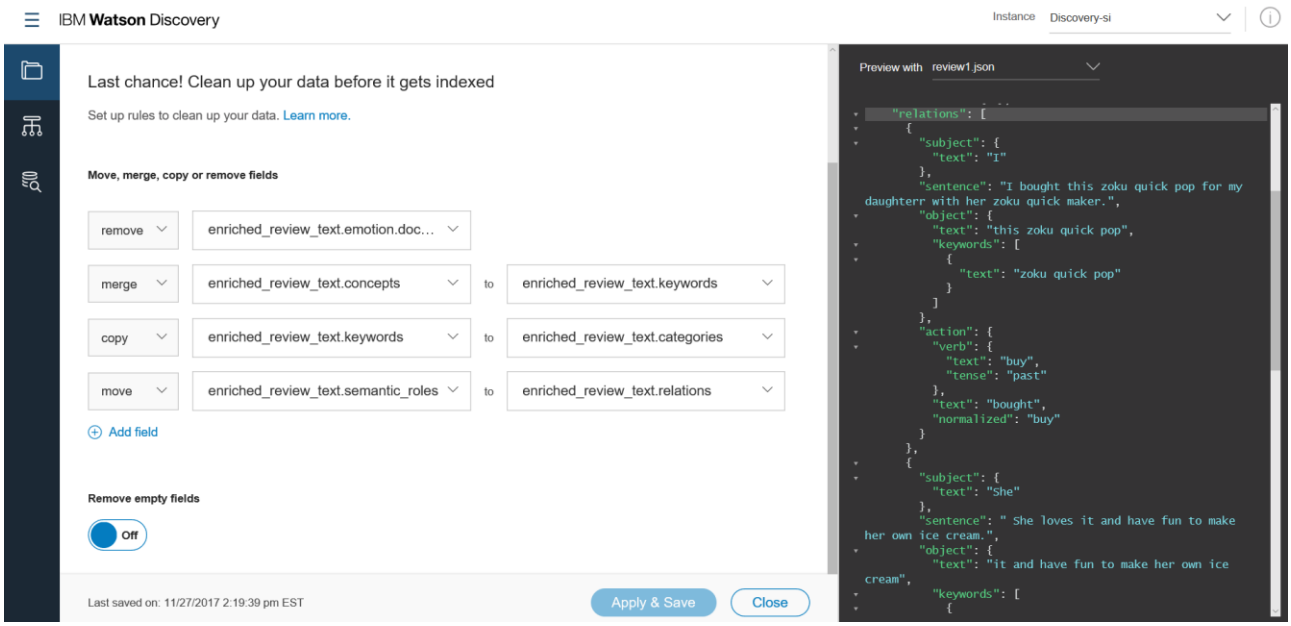
9. Click “Apply & Save” to see that the “enriched_review_text.keywords” field is copied into the “enriched_review_text.categories” field



10. You can try moving one data field to another field. Click the “Add field” button and select “move” from the dropdown menu. Select the “enriched_review_text.concepts.freebase” field from the first dropdown menu and the “enriched_review_text.concepts.opencyc” field from the second dropdown menu.



11. Click “Apply & Save” to see that the contents of “enriched_review_text.semantic_roles” field replaces the content within the “enriched_review_text.relations” field and the “enriched_review_text.semantic_roles” field content is removed from review1.json



12. You can remove the normalization attributes using the “button” on to the far sight of the fields
13. Remove the normalization attributes since it is not required for Amazon use case lab
14. You can click on Close to go back to Your data tooling homepage.
15. Once the configuration setup is completed, users are now ready to upload the Amazon Product Review data files.
16. If you are a developer, please refer to section 3 to learn how to use data crawler to upload files.
17. If you are a general user, please refer to section 4 to learn how to use tooling interface to upload files.
18. Users can either use crawler tool or tooling interface but not both.

3 Data Ingestion using data crawler

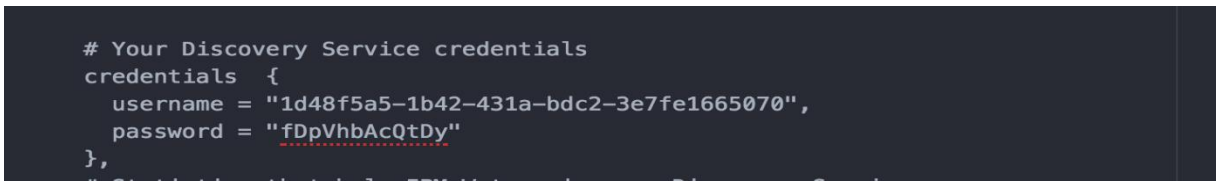
Data crawler can either upload data files located in the local file system to the discovery collection or crawl data from the data base and upload them as JSON to the collection. For this lab, we will walk through the steps of crawling data from your local file system.

3.1 Data Crawler Overview

The data crawler is useful when you have a directory or database with a large number of documents to ingest. Data crawler can be executed from command line and be configured with ease. Data crawler defaults to run 14 concurrent threads to speed up upload time. For this lab, we assume users have followed the prerequirements documents and have the data crawler already installed. If that is not the case, check the IBM Cloud documentation for instructions to install the data crawler. The documentation can be found at https://www.ibm.com/watson/developercloud/doc/discovery/data-crawler-install.html

3.2 Data Crawler Configuration

1. Locate the 'crawler-x.x.x/share/examples/config' folder in your machine
2. Copy the 'crawler-x.x.x/share/examples/config' folder into the same directory as the Amazon Product Review content folder.
3. Once copied to the new location, open the copied folder.
4. Open the config/discovery/discovery_service.conf file. Modify the discovery_service.conf file by copying and pasting your Discovery Service credentials (password and username) to the #Your Discovery Service credentials section of the file.
5. As a reminder, your Discovery Service credentials can be found in the Service credentials tab in the same window that you used to open the WDS Tooling.



6. You will also need the Environment ID, Collection ID and Configuration ID for the “Amazon Product Reviews” collection you saved in Section 2. If you missed that step, the IDs can always be retrieved from the collection dashboard.
7. Copy and paste the Environment ID, Collection ID, and Configuration ID within their respective fields identified with “#” (e.g. #The Collection in which documents should be indexed. Collection id = ...)

```
# The environment to use.
environment_id = "c90e85d6-691a-4b9c-b1c0-aba9a75c3dd6",

# The collection in which documents should be indexed.
collection_id = "9ec6362e-ae84-4d51-96c1-27a62697c2b2",

# The conversion configuration to use.
configuration_id = "91e7b8a8-3ed2-4cbe-9ca3-3b76e6bd5675",

# Or you can provide a file with your config
#configuration = "file://PATH_TO_CONFIG"
# Check if the document was successfully processed at the endpoint.
# This will reduce the perceived performance of the crawler but will
# produce reliable notification of a successful document upload and conversion.
# When enabling this, it is wise to increase concurrent_upload_connection_limit
# to your WDS ingestion conversion concurrency limit in order to fully utilize
# the resources available to you.
check_for_completion = false

# Your Discovery Service credentials
credentials  {
  username = "1d48f5a5-1b42-431a-bdc2-3e7fe1665070",
  password = "fDpVhbAcQtDy"
},
```

8. Save and exit out of discovery_service.conf file.

9. Open the crawler.conf file located in the top directory of the configuration folder. Under the output_adapter section, make sure that class is set to "**com.ibm.watson.crawler.discoveryserviceoutputadapter.DiscoveryServiceOutputAdapter**" and the config is set to "discovery_service"

```
output_adapter {
  # Select the Output Adapter to use by specifying its class. See crawler.conf(5) for more information or read below.
  class = "com.ibm.watson.crawler.discoveryserviceoutputadapter.DiscoveryServiceOutputAdapter",

  # Select the configuration key to be passed to the Output Adapter. The string here must correspond to a key within this configuration object.
  config = "discovery_service",
```

10. Save and close the crawler.conf file.

11. Navigate to the seeds directory folder within your configuration folder.

12. Open config/seeds/filesystem-seed.conf

13. Modify the value attribute directly under:name="url" attribute to the file path of the Amazon Product Review folder that you want to crawl. This url will be specific to the location in which you stored the content. Example: value="sdk-fs:///Users/chethan/Downloads/WDS/Crawler/data/" Note that the third slash is needed as well as a slash at the end of the directory url (e.g. data/).

```
# The third / is necessary! Do not prepend file:// to the listed paths.
{
  name ="url",
  value="sdk-fs:///Users/chethan/Downloads/WDS/Crawler/data/"
},
{
  name = "hops"
```

14. Save and exit the file.

15. Open config/crawler.conf and make sure the crawl_config_file is set to "**connectors/filesystem.conf**" and the crawl_seed_file is "**seeds/filesystem-seed.conf**"

```
# The configuration for the Connector Framework.
connector_framework = {
  # The file containing the connector configuration for the crawl.
  crawl_config_file = "connectors/filesystem.conf",

  # The file containing the seed URL for the crawl.
  crawl_seed_file = "seeds/filesystem-seed.conf",

  # The file containing the encryption key used for decrypting passwords.
  # Generate a new key with the `vcrypt` command included with Crawler.
  id_vcrypt_file = "id_vcrypt",
```

3.3 Uploading Data to WDS collection

1. Open terminal/command line and navigate to top level folder that contains the Data Crawler config folder

2. Run the command **crawler testit** to verify that crawler is crawling the URL as expected

   Ignore the error message "*Could not find path configuration in the Discovery config. There was an error trying to find required path in the Discovery config*". The testing is successful if the command executes without any error other then the mentioned one. If this error message does not show – please get the attention of the instructor.

3. Run the command **crawler crawl** to push your files to discovery service

   You will get the same error message that you got while testing crawler. If there is no other error message and the commands executes successfully, then you can see that the documents are uploaded to the collection. You can revert back to the Watson Discovery Service tooling interface.

4. You can see the number of documents uploaded by accessing the collection dashboard from the tooling interface. This step will be completed in 2-5 minutes.

5. The crawling will come to a halt. You can now exit from the terminal.

4 Data Ingestion using Discovery Tool and API calls

Data files can also be added to the collection using either the discovery tool or the API calls. Before uploading any documents, it is suggested that you create configuration and add enrichment attributes into the configuration (this should have already been completed in Section 2). Any data files uploaded before configuration updates will not reflect the changes.

4.1 Adding documents to the collection

The system detects the type of the document being uploaded and will only accept the following types:

i. text/html (.html)

ii. application/pdf (.pdf)

iii. application/JSON (.JSON)

iv. application/msword (.doc)

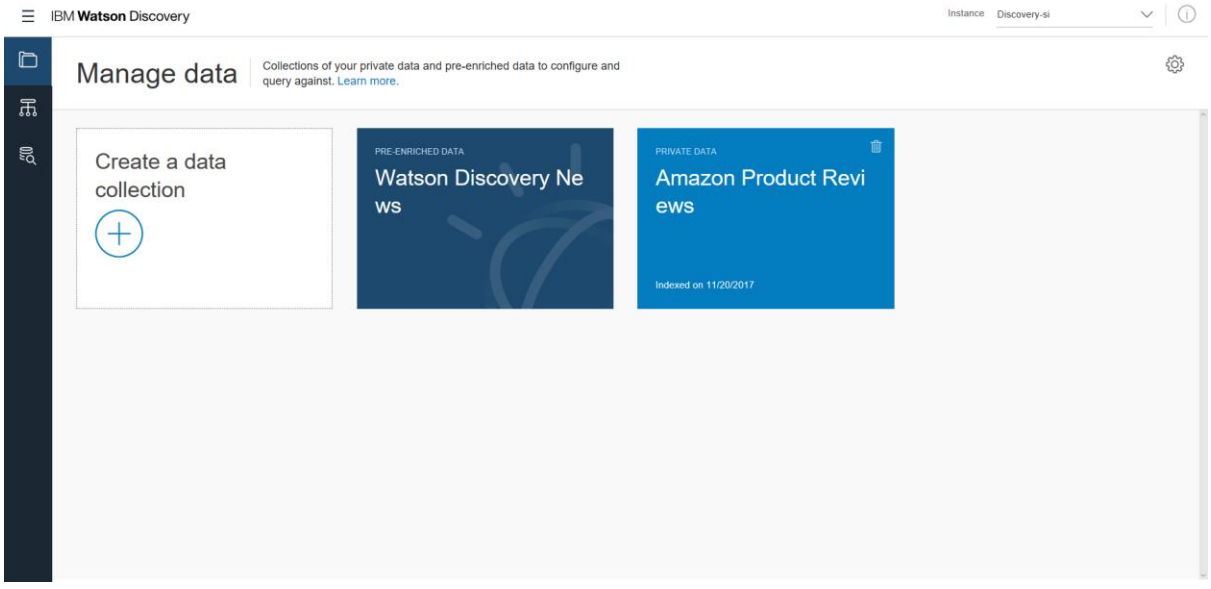
v. application/vnd.openxmlformats-officedocument.wordprocessingml.document (.docx)

vi. application/xhtml+xml (.xhtml)

4.1.1 Uploading documents using the Tooling interface

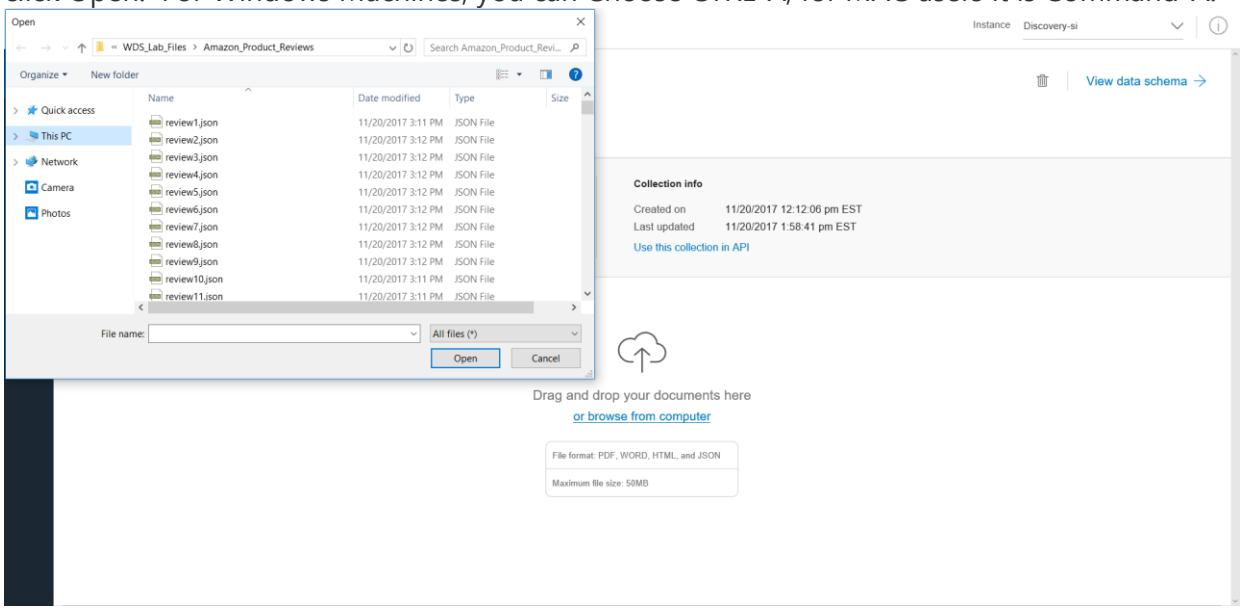
The tooling should not be used to upload large amounts of data as it is single threaded, but it is a good way to get started when you have smaller sets of data.

1. Open the Amazon Product Reviews collection from Your Data tooling homepage by clicking on the Amazon Product Reviews light blue tile.



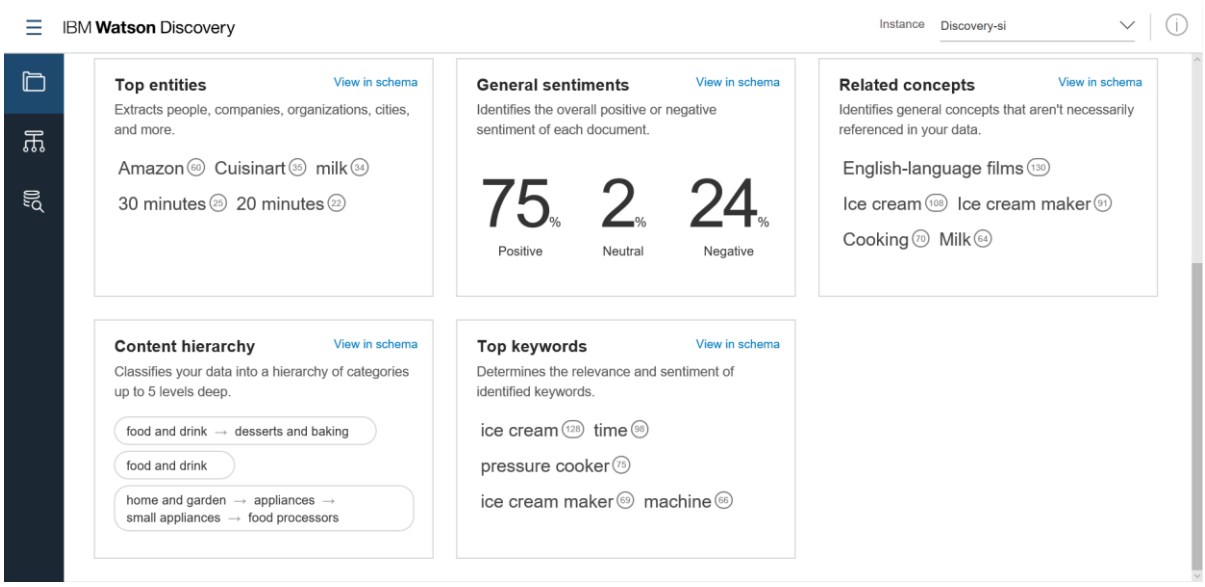
You can ingest content by either dragging and dropping your document in the "Add data to this collection" box or by clicking "browse from computer".

2. Click on 'browse from computer", navigate to the Amazon Reviews Content folder (this was part of the Zip folder which you have previously downloaded in Section 2), select all Amazon Reviews, and click Open. For Windows Machines, you can Choose CTRL-A, for MAC users it is Command-A.



3. Once you click Open (or drag and drop depending on the method you used), the tooling will begin the ingestion process. It will take 5-10 minutes to upload the Amazon Review content to your collection.

4. Once the selected documents are uploaded successfully, you will receive a popup saying, "Documents Uploaded". You can also refresh the page (by pressing Ctrl + R for Windows or Command + R for Macs) to see the number of documents available, processing and the failed uploads.



