IBM Developer
SKILLS NETWORK

Lab 2: Creating a Discovery Collection (1 hour)

Objective for Exercise:

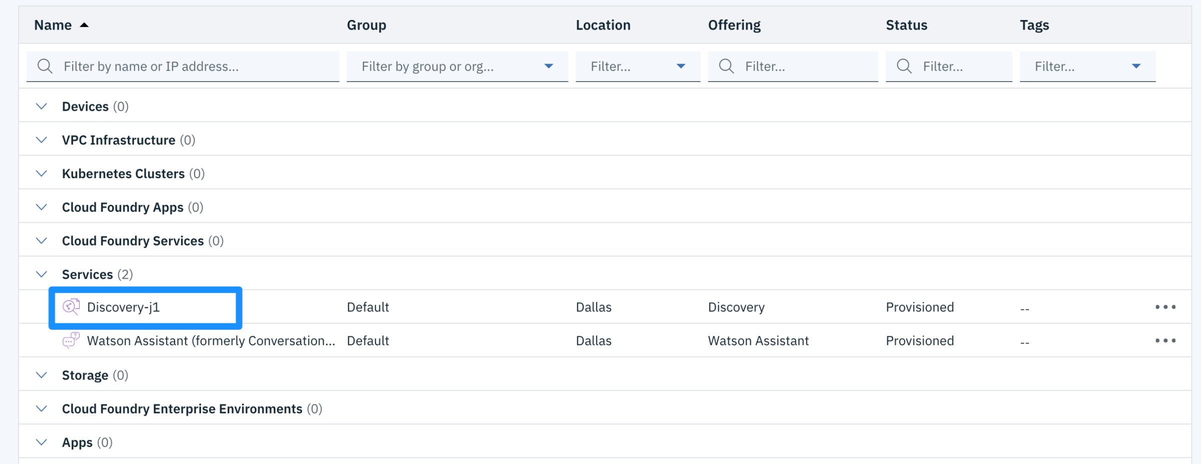
- How to use a Watson Discovery Service
- How to upload Documents to a new Discovery collection.

Since you're taking this course you should already be familiar with Watson Assistant. What you might not be familiar with is Watson Discovery. In Lab 1 you created a Watson Discovery service instance in your IBM Cloud account (if you haven't, go back and consider doing that now); here we'll create a Discovery Collection and, in the process, learn more about this powerful insight tool.

Explore the default News Collection

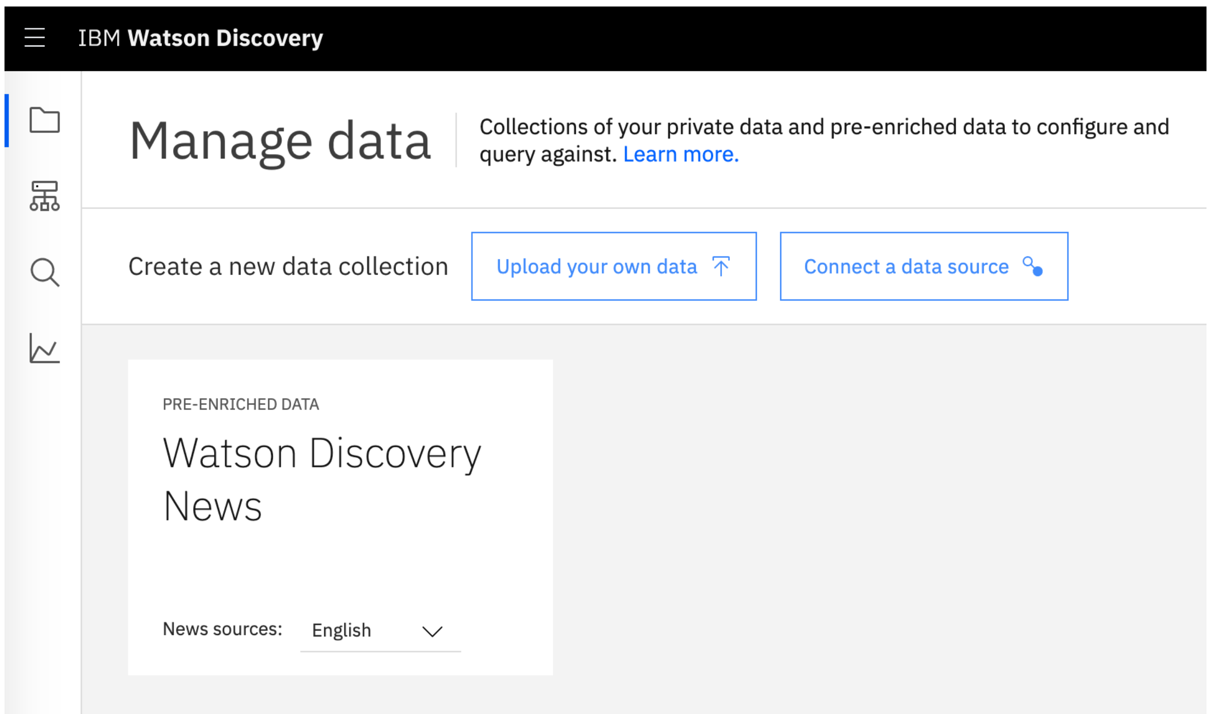
We'll start off by launching the Discovery tool and exploring the default News collection.