4.1.2 Process to Upload documents using API calls (Optional)

Please note: this is not a step to be performed during the lab. It is additional information that users could find helpful. Developers can ingest documents using direct API calls. Documents can be uploaded into the collection using Node JS, python, java or any other programming language. The client should know the file handling process and creating HTTP post request process before trying out the API calls method

1. To upload data using the Node JS application, use the following code

```
var watson = require('watson-developer-cloud');
var fs = require('fs');

var discovery = new DiscoveryV1({
  username: '{username}',
  password: '{password}',
  version_date: '2016-12-01'
});

var file = fs.readFileSync('{/path/to/file}');

discovery.addDocument(('{environment_id}', '{collection_id}', file),
function(error, data) {
  console.log(JSON.stringify(data, null, 2));
});
```

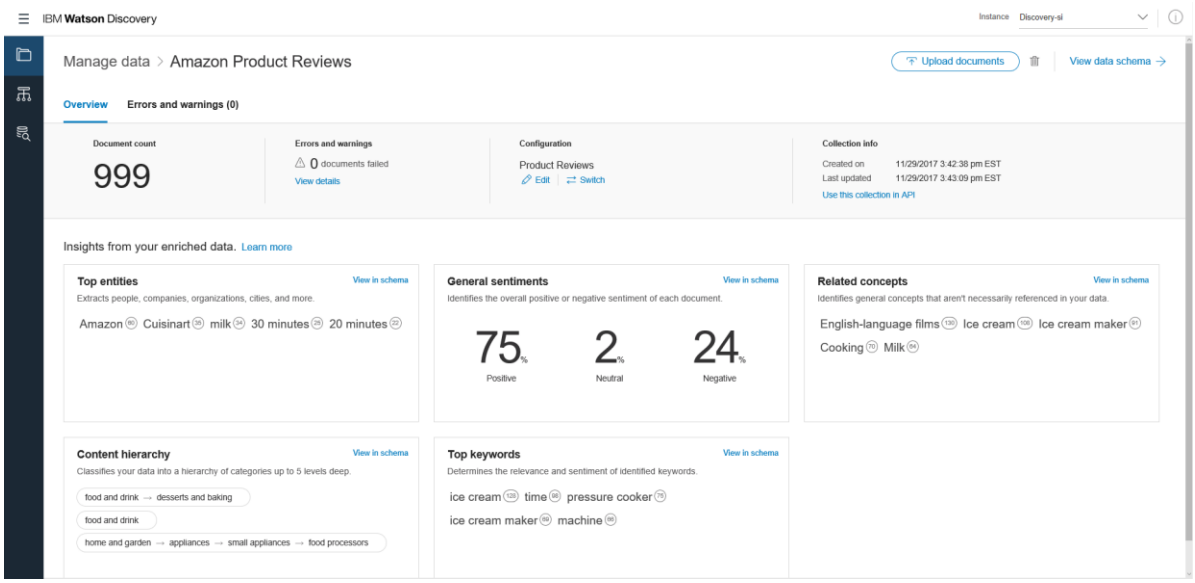
2. If you are using java or C++ make sure you do not set the document type to “multipart/form-data” as you will encounter the error saying the format is not supported. Instead use the document type which is exactly matching with that of the document format you are uploading. For example, for an html document use “text_html”. A java example to upload document is shown below

```java
package com.ibm.watson.developer_cloud.discovery.v1;

import java.io.FileInputStream;
import java.io.InputStream;

import com.ibm.watson.developer_cloud.discovery.v1.model.document.CreateDocumentRequest;
import com.ibm.watson.developer_cloud.discovery.v1.model.document.CreateDocumentResponse;
import com.ibm.watson.developer_cloud.http.HttpMediaType;

public class WDSDocumentUploader {

    public static void main(String[] args) {

        try {

            Discovery discovery = new Discovery("2016-12-01");

            discovery.setEndPoint("https://gateway.watsonplatform.net/discovery/api");
            discovery.setUsernameAndPassword("Username", "Password");

            InputStream documentStream = new FileInputStream("sample.html");

            CreateDocumentRequest.Builder builder = new CreateDocumentRequest.Builder(
                    "environment id", "collection id");
            builder.inputStream(documentStream, HttpMediaType.TEXT_HTML);
            CreateDocumentResponse createResponse = discovery.createDocument(builder.build()).execute();

        } catch(Exception ex) {

        }

    }

}
```

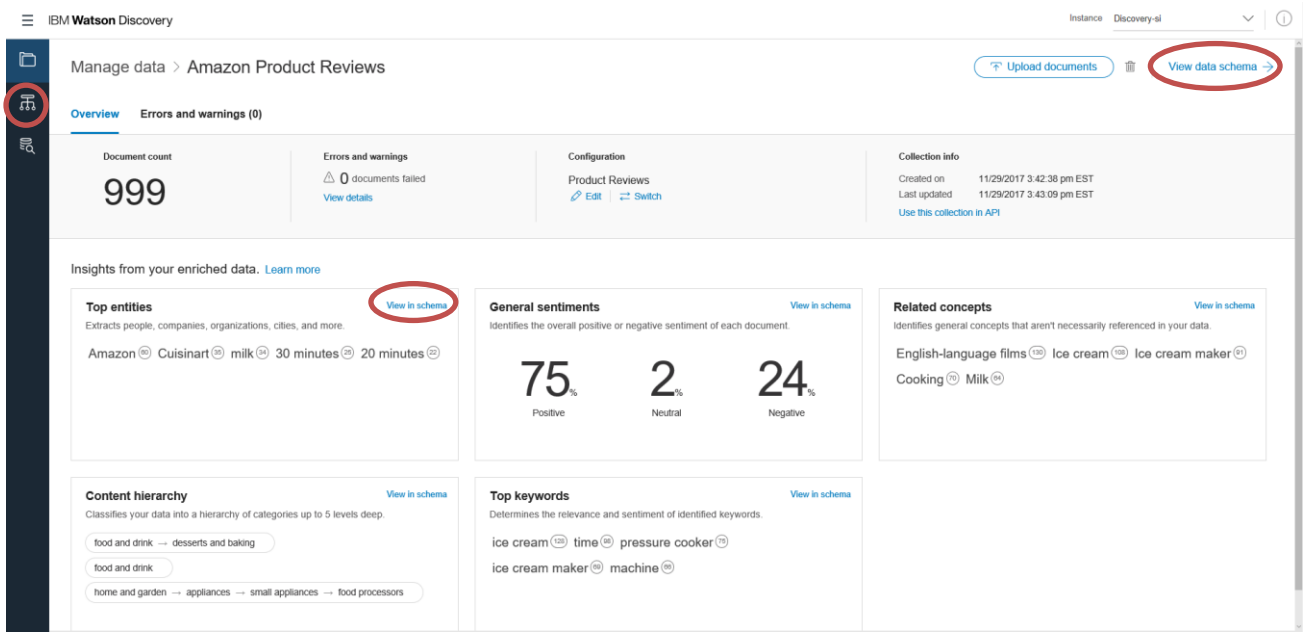
5 Obtaining data insights/Querying the collection

After the discovery instance is configured and amazon review data is uploaded and enriched, the collection will be queried to gain insights into the data. The data schema will also be provided to gain an overall view of the collection and documents. NLP is running behind the scenes to provide clear data insights.

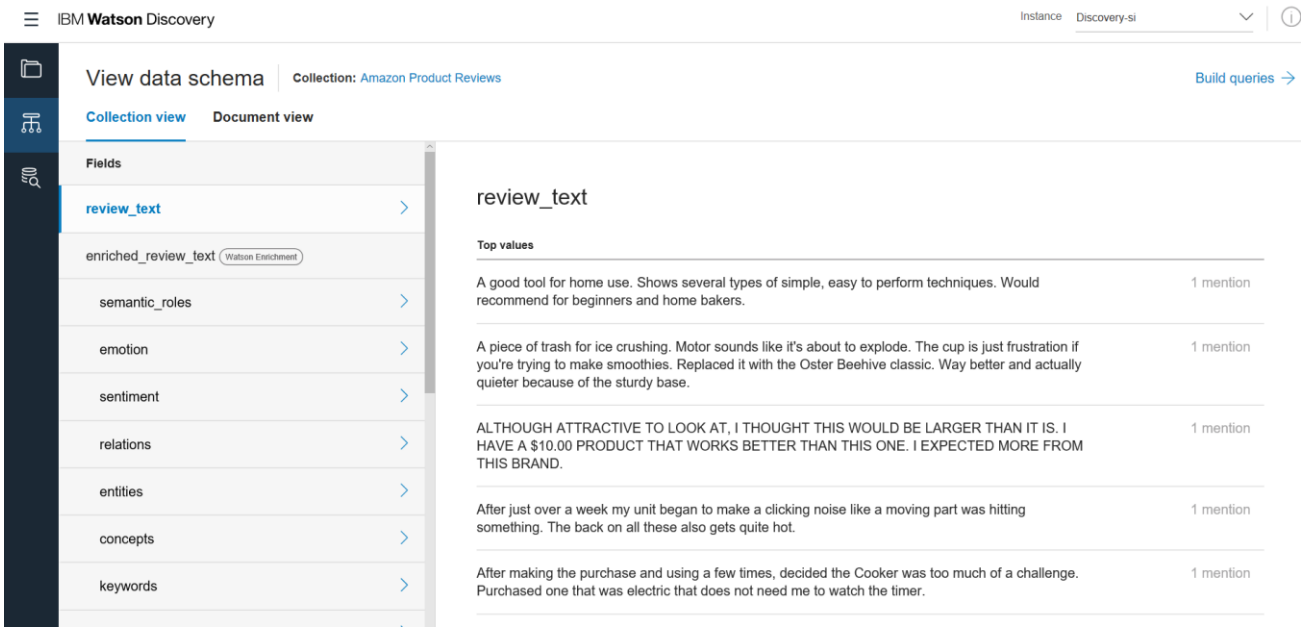
1. In the "Overview" section, you can view the Amazon review analysis results. Data insights obtained for Amazon review data is shown in the below figure.

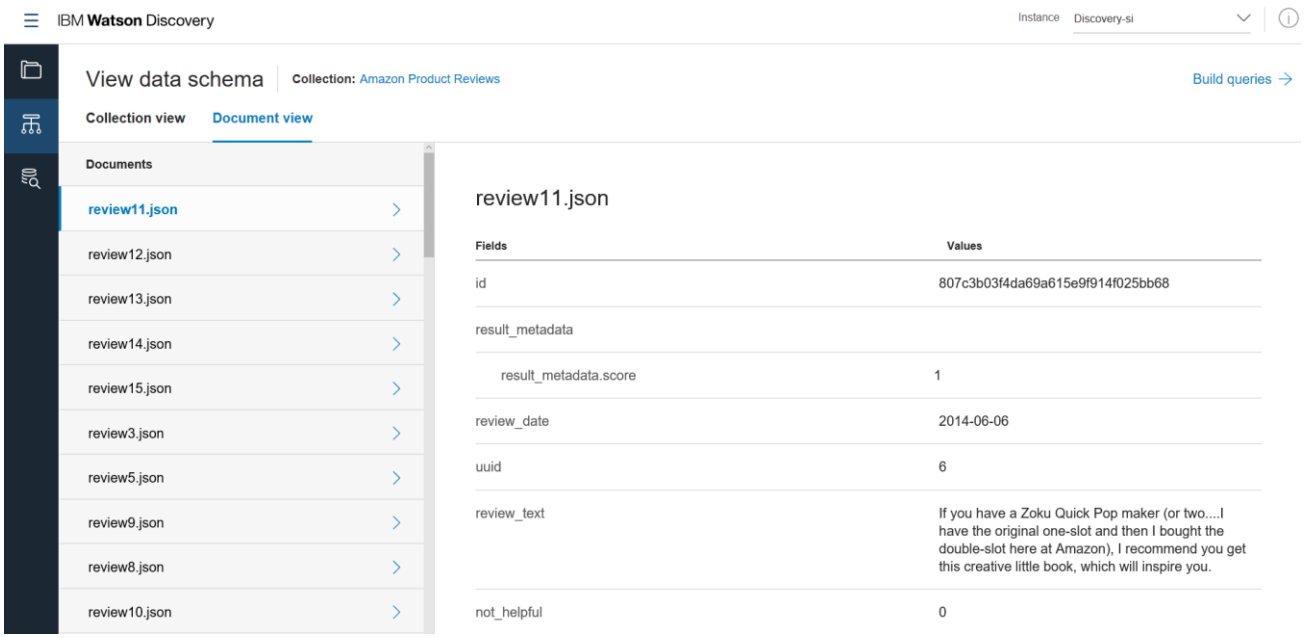


2. Click "View data schema →" in the top right corner of the "Overview" section or in the top right corners of any of the data insight boxes for exploration purposes of the collection and documents. You also can click the data schema icon on the lefthand side bar, but if you do that, you must reselect the collection you are working in which is "Amazon Product Reviews." The links are circled in red below.

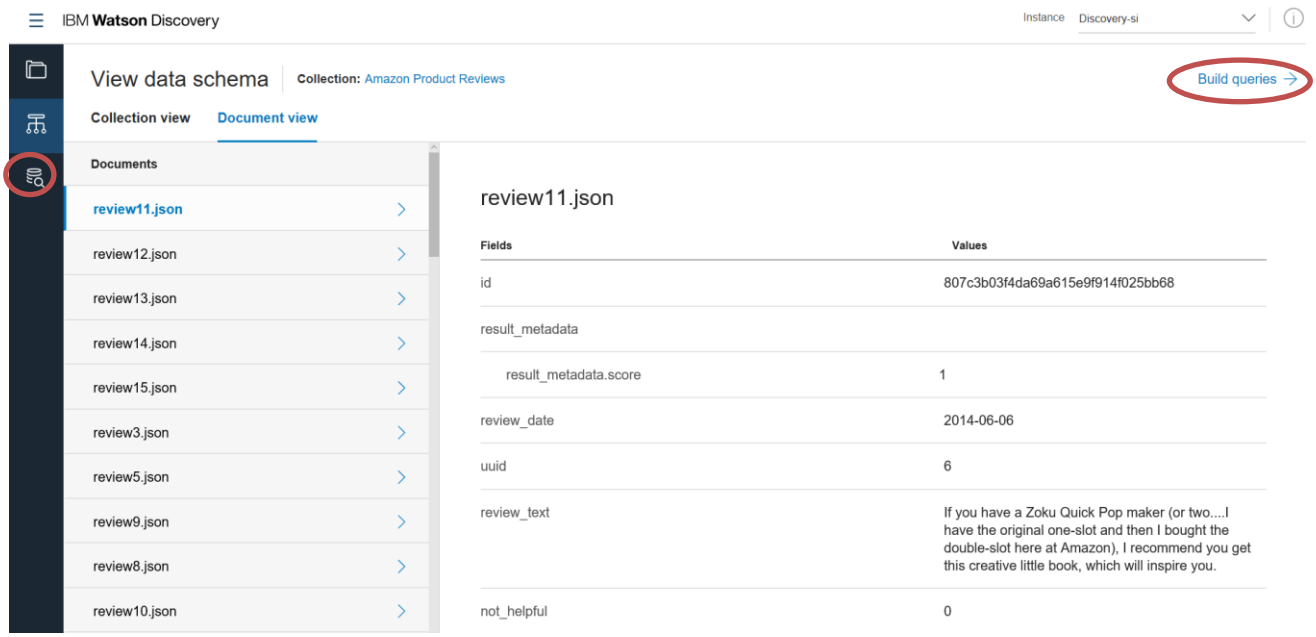


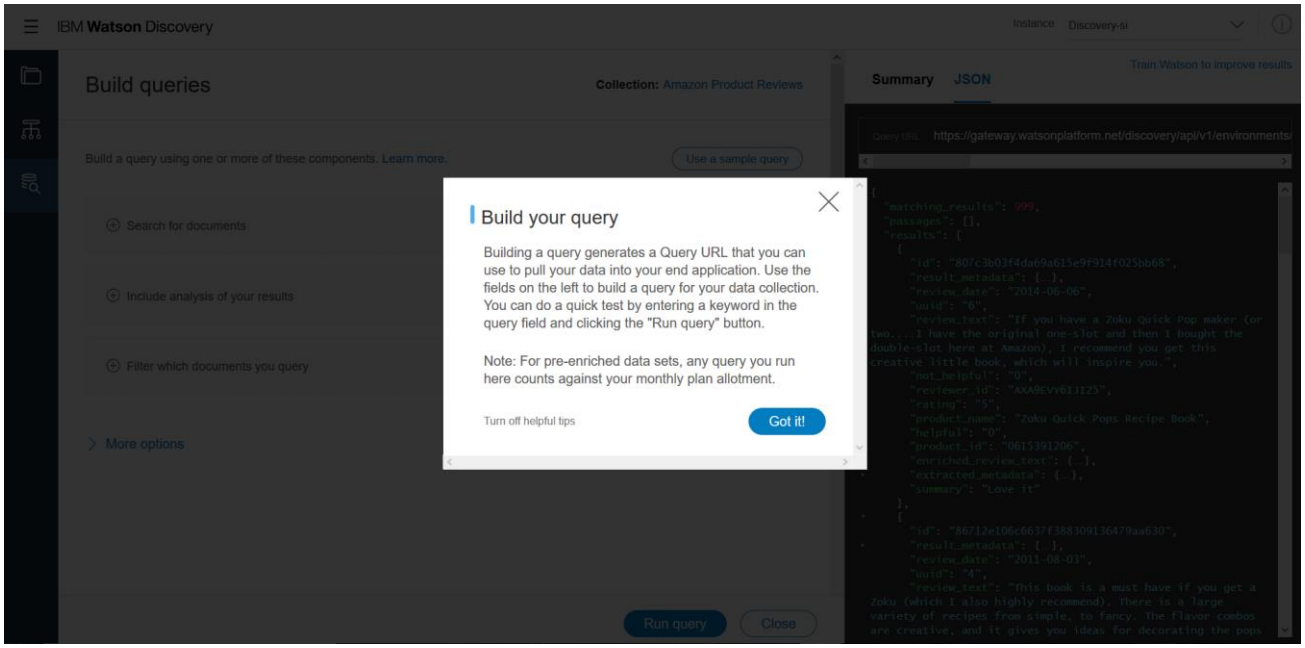
3. You can view the data schema in the "Collection view" or the "Document view." "Collection view" will display all fields in your collection and "Document view" will display a sample of documents from your data set with all fields included in the left side. Both representations are included below.



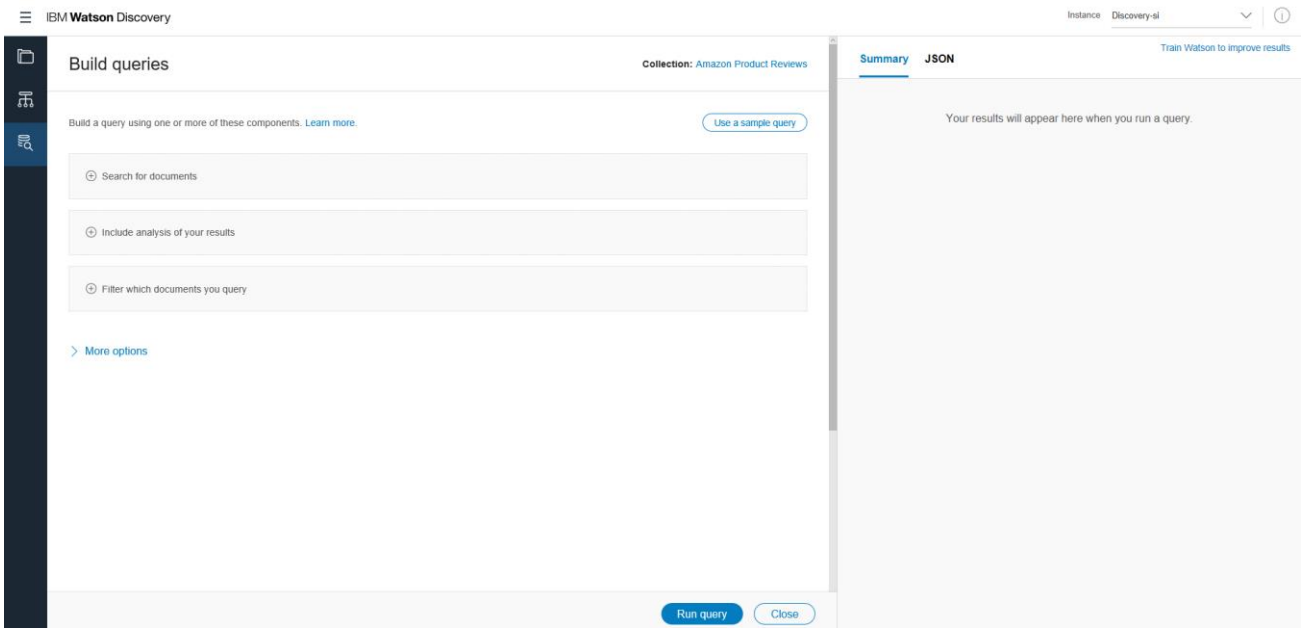


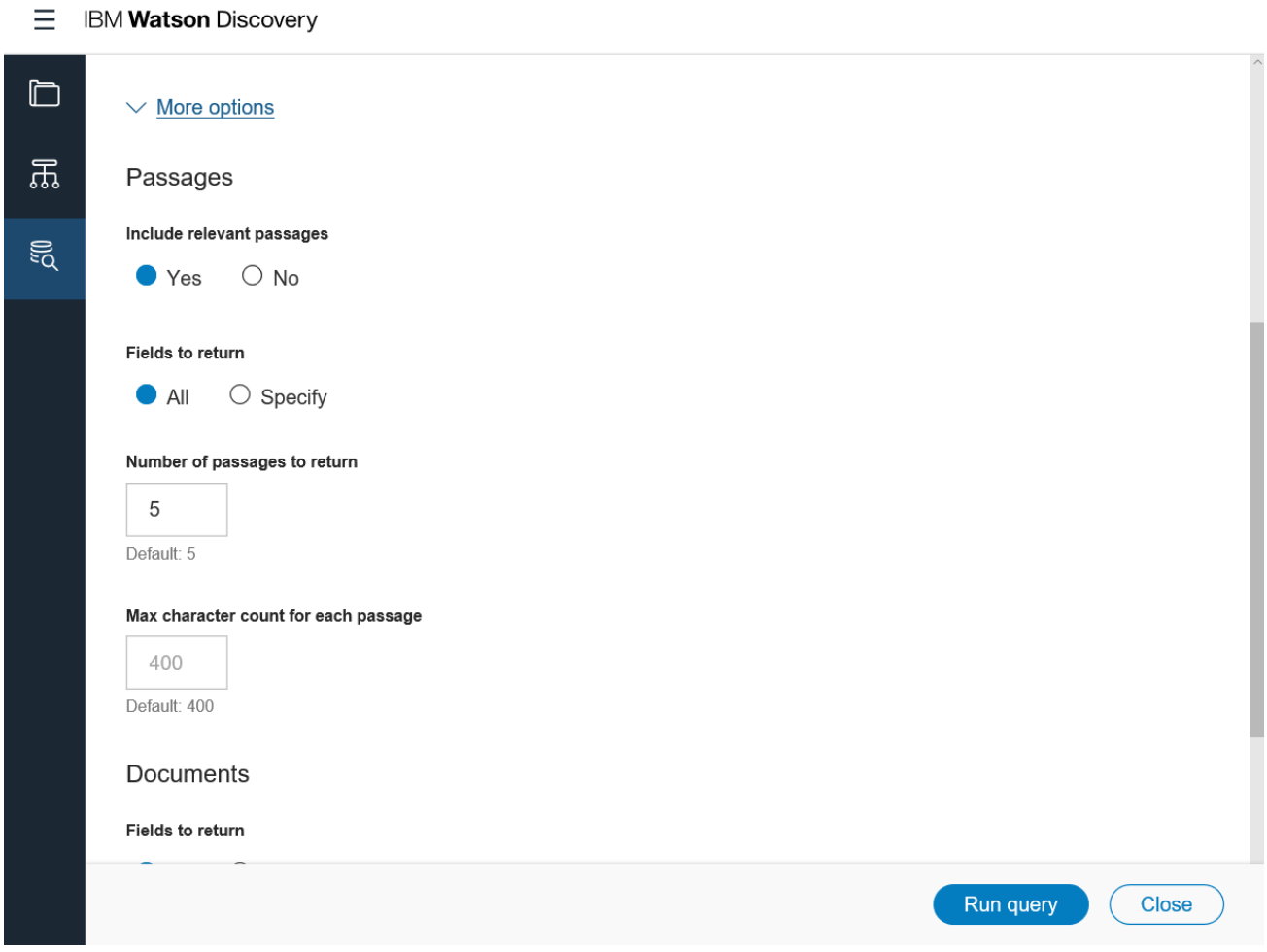
4. To build a query, you can either click "Build queries →" in the top right corner or click on the build queries icon on the lefthand side bar. Both options are circled in red below. If you click the icon, you will have to choose the collection you want to query which is "Amazon Product Reviews."





5. In the "Build queries" page, you can choose to "Search for documents," "Include analysis of your results," or "Filter which documents you query." Click "More options" to turn on passage retrieval.





You can click "Use a sample query" to check out a few pre-built queries.

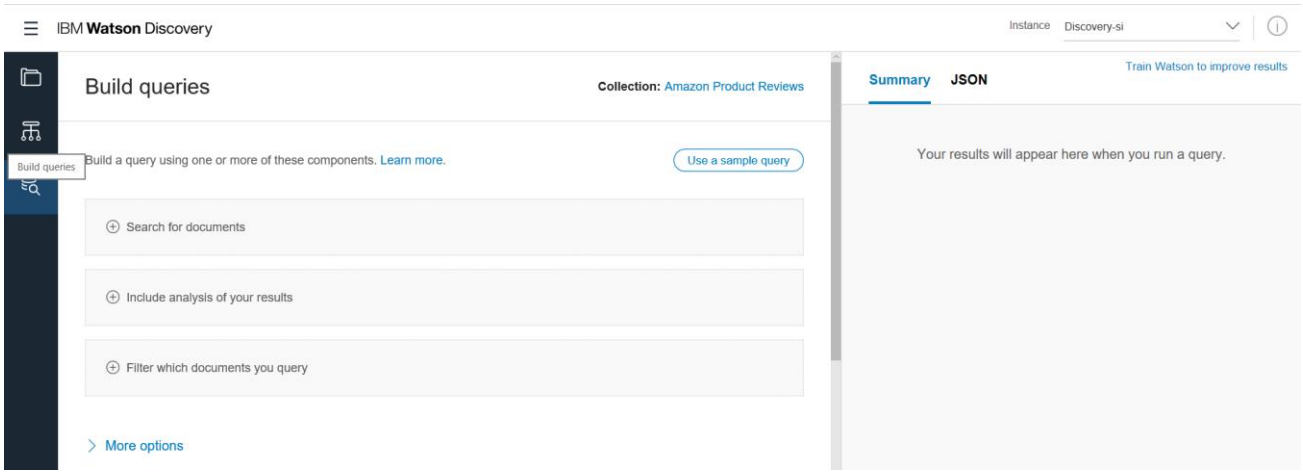
Below sections helps in understanding specific analysis performed on the dataset

5.1 Search for Documents

In this section, we will walk through custom queries that a developer would configure within their application in order to gain insight from the data. The application end user is Henry, a Business analyst hired by Amazon. Henry is responsible for consolidating product feedback and presenting it back to the product development team for continuous enhancements. Going through all the reviews by himself would be a tedious task. He therefore takes advantage of the WDS based application which can sort through huge number of reviews and provide valuable insight. As a developer, we need to help Henry generate actionable feedback by surfacing insight from the customer data. Below are some examples of the custom queries that a developer would create to help Henry achieve his goal.

The Discovery application provides users an option to build their own customer query and execute them. The query results will be shown in the right window of "Build your own query" page. In this example, passage retrieval is turned on.

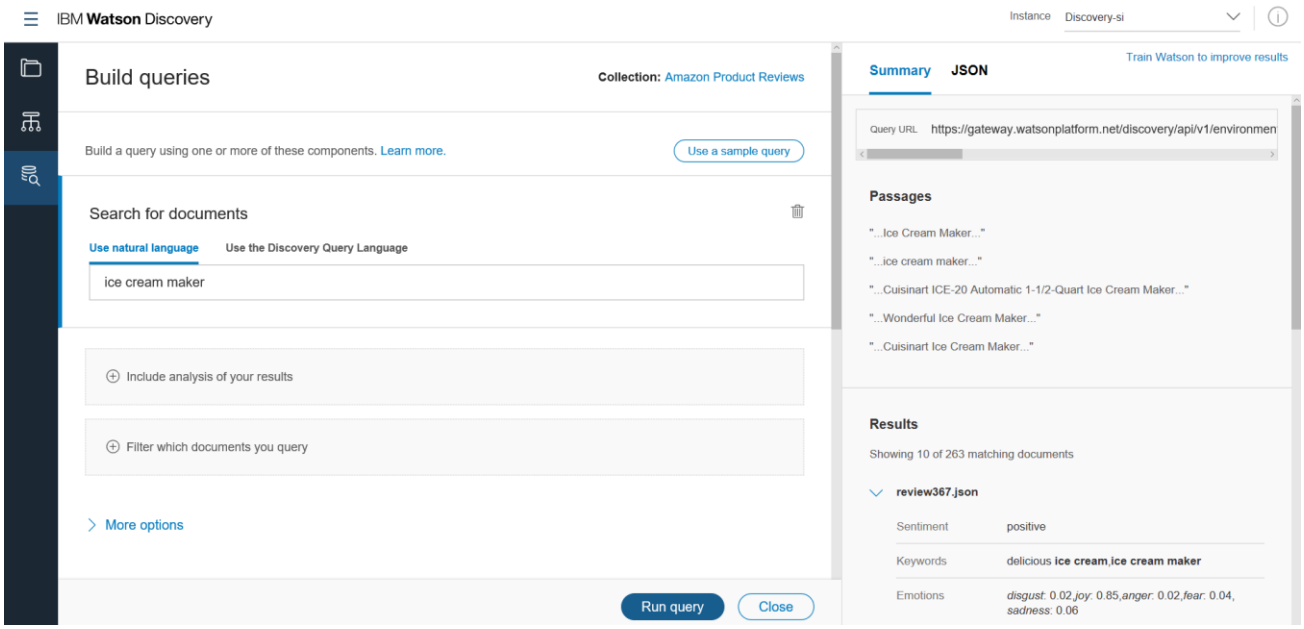
1. Click on "Build queries" button on the lefthand bar and select the "Amazon Product Reviews" collection. In "Build queries" page, text fields are provided for filling in the keyword, count, sentiment types and other parameters against which the queries will be executed



2. Click "More options" to make sure passage retrieval is included.

3. Click "Search for documents" to build queries.

4. The below figure shows an example of a custom query, "ice cream maker" executed against the Amazon review data



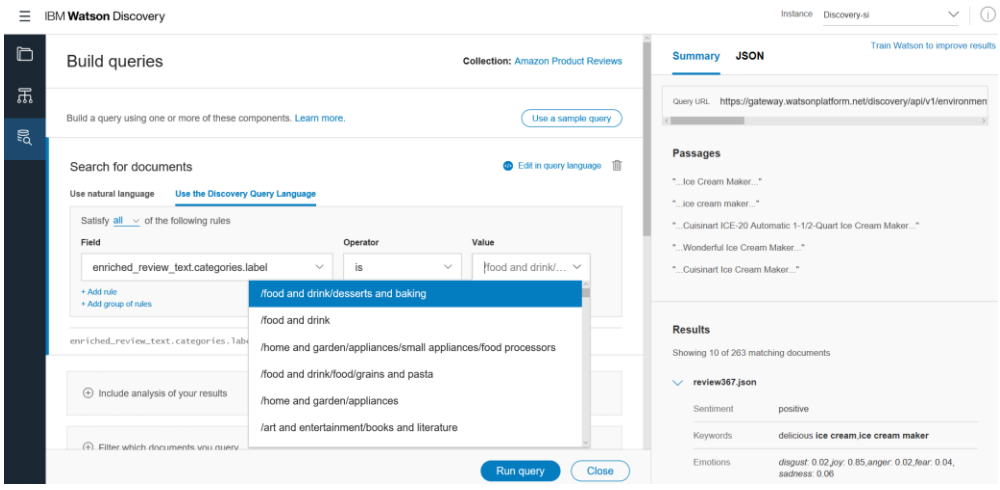
You can view the Summary or the corresponding JSON of the results on the righthand side.

In this query, the enriched reviews are first filtered against "ice cream maker." The query searches for "ice cream maker" in all the fields of the review unless you specify a particular field to look for. By default, the query returns 10 matches, but as a user you can change that number by editing the field "Number of results to return".

The custom queries are powerful and can get deeper insights which can be exploited by users to understand where does their business stand and what could be done to reach a better ground.

5. Click the trash icon to clear the query.

6. On the Build your own query page "Search for documents" section, click the "Use the Discovery Query Language" button and add to the field, operator, and value.



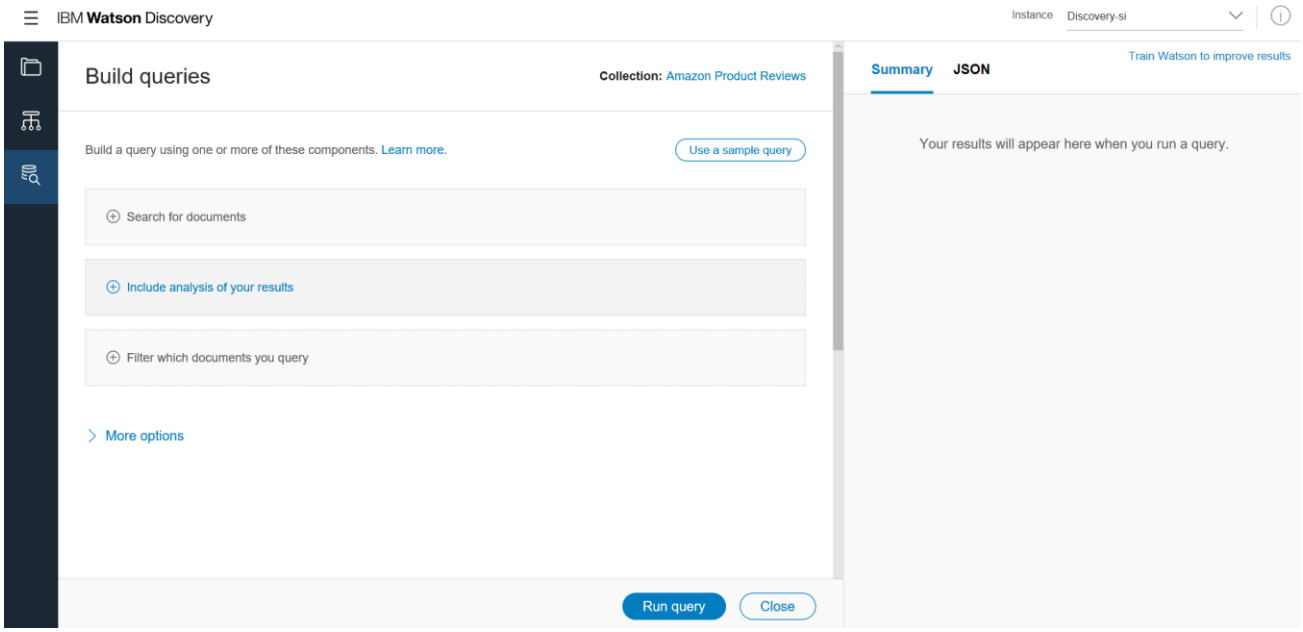
This shows another way you can build querys.

7. To clear some fields from dropdown, click on the cross mark next to each field if you have multiple selected. To clear the query click on the trash icon in the upper righthand side of the white space.

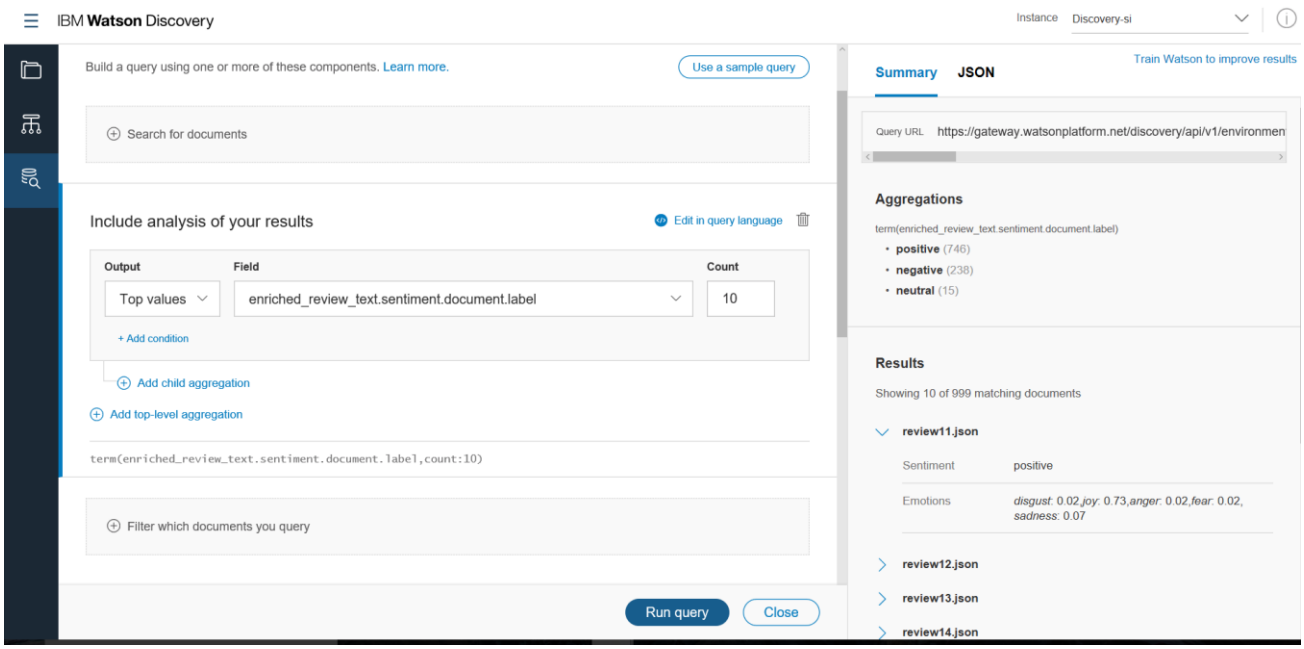
5.2 Include Analysis of Your Results

Henry may want to see a further analysis of the distribution of sentiment in reviews. He also may want to dive deeper into the keywords to see if there are negative trends with certain products. Section 5.4 covers keyword analysis using queries. To do this, he can use aggregrations in the build queries page. Below is a quick briefing on how to use the analysis component.

1. Click "Include analysis of your results"



2. Type or select "enriched_review_text.sentiment.document.label" in the Field section and keep the count at 10.



Henry sees the distribution of positive, negative, and neutral reviews and can view the sentiment and emotions of 10 results.

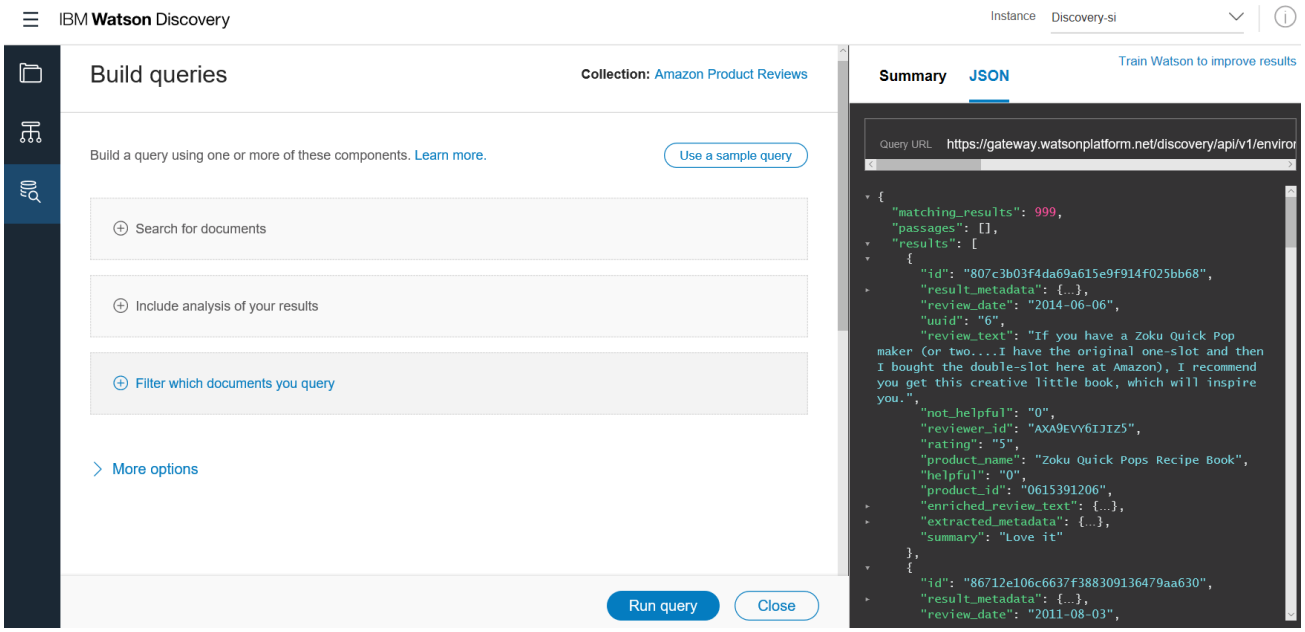
3. The analysis can be removed by clicking the trash icon.

5.3 Filtering Documents You Query

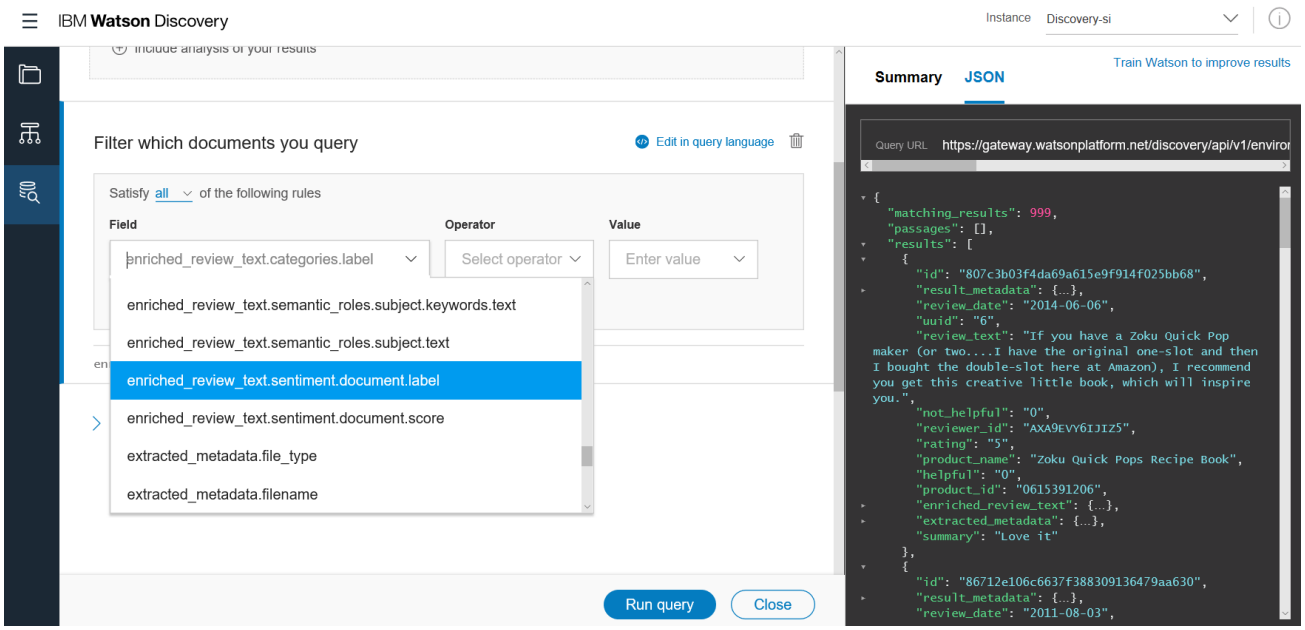
Henry might want to start his analysis by trying to understand how much positive and negative feedback is provided for a specific product. For example, he would want to know if the Cuisinart products are not doing so well so that he can take action and mitigate any risks. As a developer, we need to start by understanding what part of the data provides this insight. The enriched review JSON returned by Watson Discovery Service has lot of fields and we are only interested in querying the particular field that provides an answer to Henry's question.

The query results usually contain all the fields present in the original JSON files along with the fields generated under the enriched attribute. We can specify in the query to return just the fields we want in the result output.