1. From your Dashboard, **click on the Discovery service you created** in Lab 1. (Click on the name, not the icon next to it.)



Name ▲	Group	Location	Offering	Status	Tags	
Filter by name or IP address...	Filter by group or org...	Filter...	Filter...	Filter...	Filter...	
Devices (0)						
VPC Infrastructure (0)						
Kubernetes Clusters (0)						
Cloud Foundry Apps (0)						
Cloud Foundry Services (0)						
Services (2)						
Discovery-j1	Default	Dallas	Discovery	Provisioned	--	•••
Watson Assistant (formerly Conversation...	Default	Dallas	Watson Assistant	Provisioned	--	•••
Storage (0)						
Cloud Foundry Enterprise Environments (0)						
Apps (0)						

2. A new page will appear. **Click on *Launch Watson Discovery***. You should see the Manage Data section of the tool.



IBM **Watson Discovery**

Manage data

Collections of your private data and pre-enriched data to configure and query against. Learn more.

Create a new data collection

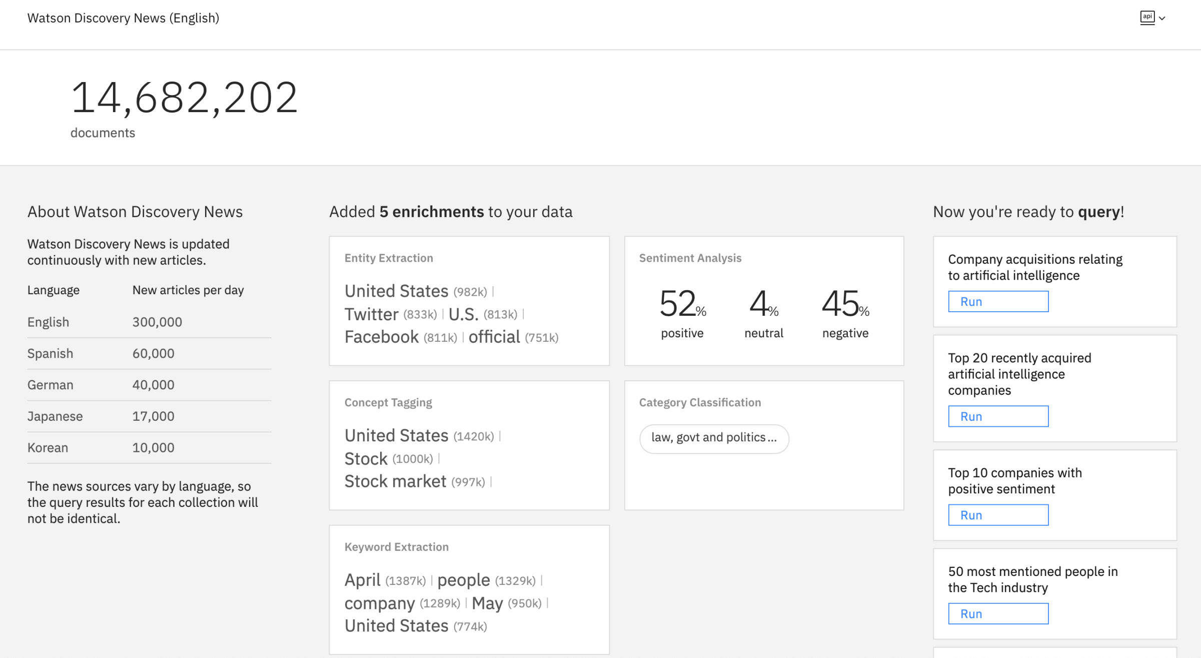
PRE-ENRICHED DATA

Watson Discovery News

News sources: English