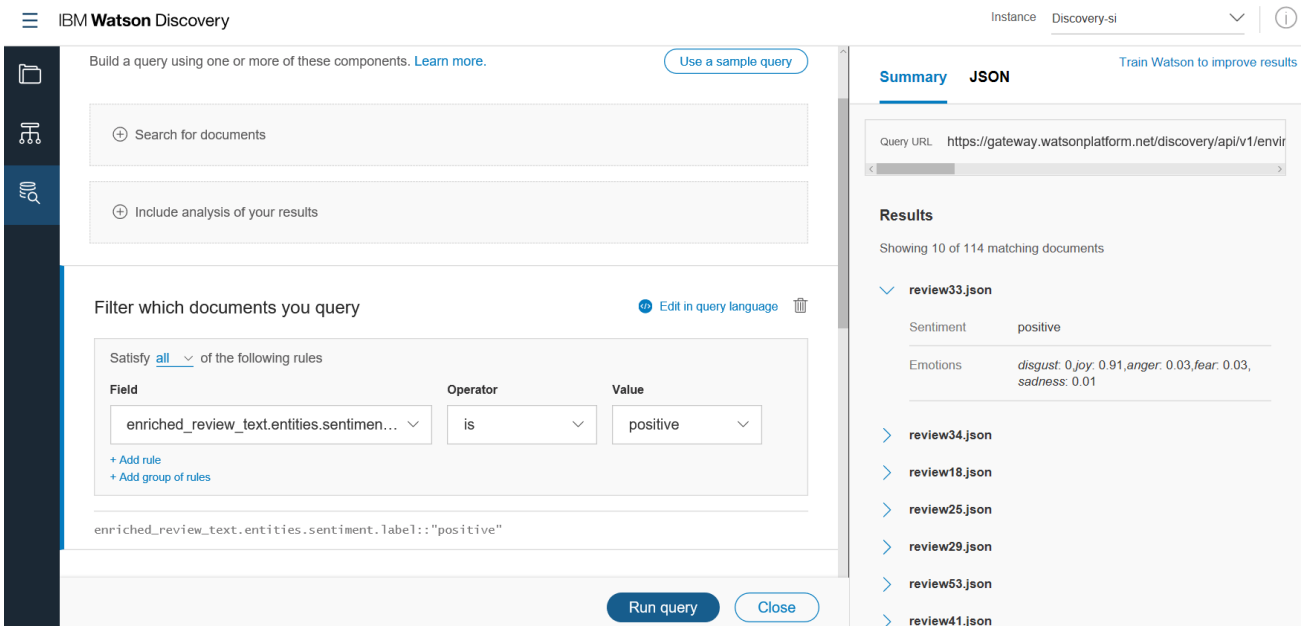
1. The filtering can be applied using the "Filter my data insights" section on the left side of the My data insights page.



2. Select the type of enrichment you want to filter on the dropdown menu: keywords, concepts, sentiment, entity or taxonomy. Select enriched_review_text.sentiment.document.label.

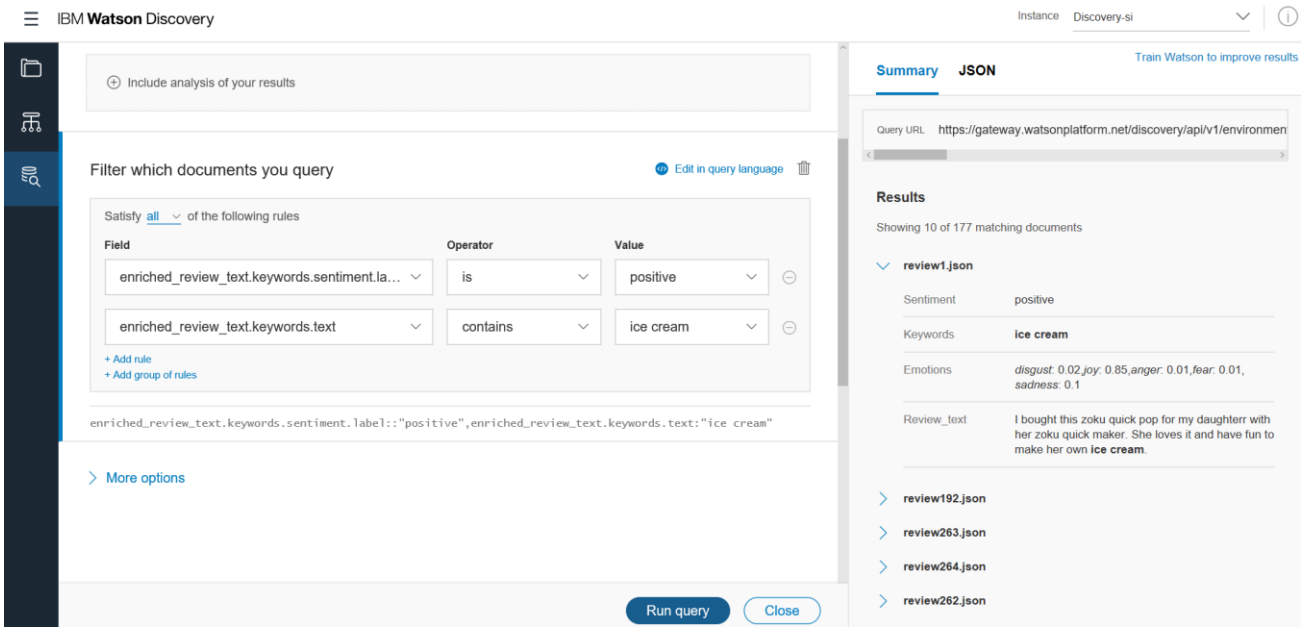


3. From the "operator" dropdown select "is" and from the "value" dropdown select "positive." Click apply filter to see the filtered insights.



4. The filtered insights show the number of matching documents along the top 10 documents emotions.

5. To filter against multiple types, click "add rule" on top of already filtered insights. In the below example, "enriched_review_text.keywords.text" "contains" "ice cream" is added so the results filter positive reviews with the mention of ice cream.



6. The insights now are filtered against all the reviews containing ice cream keyword and are classified under positive sentiment

7. The filtering can be removed by clicking the cross button next to the applied filter displayed on top of data insights.

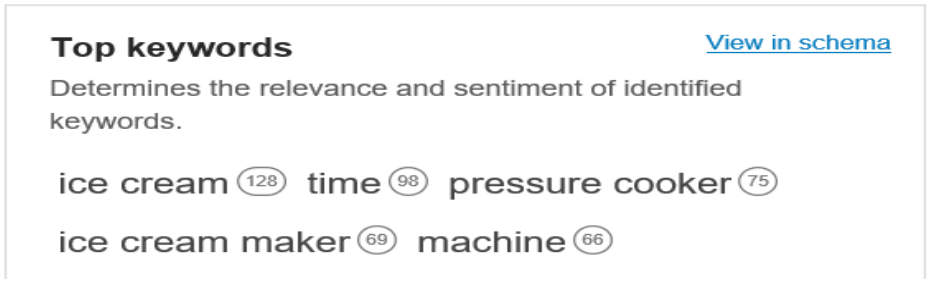
   These first steps allow developers to sift through the enriched JSON output from Watson Discovery Service. We are now ready to query the JSON fields to extract meaningful insight.

5.4 Top keywords

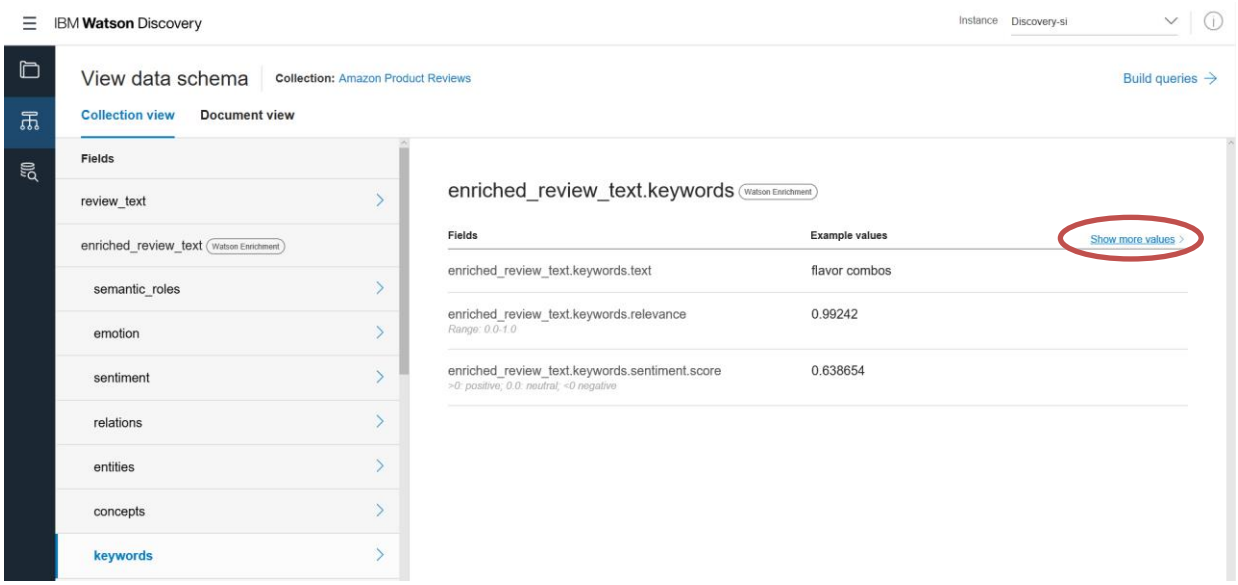
This segment displays the top keywords present in the reviews along with the count depicting the number of times they appear in all of the reviews. The Discovery service automatically identifies supported languages in your input content, and then identifies and ranks keywords in that content

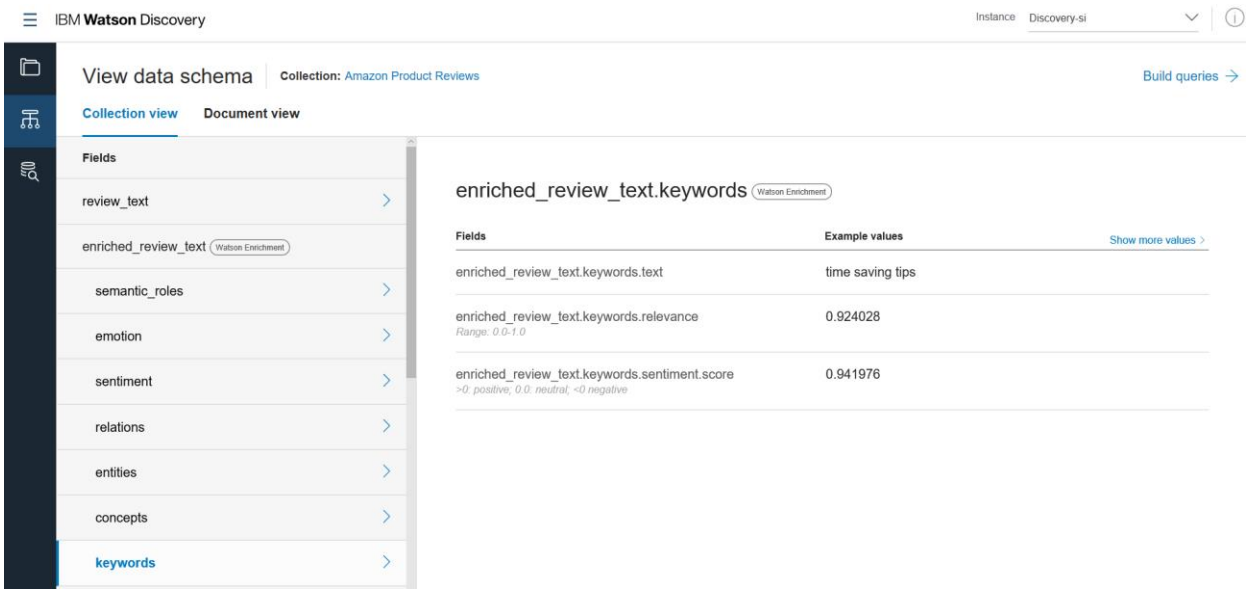
For the Amazon review data, the keywords include ice cream, time, pressure cooker, ice cream maker, and machine. This makes sense since the reviews are about kitchen appliciances and ice cream machines.

1. When looking at the "Insights from your enriched data" on the Manage data page, click the "View in schema" button on the right top of the Top keywords box. Upon clicking the button, the user will be sent to the data schema collection view of the top keywords as discussed in the overview.

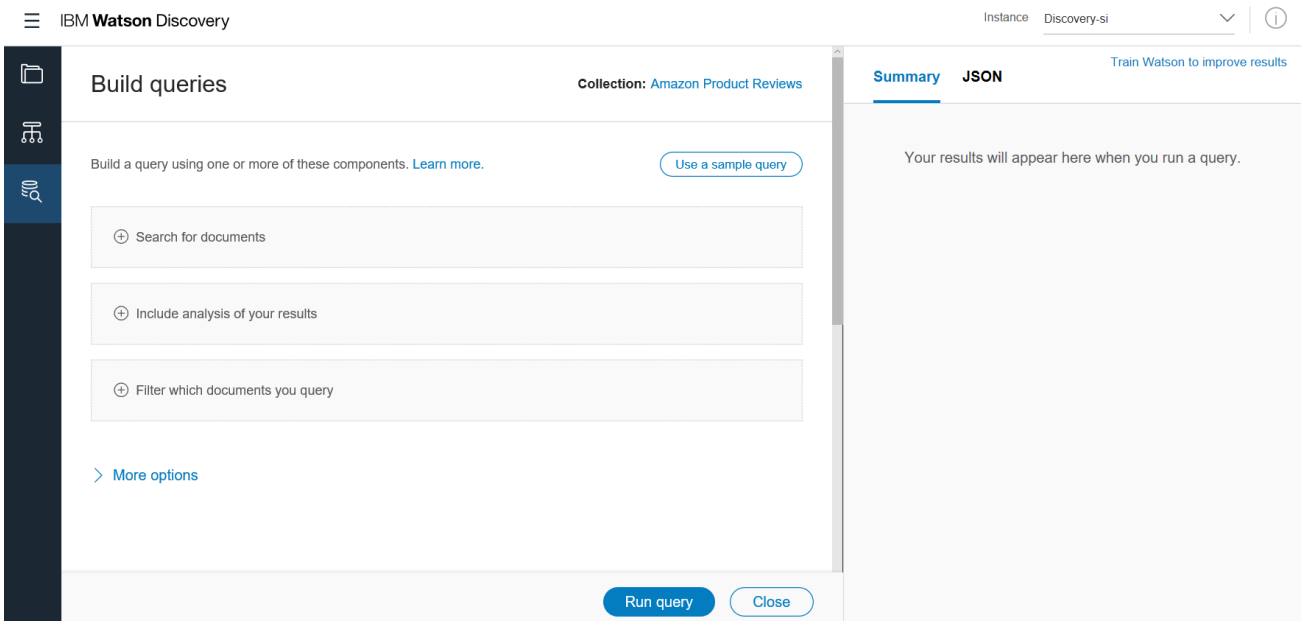


2. The top keywords query is prepopulated with a few example values and the corresponding relevance and sentiment. Click "show more values" to view different relevant examples.

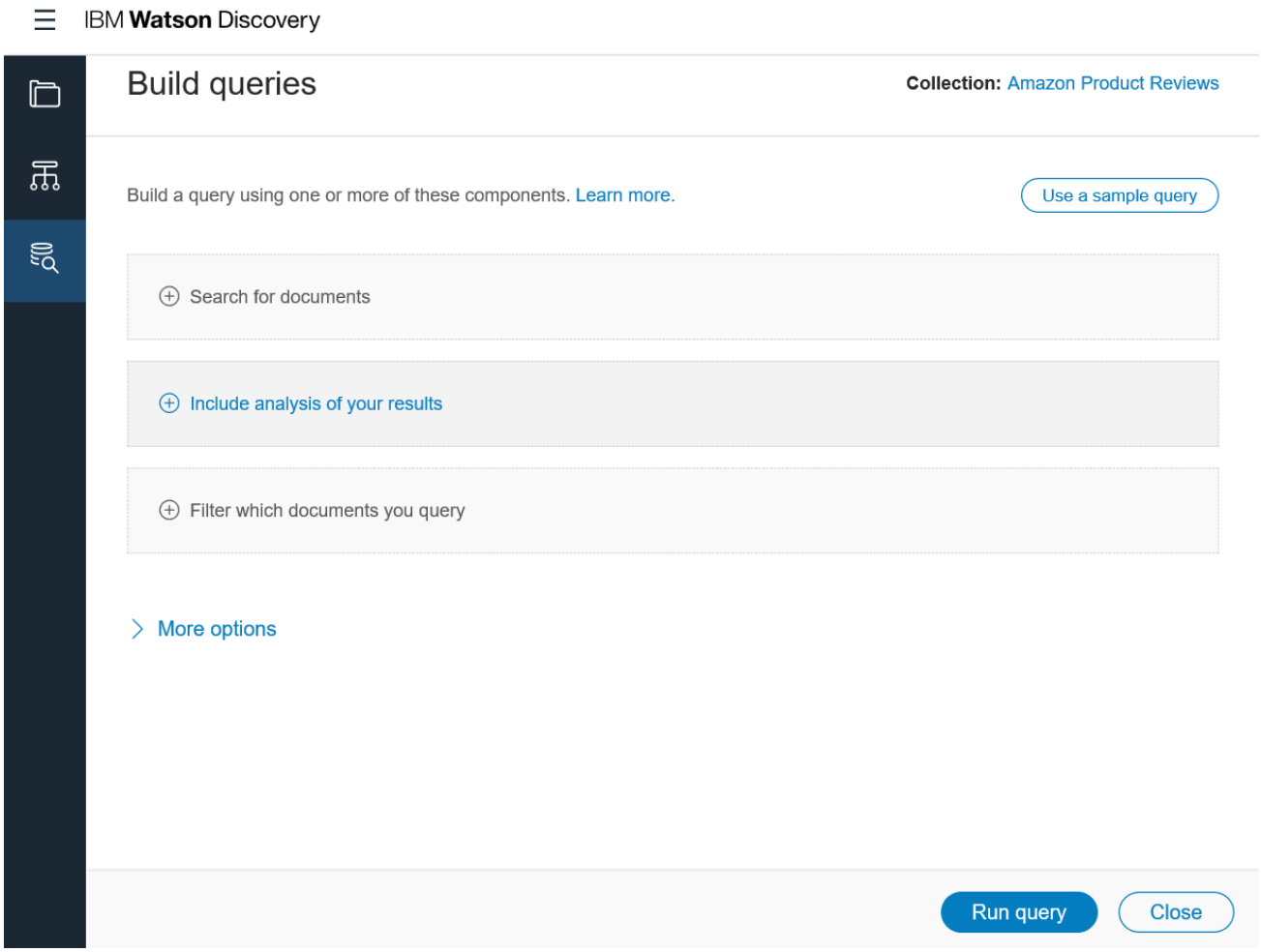




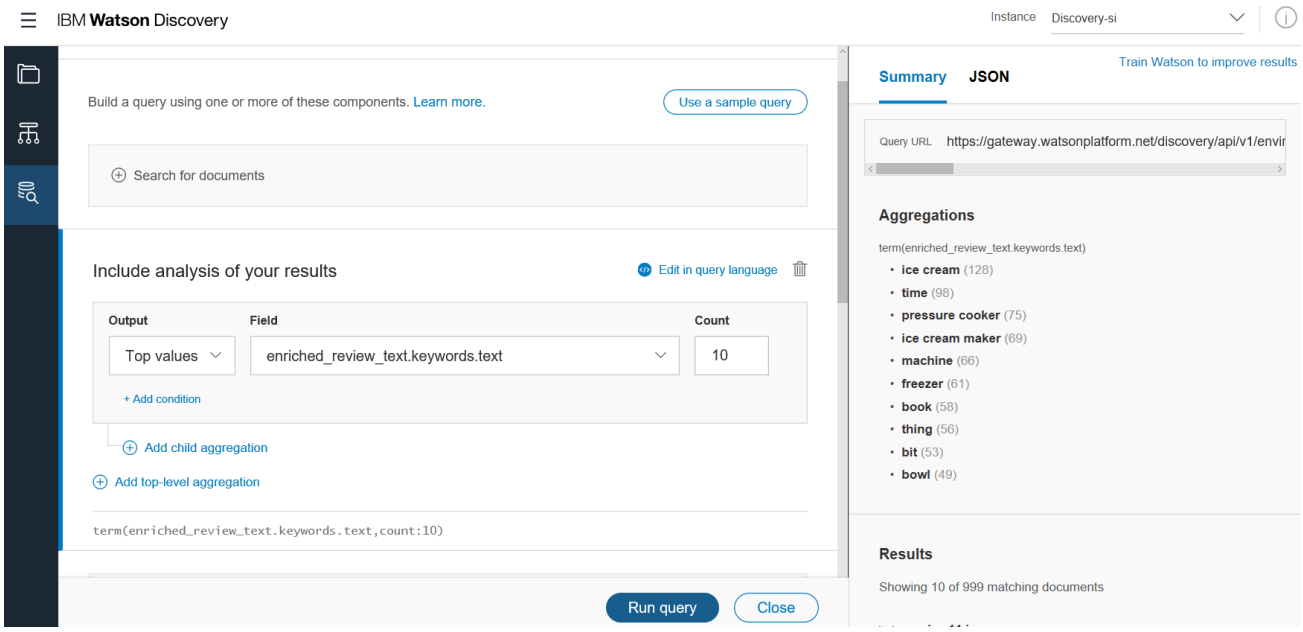
3. After viewing the different values, click "Build queries" in the top right corner. The "Build queries" page is shown below.



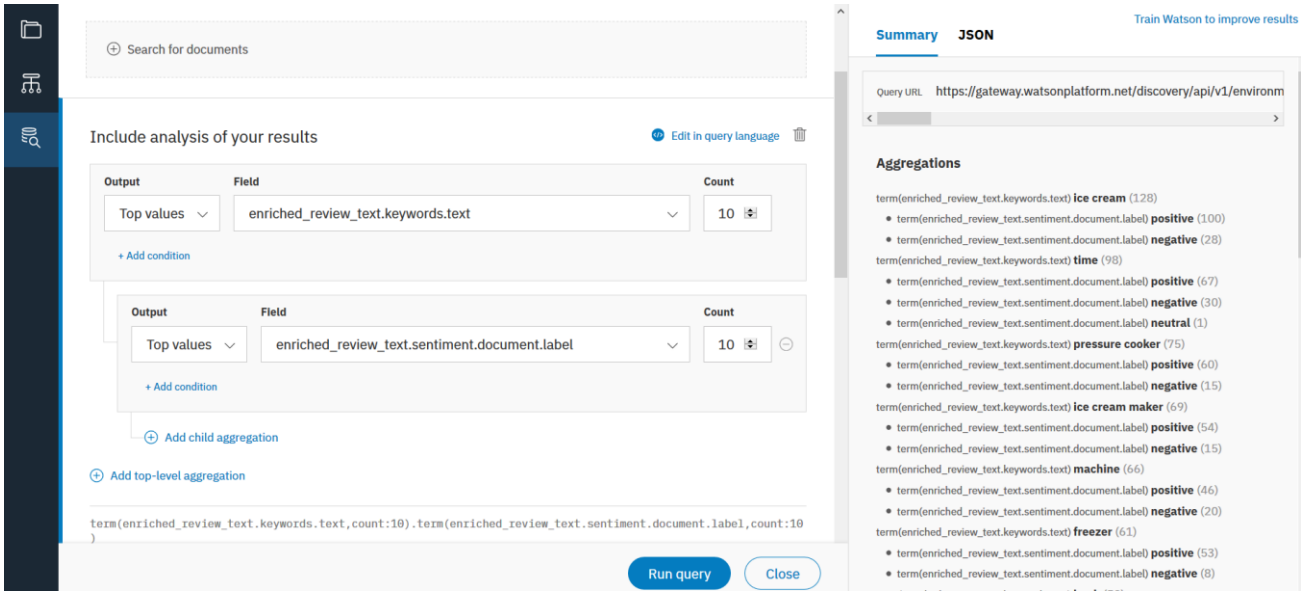
4. Click "Include analysis of your results" to build a query associated with the top keywords.



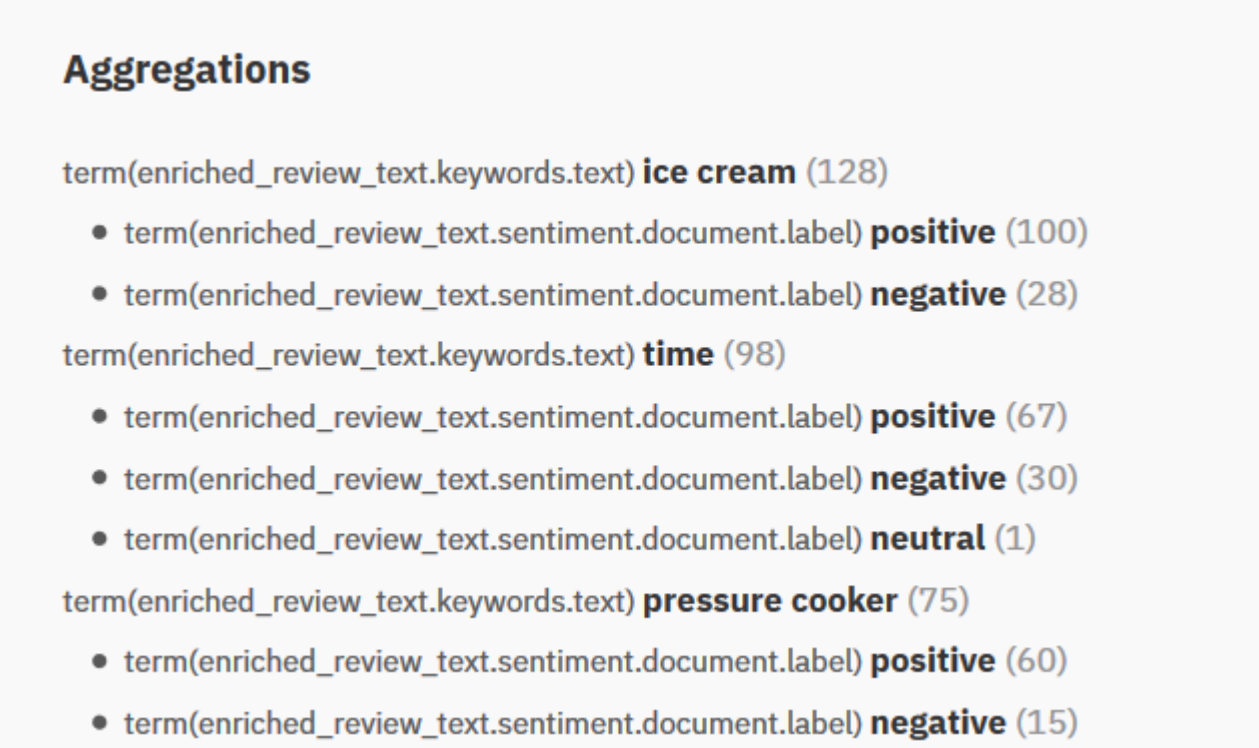
5. Type in "enriched_review_text.keywords.text" in the Field section and keep the count at 10 then click "Run query" at the bottom of the page. Below the top 10 keywords are aggregated.



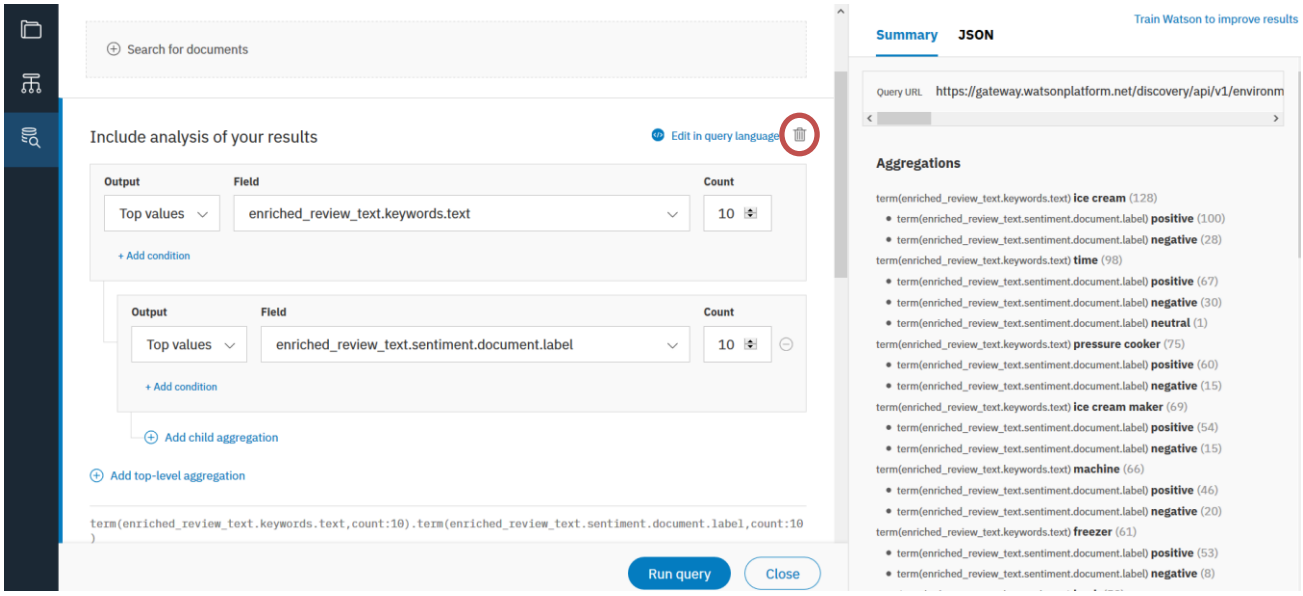
6. You can change the query by increasing the count, adding a condition, adding child aggregation, or adding top-level aggregation. Click "+Add child agreggration" and type or select "enriched_review_text.sentiment.document.label" for the field. This way you can view the distribution of positive and negative reviews relating to the specific keyword. Click "Run Query



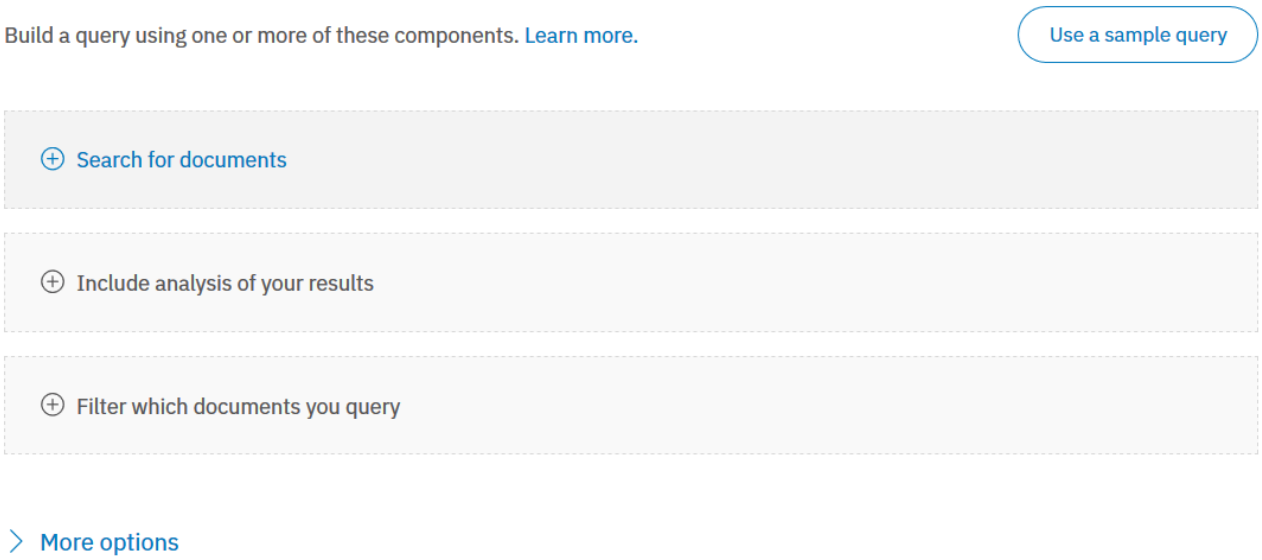
The results show how there are 100 positive reviews about "ice cream" and 28 negative reviews. For "time," there are 67 positive reviews and 30 negative reviews. It is worth drilling further into the reviews with "time" in them.



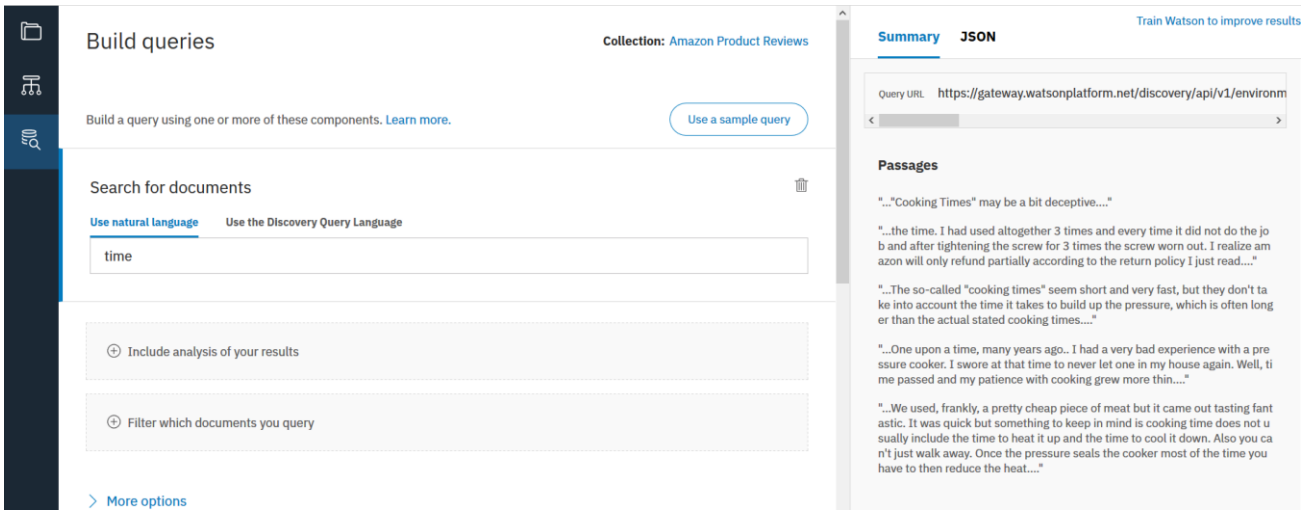
7. After reviewing the Aggregrations results, click the trash can icon to move on to the "Search for documents" section.



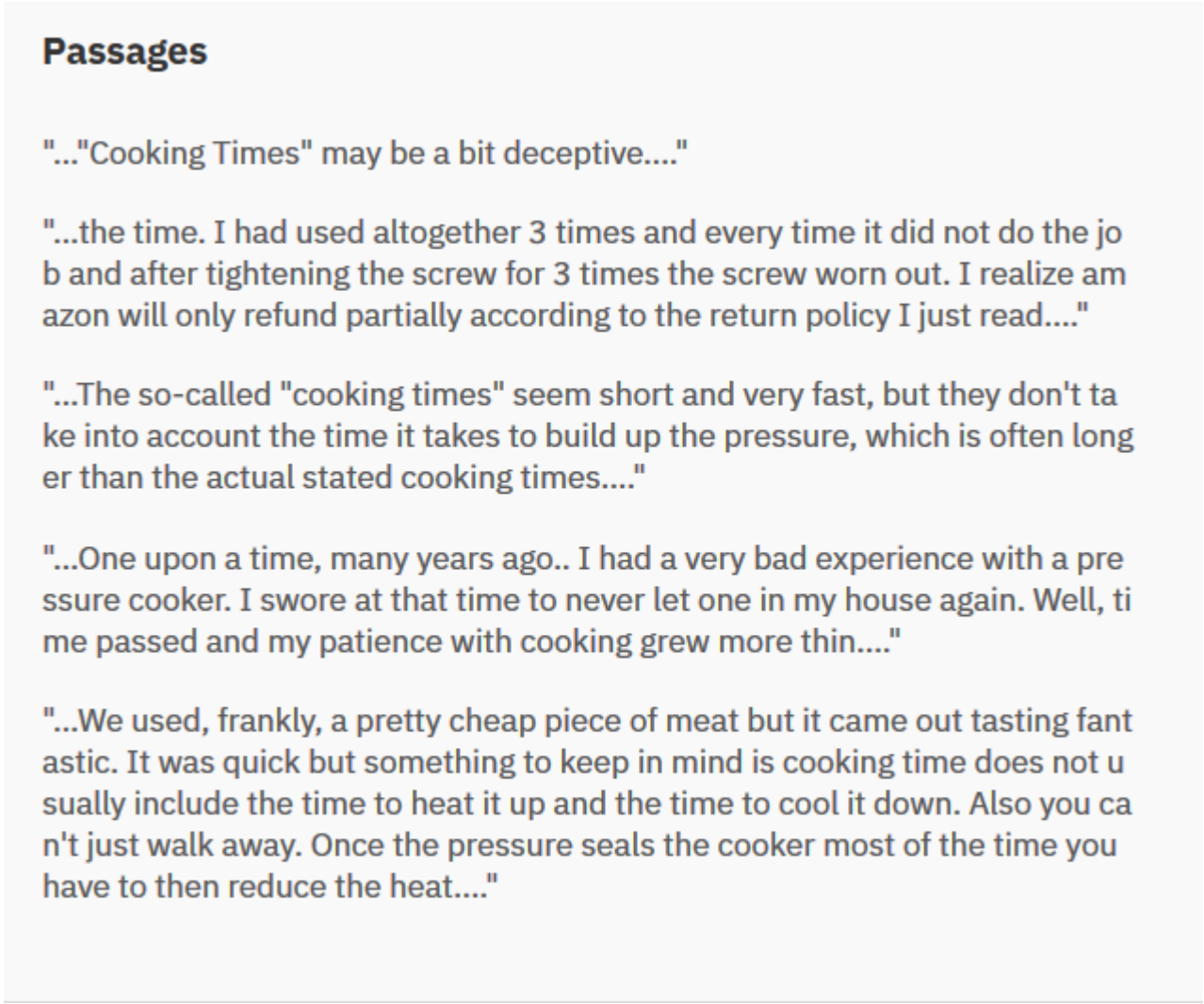
8. Make sure "Passage Retrievals" is set to on by clicking "more options at the bottom of the page. After you confirm it is on, click "Search for documents."



9. Under the "Use natural language" tab, type in "time" and click "Run Query at the bottom of the page or press enter on your keyboard.



When looking into the passages taken out from reviews, you can see th from the passages that "Cooking Times may be a bit deceptive," "seem short and very fast, but they don't take into account the time it takes to build up the pressure, which is often longer than the actual stated cooking time…" and lastly, "cooking time does not usually include the time to heat it up and the time to cool it down." These passages taken from reviews on the pressure cooker show that the pressure cooker does not do a great job offering an accurate estimate of the cooking time. This offers insights for Henry to consider adding a more accurate timer feature to the pressure cooker.



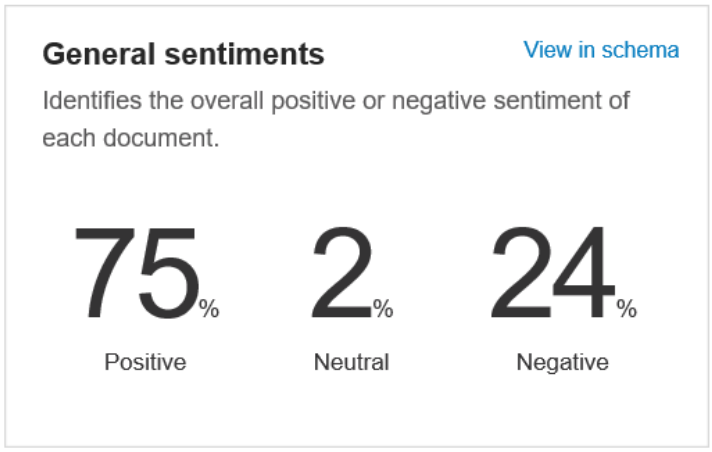
10. To return to the Query dashboard, click on the Collection name (Amazon Product Reviews) in the top right hand side of the white section.

5.5 General Sentiments Analysis

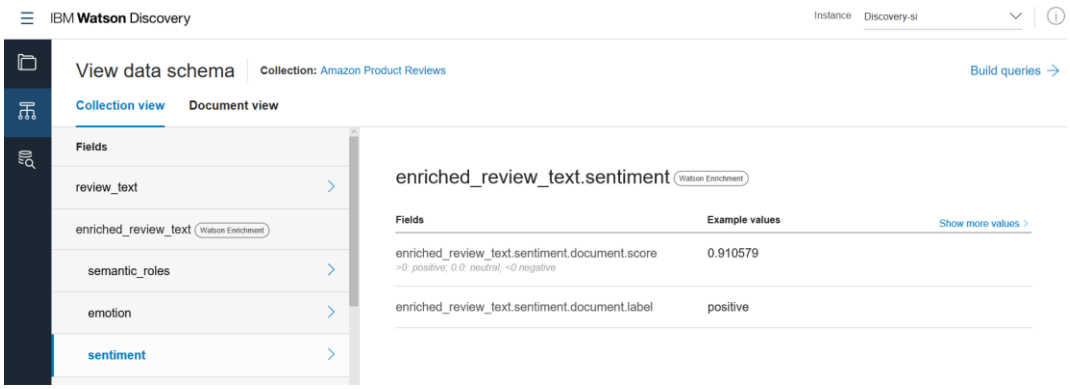
This segment displays the distribution of sentiment associated with the reviews broadly classifying them into positive, neutral and negative sentiment. Discovery service Identifies attitude, opinions, or feelings in the content that is being analyzed. A good rating score does not mean that the user is actually satisfied, since there is a possibility that the review carries negative sentiment with it. Sentiment analysis helps in identifying trends in the reviews and eliminating fake reviews

1. For the Amazon use case the classification says that 75% of the reviews fall in positive sentiment category, 2% and 24% fall in neutral and negative sentiment category respectively.