3. You'll notice that by default the Watson Discovery News collection has been pre-loaded for you. The Upload your own data and Connect a data source buttons allow you to work with your own data. For the time being, let's explore the News data collection. Go head and **click on the *Watson Discovery News tile***.

4. You should see statistical details about the collection, similarly to the image below.



Watson Discovery News (English)

14,682,202
documents

About Watson Discovery News

Watson Discovery News is updated continuously with new articles.

Language	New articles per day
English	300,000
Spanish	60,000
German	40,000
Japanese	17,000
Korean	10,000

The news sources vary by language, so the query results for each collection will not be identical.

Added **5 enrichments** to your data

Entity Extraction
United States (982k) | Twitter (833k) | U.S. (813k) | Facebook (811k) | official (751k)

Sentiment Analysis
52% positive
4% neutral
45% negative

Concept Tagging
United States (1420k) | Stock (1000k) | Stock market (997k) |

Category Classification
law, govt and politics...

Keyword Extraction
April (1387k) | people (1329k) | company (1289k) | May (950k) | United States (774k)

Now you're ready to **query**!

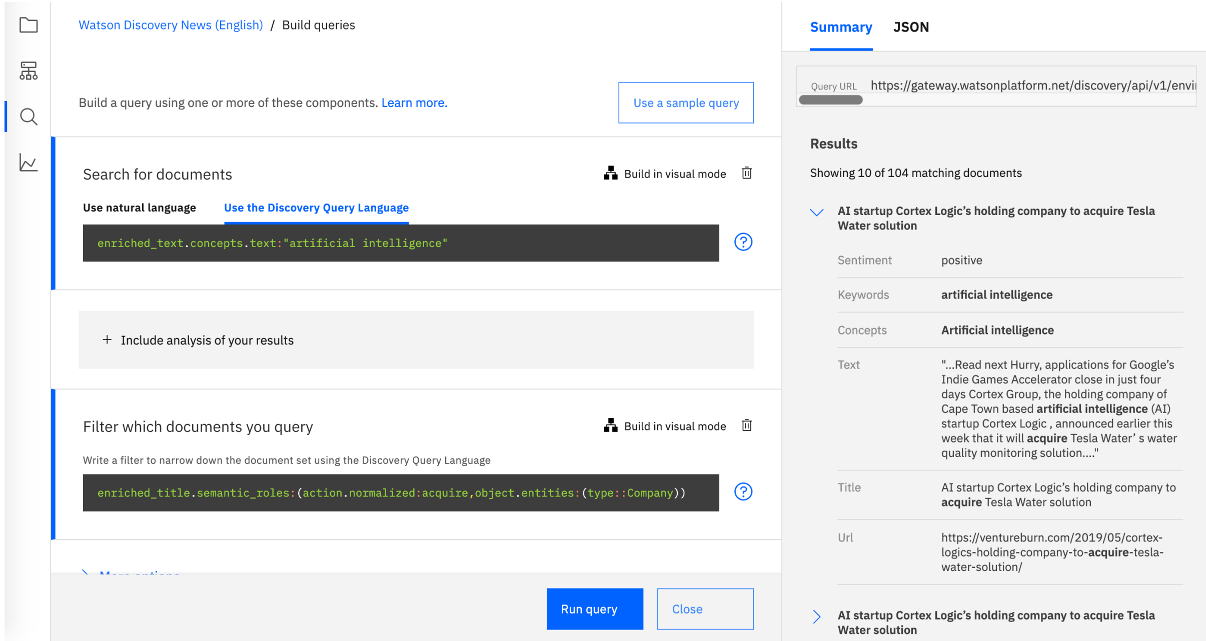
Company acquisitions relating to artificial intelligence
Run