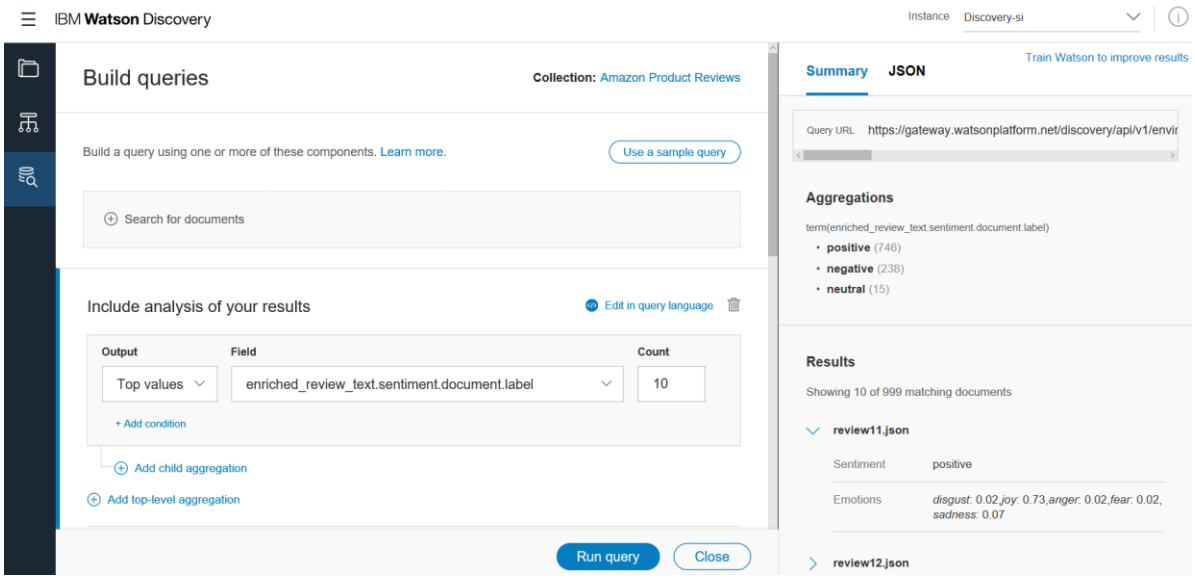
2. Click on the “View in schema” next to General Sentiments tile to navigate to view your data schema and then build your own query page.



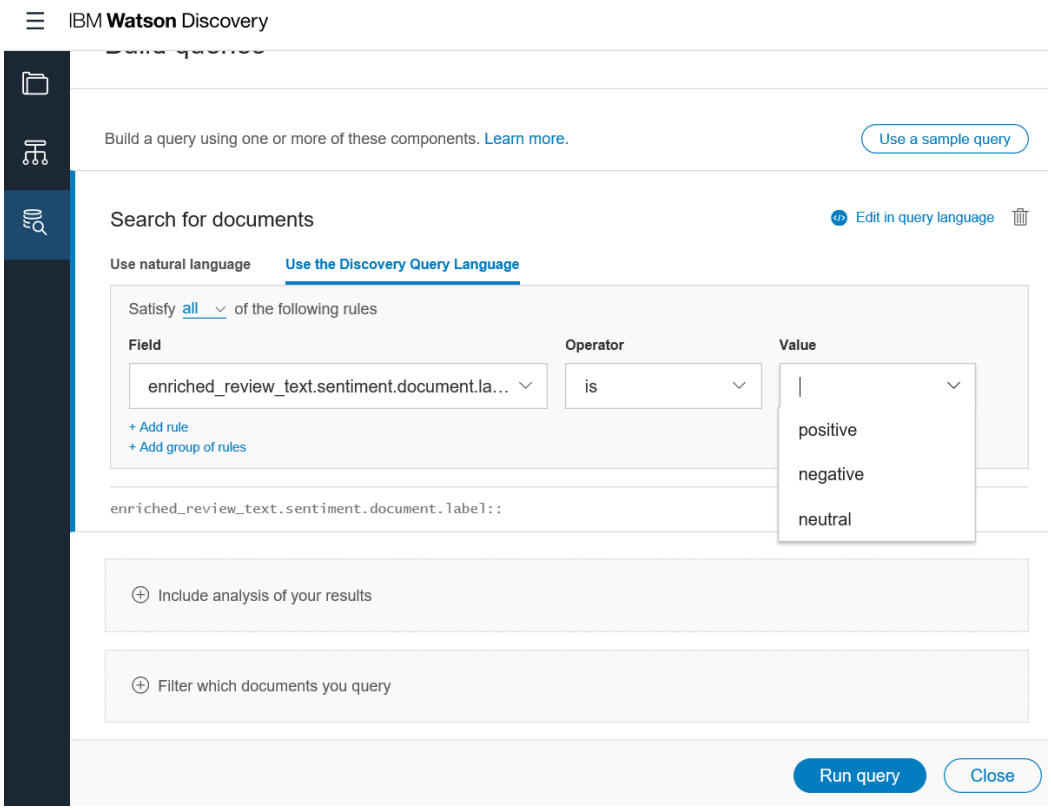
3. Example values are prepopulated. Click "show more values" to view different relevant examples. After viewing the different values, you can click "Build queries" in the top right corner to build out queries and gain further data analysis.



4. Follow the steps you did in the last section to include an analysis of your results, but this time type in "enriched_review_text.sentiment.document.label." The query results show the number of reviews with positive, negative and neutral sentiments.



5. You can change the query by increasing the count or you can filter against categories, entities, concepts, emotion, keywords and more. You also can search for documents using the Discovery Query Language and selecting a specific sentiment value.



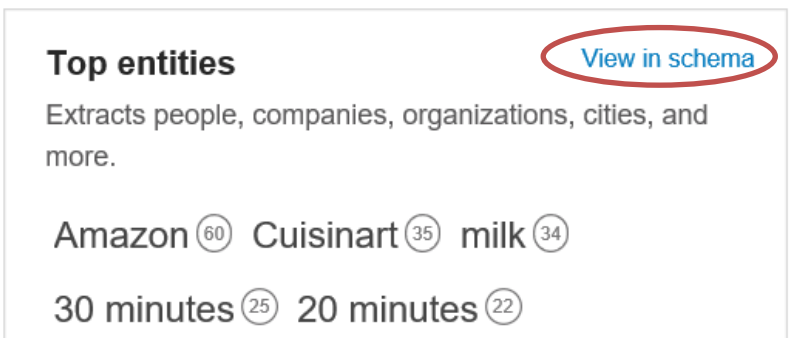
6. To return to the Query dashboard, click on the Collection name (Amazon Product Reviews) in the top righthand side of the white section.

5.6 Top Entities

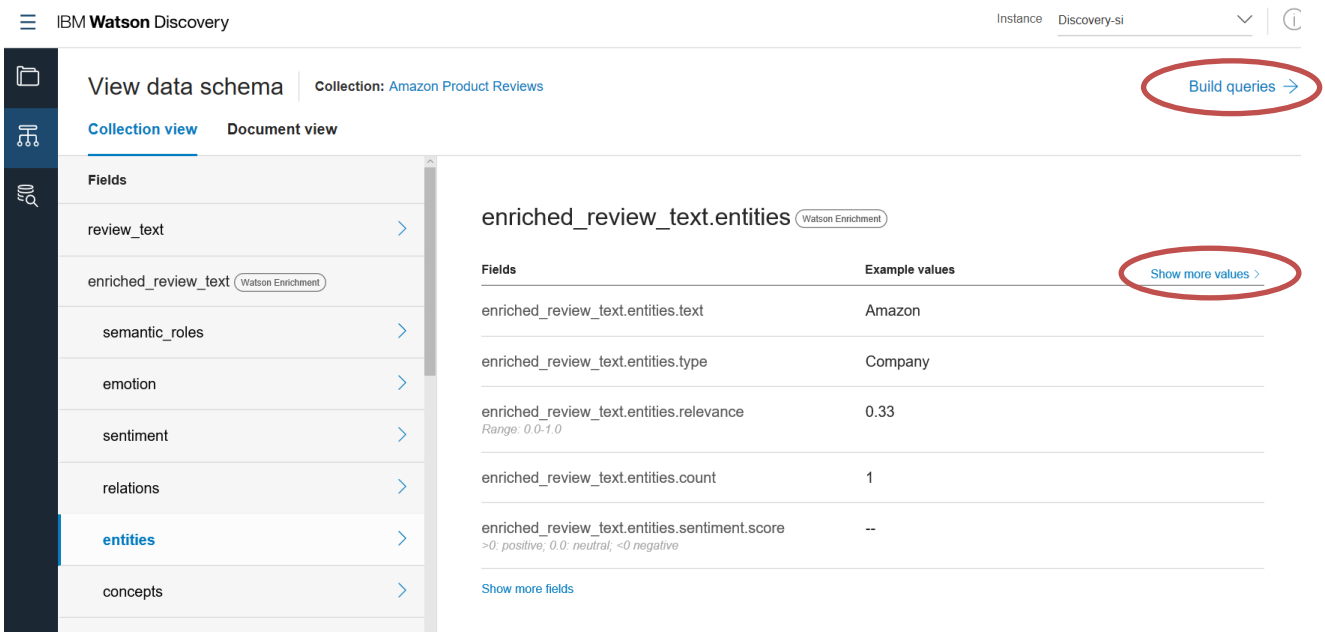
This segment displays the top entities present in the review documents: Entities like person, place or organization. Entity extraction adds semantic knowledge to content to help understand the subject and context of the text that is being analyzed

1. For Amazon review data, the top entities identified in the Top entities tile were pressure cooker, Cuisinart, Amazon, Milk etc.,

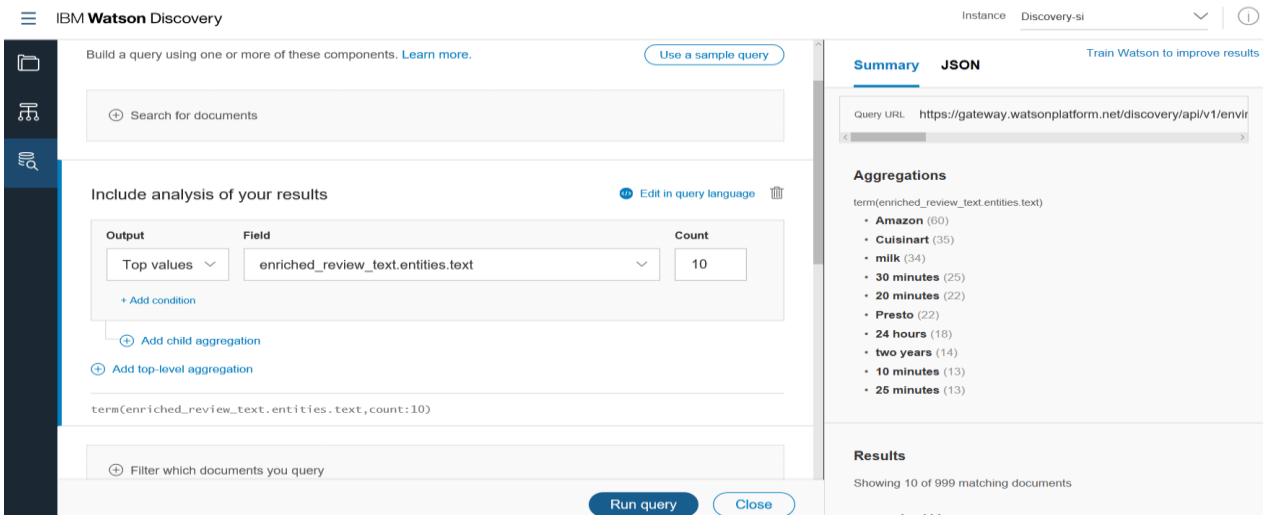
2. Click on the "View in schema" link to navigate to view the data schema and then build your own query.



3. The data schema sections show the example entities from the review documents and shows the relevance, count, and score. Click "Show more values" to view the other top entities and click "Build queries" in the top right corner after to build out your own queries.

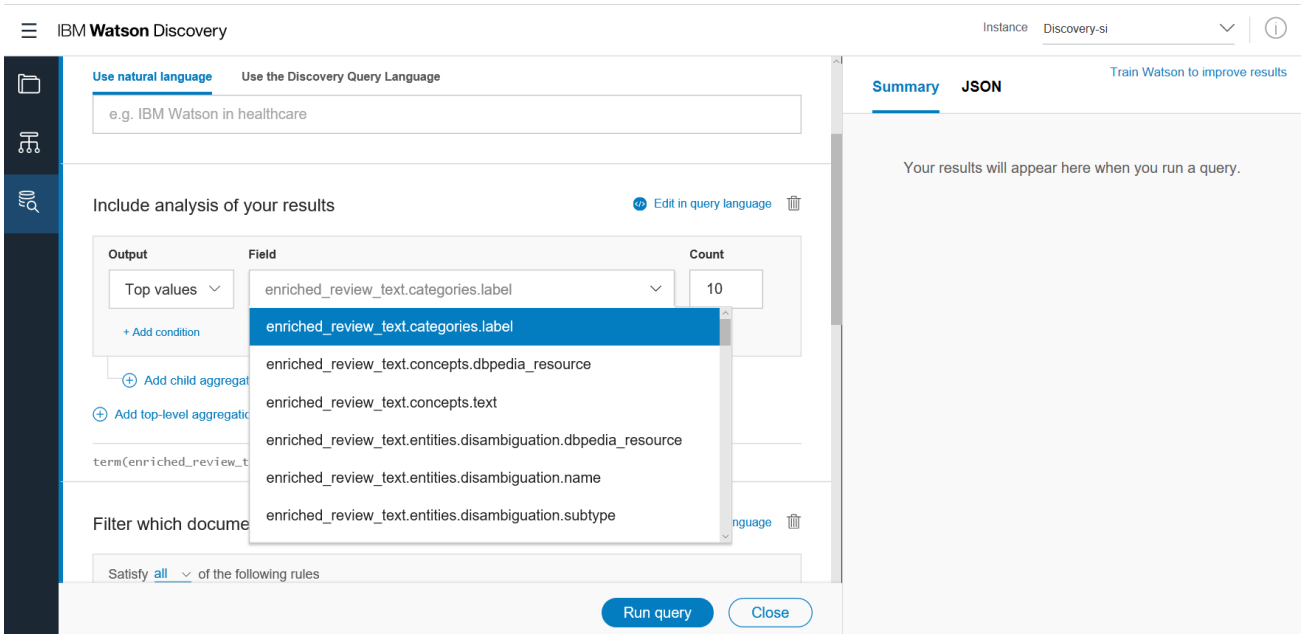


4. You can run through steps similar to the previous lab sections for the entities enrichments.



5. To return to the Query dashboard, click on the Collection name (Amazon Product Reviews) in the top right hand side of the white section.

6. You can view the other entity sub fields by searching the various options under the field section. It worth trying the different fields to gain a more in-depth analysis of your data.



More information to help with building your own query can be found at https://www.ibm.com/watson/developercloud/doc/discovery/using.html and https://www.ibm.com/watson/developercloud/doc/discovery/query-reference.html

6 Useful Considerations: Data Collection

6.1 Collecting Content for Watson Discovery Service

During this lab exercise, we have facilitated the configuration steps by providing you with sample content data. However, when working with your data, you will have to perform two additional activities: data collection and data preparation. Data collection activities are required to consolidate the data. Data preparation refers to the activities required to take the source data and prepare it in order to be supported for ingestion in Watson Discovery Service.

6.1.1 Data formats supported

In this Lab, we provide you directly with a folder containing a set of clean Amazon Product Reviews. However, in real engagements, data will mostly likely be crawled from other sources such as internal databases. Imagine for that scenario that we were storing and maintaining Amazon Product Reviews within a database. Here are some considerations we would want to keep in mind:

1. Watson Discovery Service supports three ways of uploading data into your instance: through the tooling web interface, via API call, and by using the Data Crawler tool.

2. All ingestion methods support uploading Word, JSON, Pdf, or Html files;

3. Each Amazon Product Review should be converted into a single JSON document with the document structure clearly headlining the property name and its values. This step will allow us to enrich our data effectively and be able to extract the insight we need to meet the business scenario.

4. In this scenario, you might want to store the Amazon Product Reviews within a database. The database should contain the properties as columns and each review should be a row entry in the table.

5. An example database for the Amazon Product Review use case is shown below. The structure was created to be supported by the Data Crawler.

uuid	reviewer_id	product_id	helpful	not_helpful	review_text	rating	summary	product_name	review_date	time_inserted
2	A1JVQTAGHYOL7F	615391206	0	0	I bought this zoku quick pop for my daughterr with her zoku	5	zoku	Zoku Quick Pops Recipe	6/17/14	12/13/16
4	A2MHCTX43MIMDZ	615391206	14	18	This book is a must have if you get a Zoku (which I also hig	5	Creative Combos	Zoku Quick Pops Recipe	8/3/11	12/13/16
5	AHAI85T5C2DH3	615391206	0	0	This cookbook is great. I have really enjoyed reviewing all o	4	A must own if you own the Zoku n	Zoku Quick Pops Recipe	6/6/14	12/13/16
6	AXA9EVY6IJIZ5	615391206	0	0	If you have a Zoku Quick Pop maker (or two....I have the or	5	Love it	Zoku Quick Pops Recipe	6/6/14	12/13/16
7	A1SW2D234X11MS	615391206	2	3	This book is so beautifully illustrated and easy to follow. It h	5	Beautifully illustrated recipe and ir	Zoku Quick Pops Recipe	7/1/12	12/13/16
9	A1V13XAWO98C1J	615391206	13	13	The Zoku accessories are pricey, but the recipes have the e	5	If you have the Zoku maker, buy t	Zoku Quick Pops Recipe	9/21/11	12/13/16
10	A3842PXNQ1QRM9	615391206	0	0	High quality book. Clear and helpful information, and then s	5	Satisfaction opinion	Zoku Quick Pops Recipe	5/31/14	12/13/16
11	A2BVT5QJSEEQI8	615391206	3	3	The recipes in this books are tedious to make. I found that e	4	I liked this book, but.....	Zoku Quick Pops Recipe	12/25/12	12/13/16
12	A3R9X003XW0LNR	689027818	0	0	This beautifully illustrated book featuring ten colorful,glittery	5	A sweet book for babies and todd	Good Night, Sweet Butter	7/26/09	12/13/16
14	A14BTJRH9VNLJJ	689027818	1	2	This is the story of a group of butterflies looking for a place	5	My little one absolutely loved this	Good Night, Sweet Butter	4/5/13	12/13/16
18	A3LHQ0LXICBLY2	912696591	1	1	I should not have bought it because it comes free from my c	5	I should not have bought it.	Wilton Decorating Cakes	6/6/13	12/13/16
19	A39RRTO760KGOY	912696591	0	0	had no idea about decorating cakes,by this booklet ,now i c	5	follow the instructions	Wilton Decorating Cakes	5/26/12	12/13/16
20	AZF13R8HNB59E	912696591	0	1	excellent book, shipped on time, packaged nicely, helpful tip	5	book	Wilton Decorating Cakes	6/11/12	12/13/16
23	A2S3ZE0RI9I60A	912696591	0	0	helpful information on decoraing basics. Good book for beg	4	This is a good book for anyone int	Wilton Decorating Cakes	12/12/13	12/13/16
24	A33T8O72UNBVR	912696591	1	1	Got this for my girlfriend who runs a baking business and sh	5	Handy to Have	Wilton Decorating Cakes	6/7/13	12/13/16
28	A2I5YZBMLU2R7O	912696591	0	1	I am very happy I purchased this book as it containes so mu	4	good tips	Wilton Decorating Cakes	5/6/13	12/13/16
30	A1LAAZUR8YFRW8	912696591	0	0	I have always wanted to learn how to decorate a beautiful c	4	Great For Beginners!	Wilton Decorating Cakes	8/15/13	12/13/16

6. The Data Crawler tool can be used to push large numbers of records into the Watson Discovery Service collection;

7. The Data Crawler uses data stored in a database or the data files stored in file system and creates JSON documents out of each data entry and pushes it into the Watson Discovery Service instance;

8. The JSON documents created using the crawler tool will have a slightly different format from that of the ones uploading using tooling or API calls. The attributes present in the Amazon Product Review will be placed under a parent node defined as “metadata”, in the JSON document

9. If you have crawled your data from a database, and you are trying to enrich your data fields, a prefix of ‘metadata” will have to be added. The following figure shows the sample JSON document created by the Data Crawler after crawling data from database


```
{
    "id": "f29187a2eb6a7dcb3ea074725cb1167b49eb9785b947abb69f7046cc58b9ef5d",
    "score": 1.2738007,
    "metadata": {
      "review_date": "2013-07-27",
      "uuid": "347",
      "time_inserted": "2016-12-13",
      "review_text": "Worked great for the first year we had it, when we used it maybe 3-4 times.But after that, it seemed like nothing would
      ever really freeze up enough. I looked into the problem and decided that I might have damaged the cooling capability of the bowl by
      cleaning it with hot water (which is a no-no).So, I bought a replacement bowl, and once again, for the first year with a few uses it
      seemed to be OK.But now once again, it's just not getting cold enough - and this time I *know* it's not my fault since I have been very
      careful to only wash the bowl in cold water.",
      "not_helpful": "0",
      "id": "f29187a2eb6a7dcb3ea074725cb1167b49eb9785b947abb69f7046cc58b9ef5d",
      "reviewer_id": "A2AXWQVA5V9GPJ",
      "rating": "2",
      "product_name": "Cuisinart ICE-20 Automatic 1-1/2-Quart Ice Cream Maker, White",
      "helpful": "0",
      "product_id": "B00000JGRT",
      "summary": "Cooling capability wears out fast"
    }
}
```


10. The addition of the metadata field will have implications on how you configure and query the Amazon Product Reviews. On the configuration side, you will have to create a configuration file that enriches the metadata.review_text field rather than just review_text field.