Top 20 recently acquired artificial intelligence companies
Run

Top 10 companies with positive sentiment
Run

50 most mentioned people in the Tech industry
Run

5. Select one of the pre-made queries on the right and click Run. Any query of your choice will do. You should see a result somewhat similar to the one shown in figure (only the query might be a different one, depending on which one you selected).



Watson Discovery News (English) / Build queries

Build a query using one or more of these components. Learn more.

Use a sample query

Search for documents

Build in visual mode

Use natural language | Use the Discovery Query Language

```
enriched_text.concepts.text:"artificial intelligence"
```

+ Include analysis of your results

Filter which documents you query

Build in visual mode

Write a filter to narrow down the document set using the Discovery Query Language

```
enriched_title.semantic_roles:(action.normalized:acquire,object.entities:(type::Company))
```

Run query | Close

Summary | JSON

Query URL https://gateway.watsonplatform.net/discovery/api/v1/envi

Results

Showing 10 of 104 matching documents

AI startup Cortex Logic's holding company to acquire Tesla Water solution

Sentiment	positive
Keywords	**artificial intelligence**
Concepts	**Artificial intelligence**
Text	"...Read next Hurry, applications for Google's Indie Games Accelerator close in just four days Cortex Group, the holding company of Cape Town based **artificial intelligence** (AI) startup Cortex Logic , announced earlier this week that it will **acquire** Tesla Water' s water quality monitoring solution...."
Title	AI startup Cortex Logic's holding company to **acquire** Tesla Water solution
Url	https://ventureburn.com/2019/05/cortex-logics-holding-company-to-**acquire**-tesla-water-solution/

AI startup Cortex Logic's holding company to acquire Tesla Water solution

You'll notice that the third icon on the left (the magnifying glass icon) is now selected, since we are no longer within Manage Data but rather in the Query section.

If you click + Search for documents, you'll also notice that there are three ways to query data: 1) Using natural language; 2) Using the Discovery Query Language, which is what the pre-made queries use; 3) Using visual mode.

We won't be spending much time learning the intricacies of the Discovery Query Language, since our chatbot will simply pass the natural language queries from the use to the Discovery service to retrieve relevant articles ranked in order of relevancy. However, feel free to spend some time analyzing these sample queries which use the Discovery Language, as well as reading the relevant documentation.

On the right of the page, you'll see the output of the query over two tabs: A human-readable Summary tab, and a JSON output tab(which you would work with when querying the service through its API).

Expanding the relevant documents returned by the query, will highlight how they are relevant to the query. For example, in my screenshot above, you'll notice that the query is asking for AI acquisitions in the news, so the top result (among 104 matching document) is one that features the word acquire in the title, URL, and text fields. Watson uses its AI capabilities to enrich the raw documents that are uploaded in Discovery. So much of the magic lies in its ability to classify and tag the documents so that they are easy to query. In this case, my top result for this query, has a positive sentiment, and was deemed as relevant to the concept of Artificial Intelligence by Watson.

Feel free to explore a bit more on your own, perhaps altering the query yourself or by trying a natural language query. When you're ready, move on to the next section where we'll create a collection of our own, which is much more relevant to the chatbot that we'll build.