11. A similar approach will have to be used for querying your data. For example, you will have to change your query to search metadata.review_text rather than review_text field.

6.1.2 Database creation and access

1. IBM Cloud database services can be used to store data for the crawler to crawl it and push into the WDS collection.

2. Postgres DB or DB2 is recommended if you are intending to use IBM Cloud Database services.

3. A sample Postgres DB can be created in IBM Cloud which can then host the Amazon Review Data. Below are the necessary credentials you will need to access the DB

   i. Username: xxxxxx

   ii. Password: xxxxxxxxx

   iii. URL: jdbc:postgresql://bluemix-sandbox-dal-9-portal.5.dblayer.com:18975/*dbname*

   iv. Table Name: XXXXX

4. The database can be accessed/modified using API calls or using a simple java application with jdbc connection.

5. The application needs to reference a jdbc driver jar for postgres which can be found at https://jdbc.postgresql.org/download.html

6. A sample code for accessing and creating a table in the postgres service hosted in Cloud is shown below. In a similar way, an entry can be added to the row using simple SQL commands

WDSDatabase.java

```
import java.sql.Connection;

public class WDSDatabase {

    public static void main(String[] args) {
        try {
            Class.forName("org.postgresql.Driver");
        }
        catch (java.lang.ClassNotFoundException e) {
            System.out.println(e.getMessage());
        }

        //Postgres Credentials
        String Url = "jdbc:postgresql://bluemix-sandbox-dal-9-portal.5.dblayer.com:18975/dbname";
        String Username = "Username";
        String Password = "Password";

        try {

            Connection db = DriverManager.getConnection(Url, Username, Password);
            Statement st = db.createStatement();

            st.executeUpdate("CREATE TABLE reviews(uuid SERIAL, reviewer_id varchar(20), product_id varchar(20), helpful int, "
                    + "not_helpful int, review_text text, rating real, summary varchar(150), product_name varchar(250), "
                    + "review_date date, time_inserted date, PRIMARY KEY (uuid) )");

            st.close();

        }

        catch (Exception e) {
            System.out.println(e.getMessage());
        }
    }

}
```

6.1.3 Database table schema for our Amazon Product Review use case

1. The database table “reviews” should contain all the required properties of a review arranged as table columns

2. A typical schema of the table for our use case is shown below

```
CREATE TABLE reviews(
        uuid SERIAL,
        reviewer_id varchar(20),
        product_id varchar(20),
        helpful int,
        not_helpful int,
        review_text text,
        rating real,
        summary varchar(150),
        product_name varchar(250),
        review_date date,
        time_inserted date,
        PRIMARY KEY (uuid)
 );
```

3. The column "uuid" which is of the serial datatype is used as an auto increment index for the table which is also a primary key

4. Indexing the table helps in fast retrieval of the data items

7 Relevancy Training

During this lab exercise, we have facilitated the configuration steps by providing you with sample content data and example queries. In this section, we will go over how you can improve your answer results by training the model. Relevancy training (RT) is optional; if the results of your queries meet your needs, no further training is necessary. In cases with realy hard questions to answer where there a lot of possible documents linked to queries, relevancy training is most useful. RT can be done programatically or in the tooling. In this section, we will show you how to train the model in the tooling.

7.1 Overview and Requirements

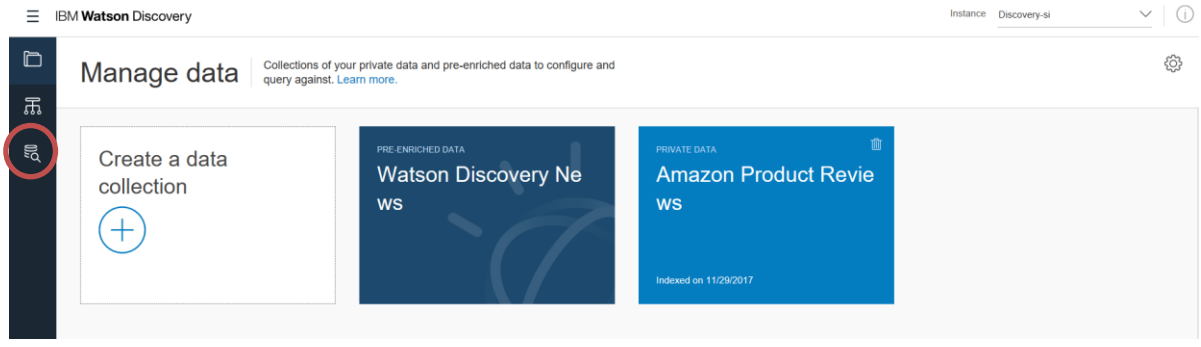
The relevance of natural language query results can be improved in Watson Discovery service with training. In order to train Watson, you'll need to provide example queries that are representative of queries your user enters and provide ratings to say which results for each query are relevant and not relevant. The requirements for Discovery to begin applying your ratings include training a minimum of 49 queries or possibly more (Watson will give you feedback if it needs more queries in order to train) and applying both Relevant and Not relevant ratings. Only rating Relevant documents will not provide the data needed.

In this section, we will only train two example queries to show the training process flow. In real life instances, a minimum of 49 queries must be trained for Discovery to begin applying rating.

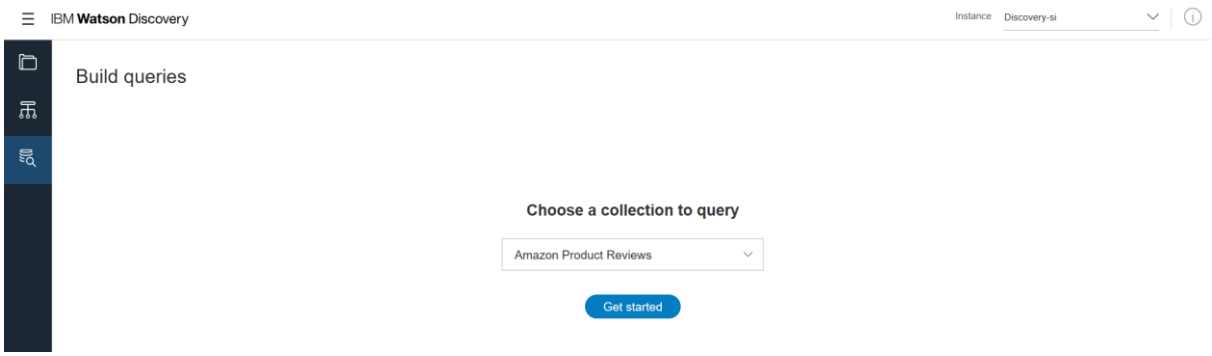
7.1.1 Train Watson to Improve Results

We will train the results for the query "time" to help the model return results to "time" that relates to to cooking times. We also will train the results to "problem" to see problems that reviewers have found. You can experiment and add more queries to train. Remember to rate them with both Relevant and Not relevant.

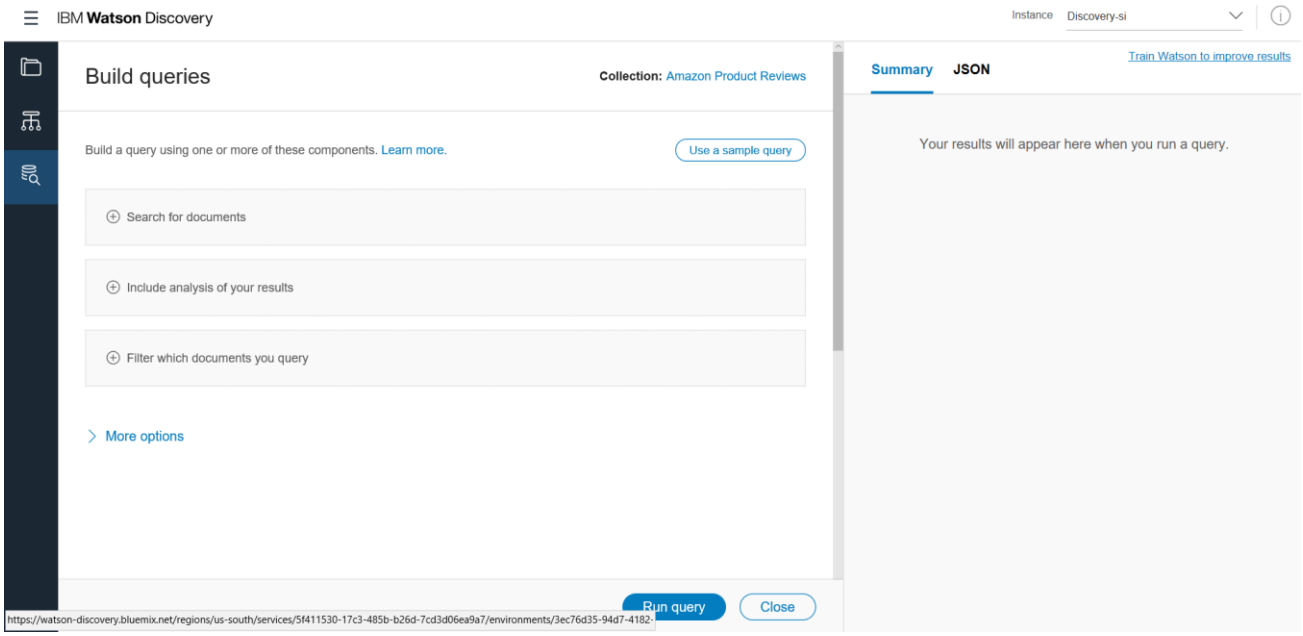
1. In order to train your model, you must head over to the "Build queries" page. To do so, click the "build queries" icon on the lefthand side bar.



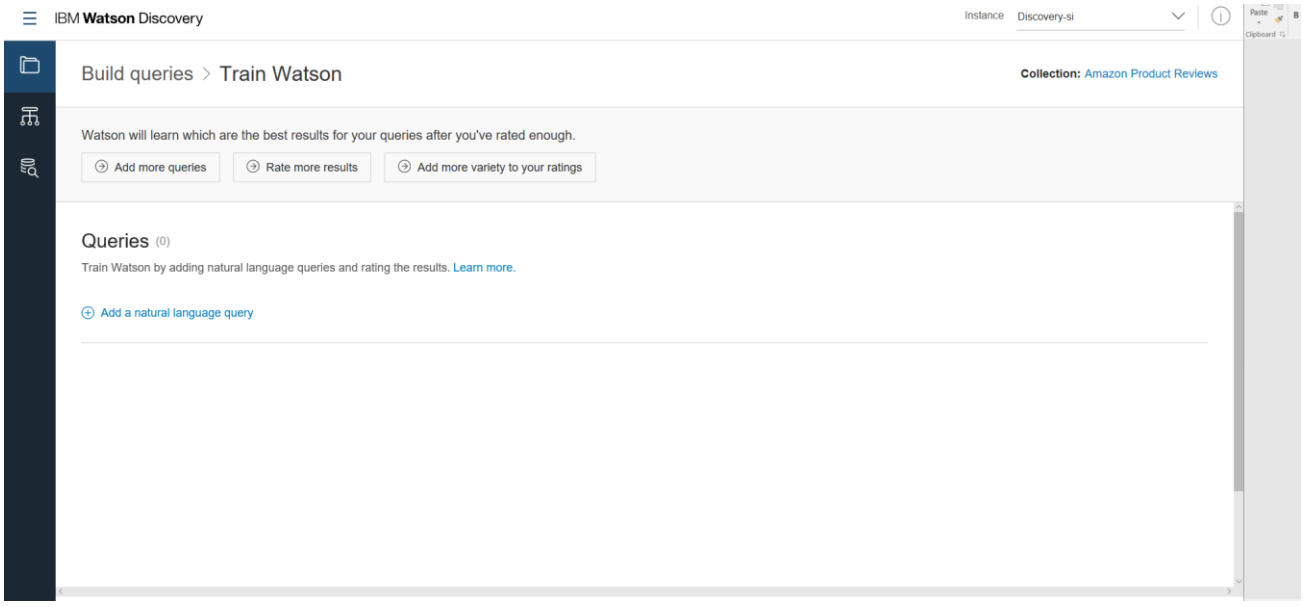
2. Choose the "Amazon Product Reviews" collection to query from the dropdown menu and click "Get Started"



3. Click “Train Watson to improve results” in the top right corner to get to the relevancy training page

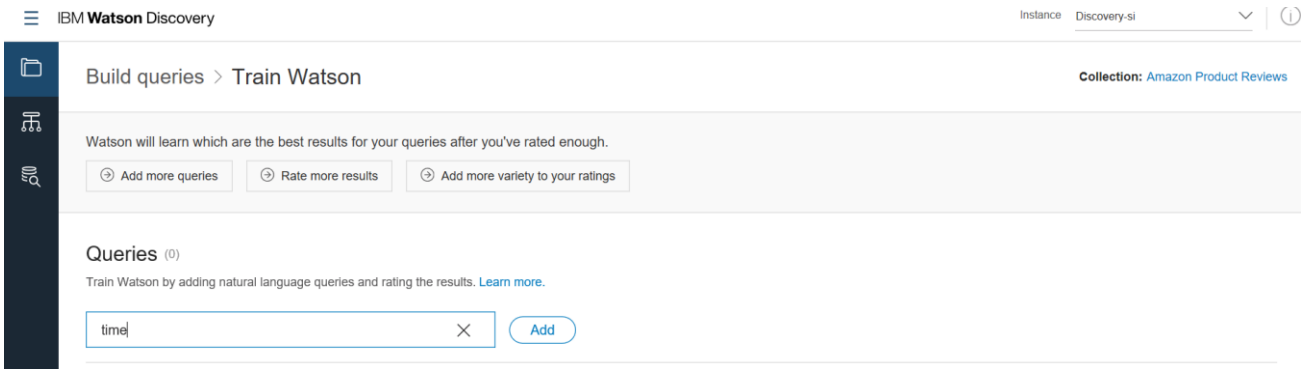


4. Click the “Add a natural language query” link that is in blue

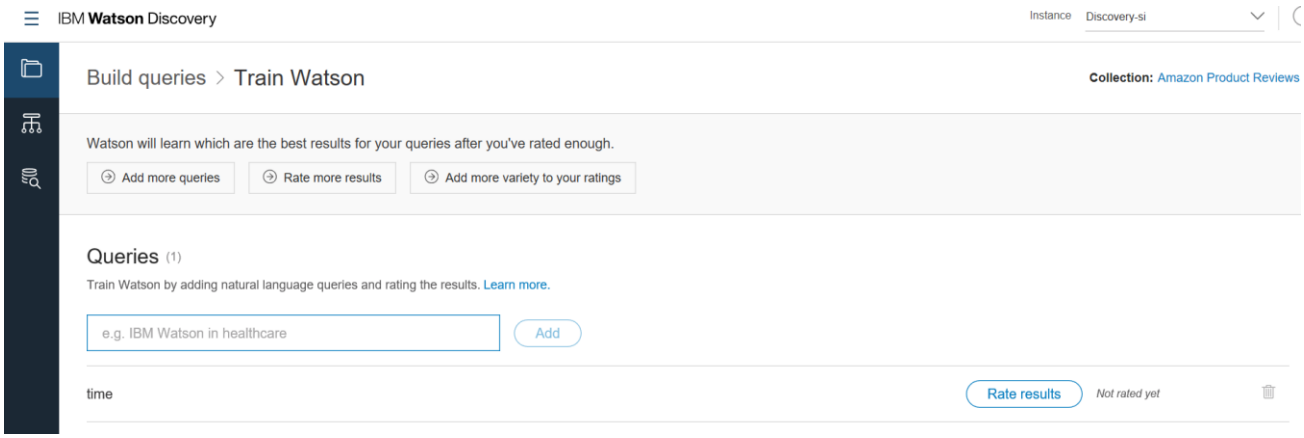


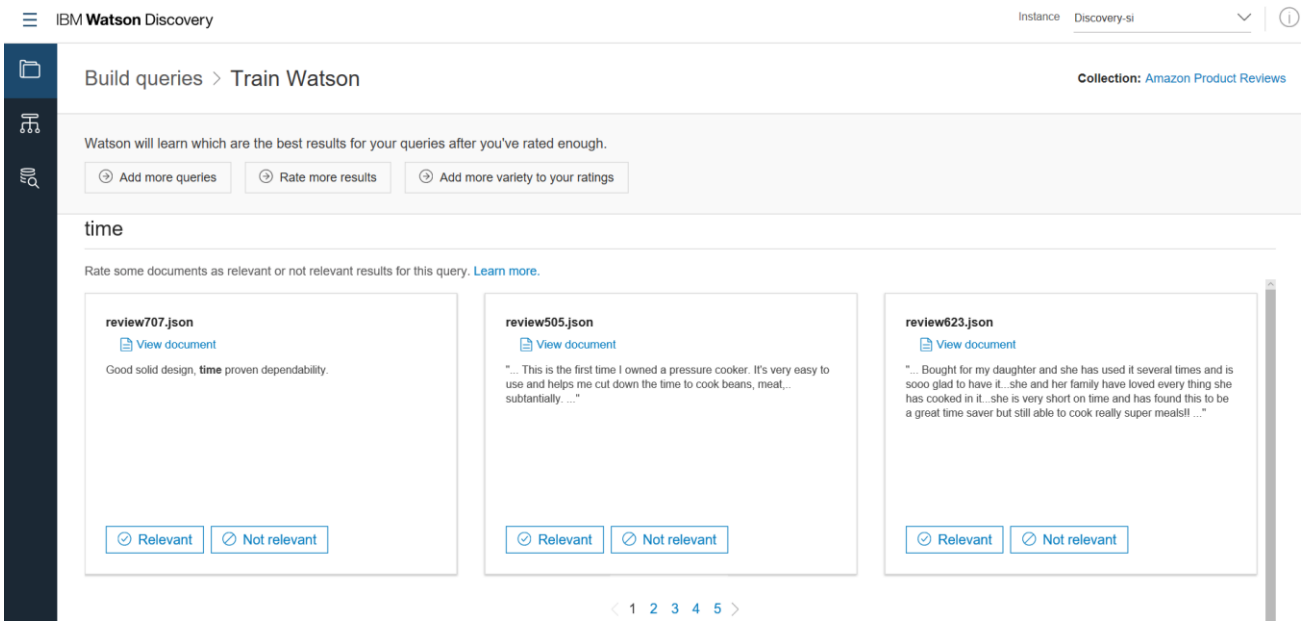
5. Click “Learn more” if you’d like view IBM Cloud documentation on improving result relevance with the tooling

6. Type in “time” into the query box and click the “Add” button

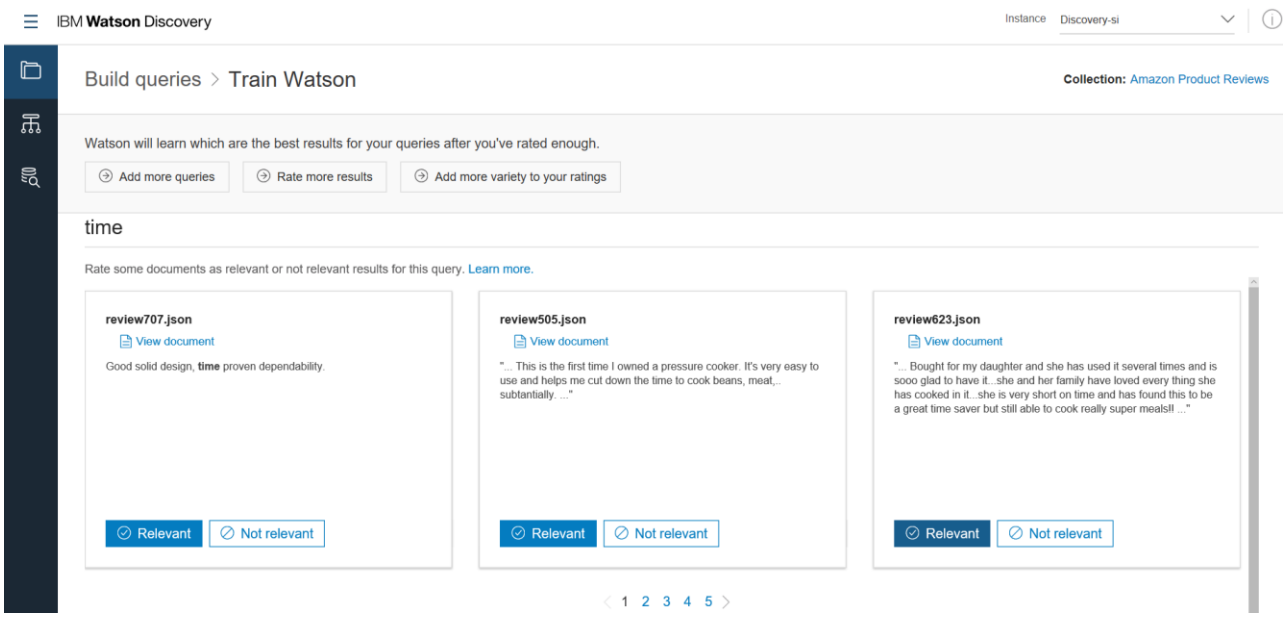


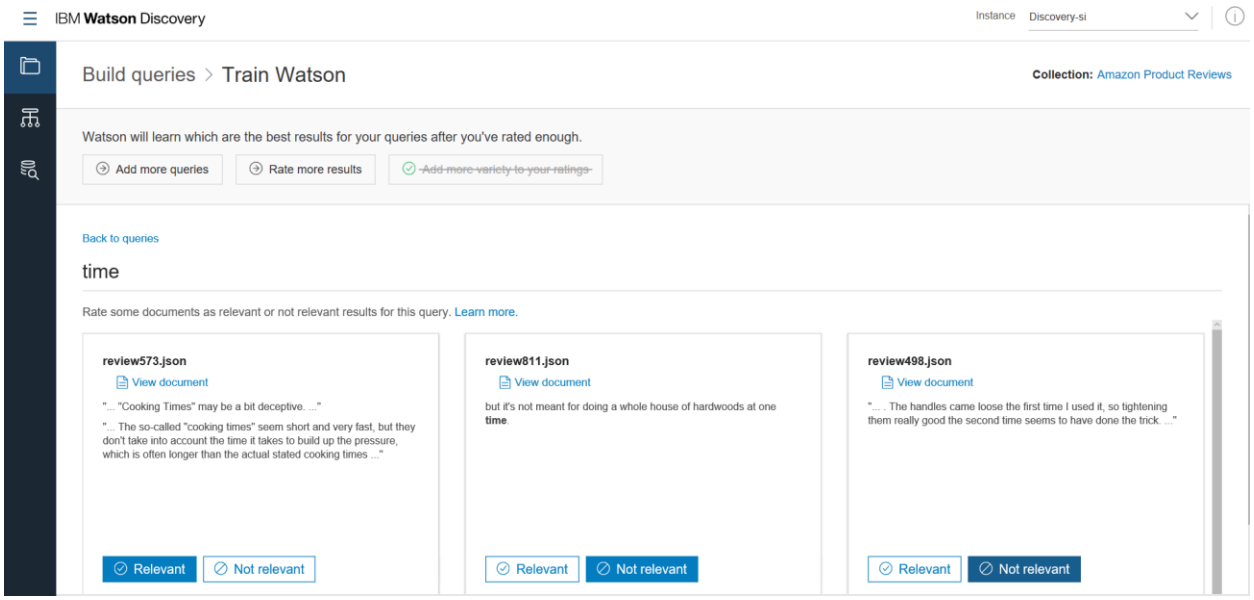
7. Click “Rate results” on the bottom right hand side in blue

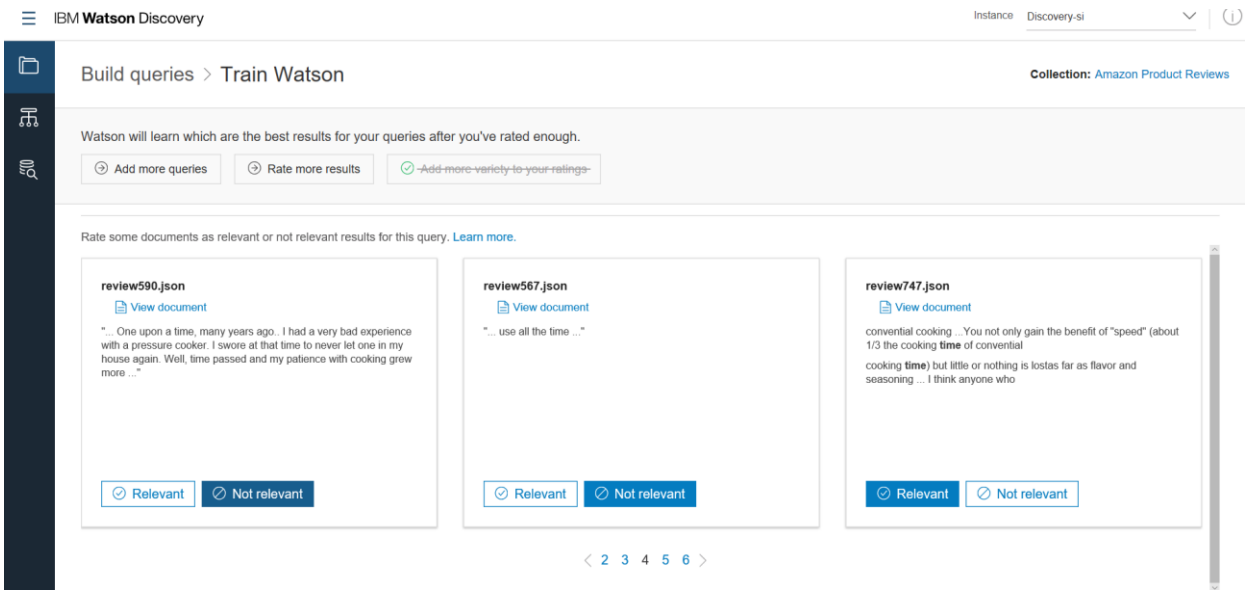




8. Rate “Relevant” or “Not relevant” based on if the “time” mention is about cooking times

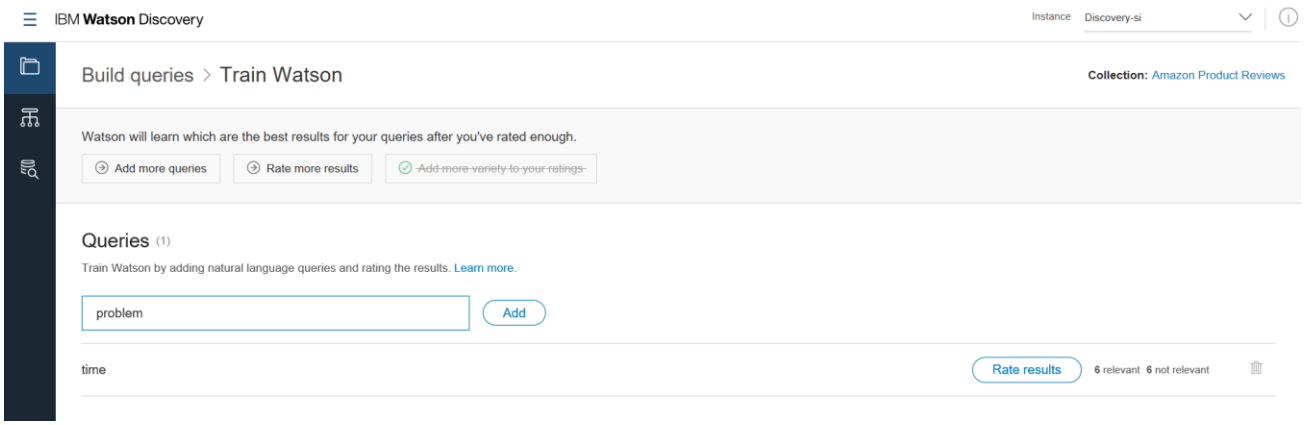


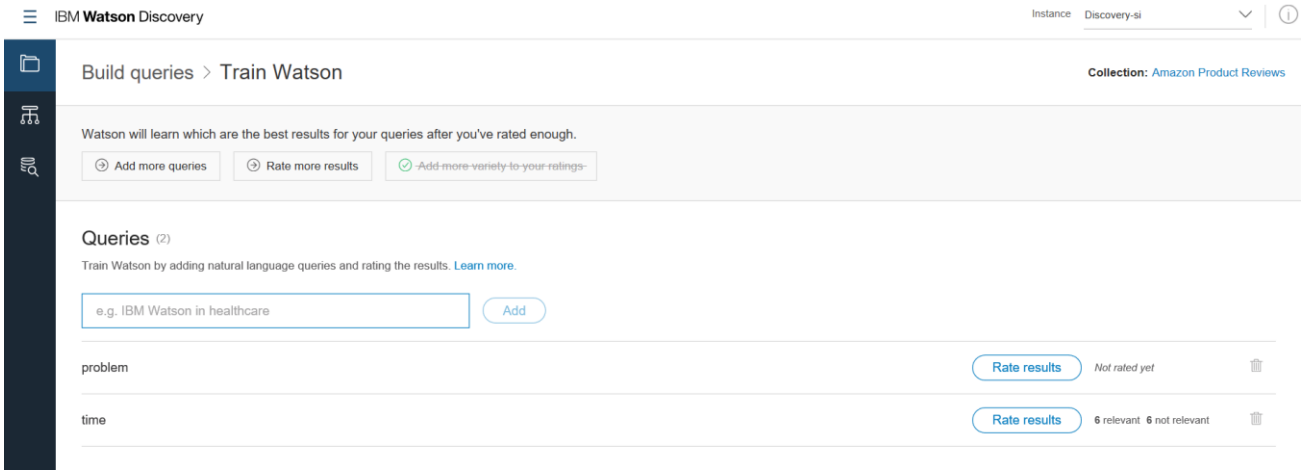




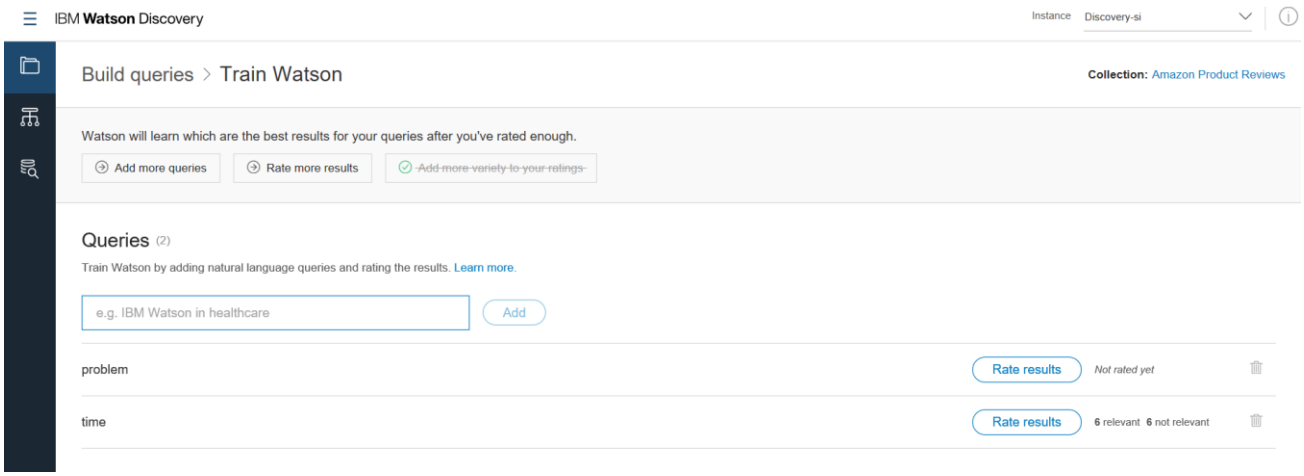
9. Click “back to queries” and then “add a natural language query”

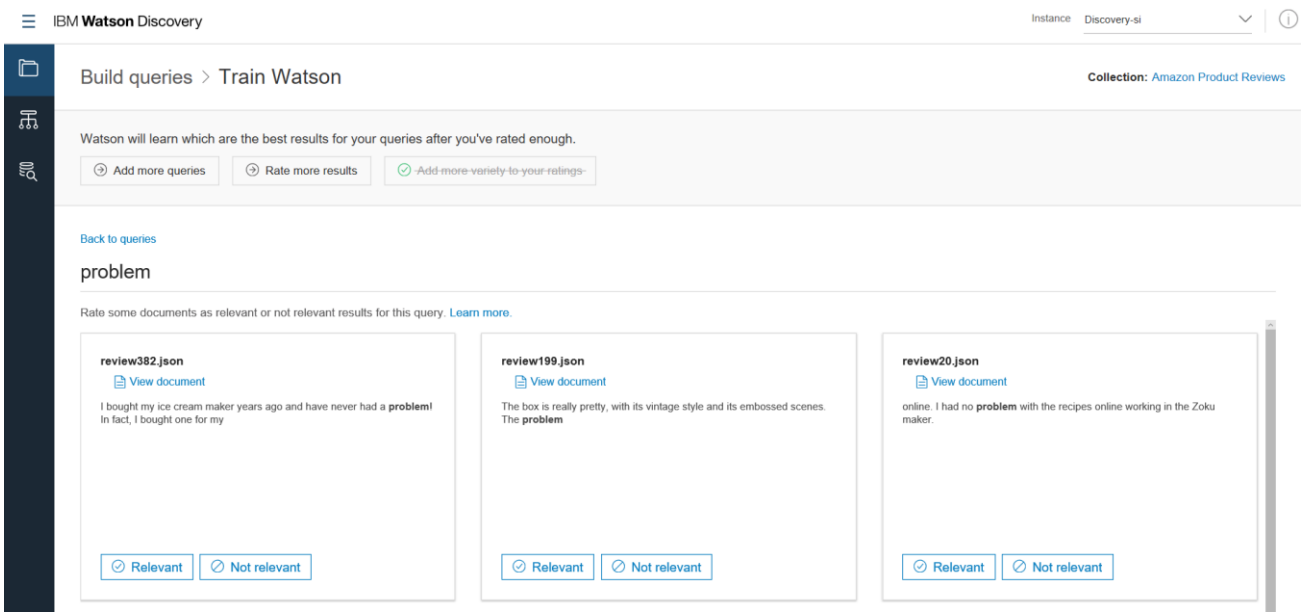
10. Type in “problem” into the query box and click “Add” to the right of the box



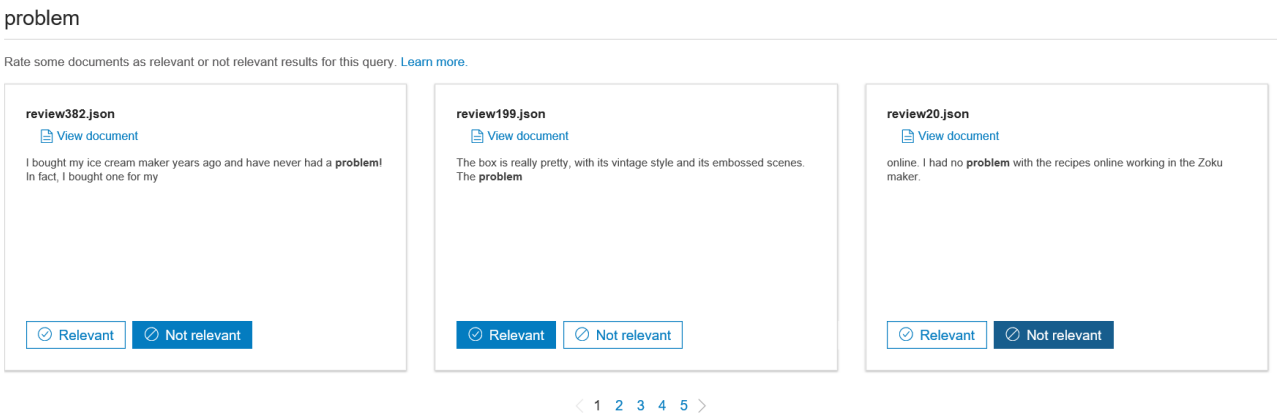


11. Click “Rate results” that corresponds to “problem” on the bottom right side

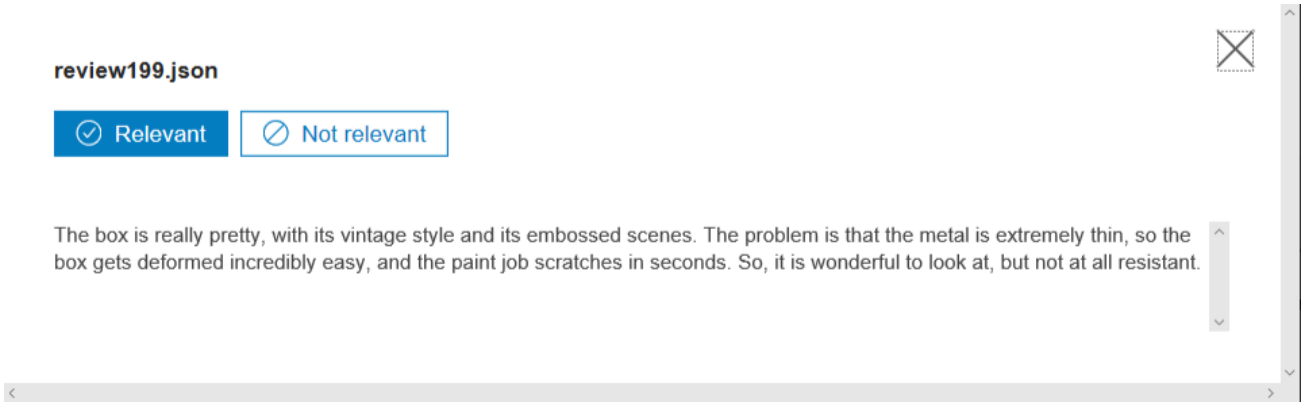




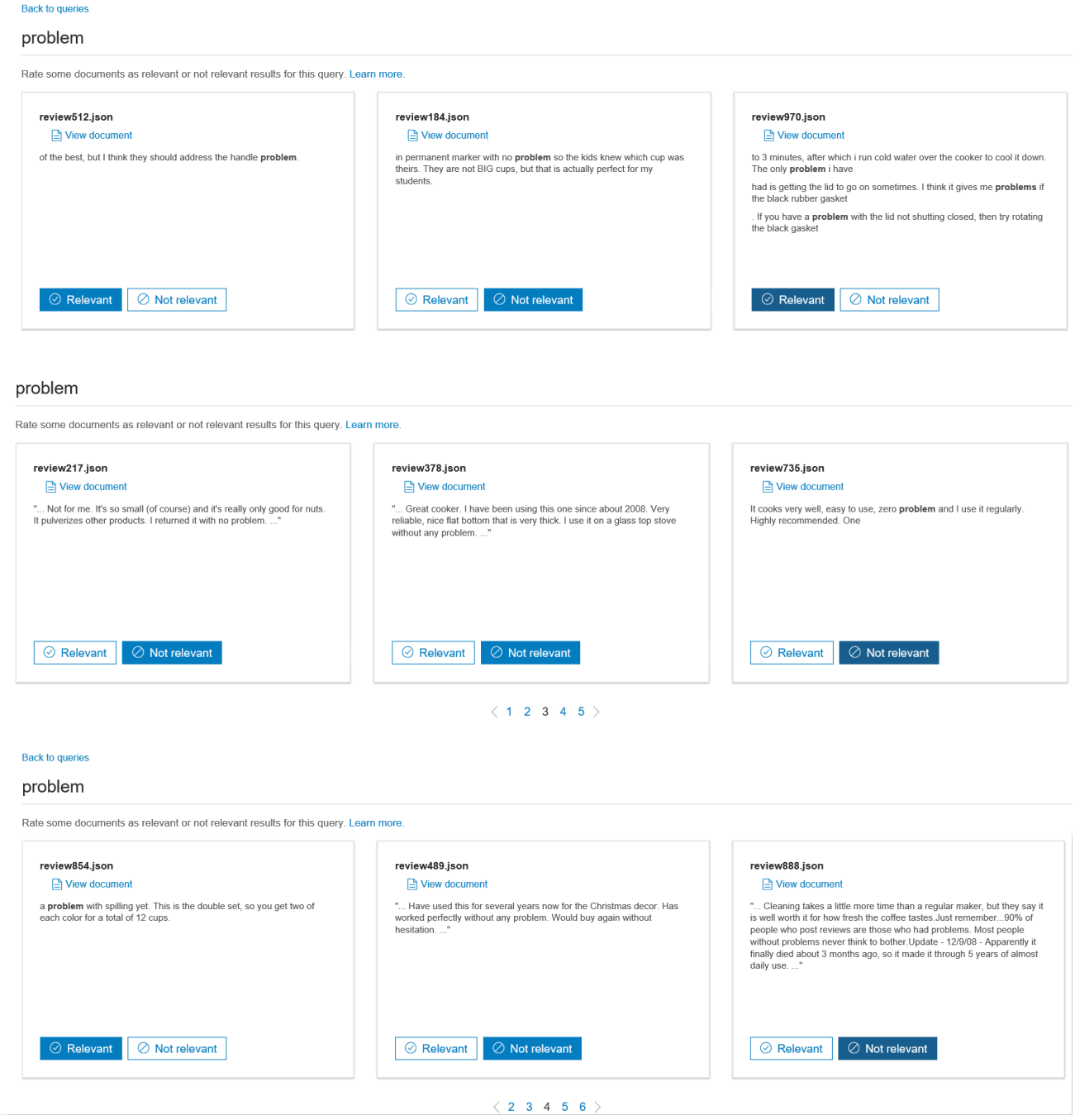
12. Rate “Relevant” or “Not relevant” based on if the “problem” mention is about an issue.



Review382.json mentions that the reviewer “never had a problem” and review20.json mentions that the review “had no problem” so both results should be rated Not relevant. Meanwhile review199 specifies a problem. To view the full review, click View document in blue under the review title.



Click the X and continue on the next page to keep rating the results



13. Click “Back to queries”