Create a Coursera course catalog collection

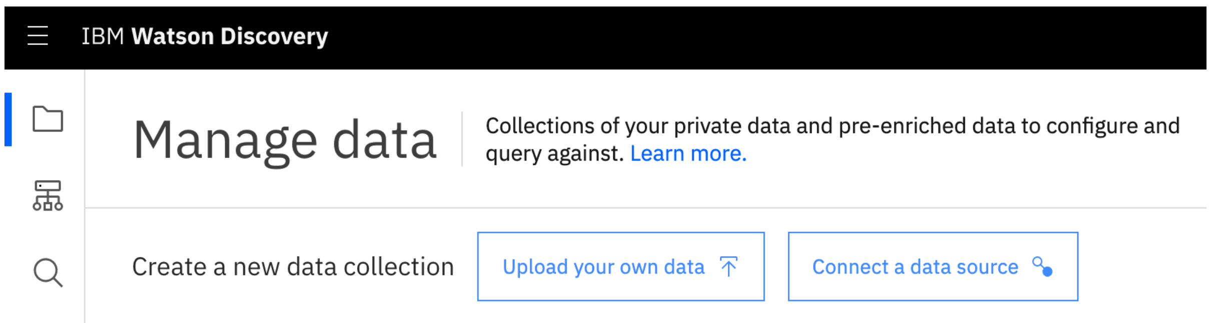
In the next module, we'll create a barebone student advisor chatbot for Coursera. It should be able to, among other functions, point people to courses relevant to the topic the user is looking for. So when the user asks, "Do you have courses on Python?" the chatbot should be able to reply with a list of courses.

The problem with the traditional approach of hardcoding courses in the node responses within Watson Assistant is that Coursera has thousands of courses. It's just not feasible to manually add all of them to a dialog.

So, instead, we'll leverage Discovery and connect it to Watson Assistant. But before we can do all that, we'll need to have a Coursera course collection uploaded within Discovery. Let's do that now.

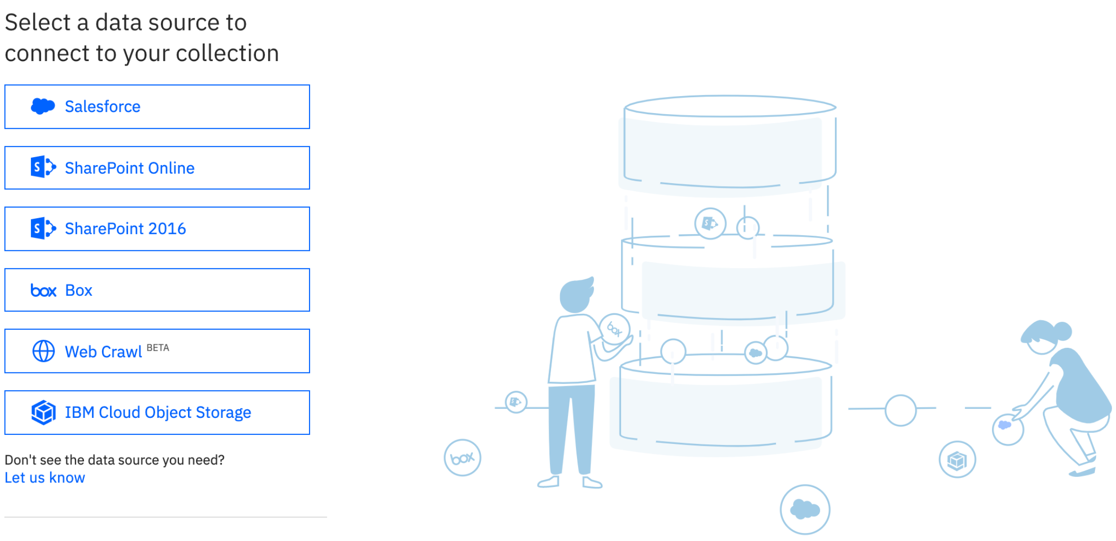
Two ways to upload data

There are two main ways to upload data to Discovery: uploading documents and connecting to a data source. If you head over to the Manage Data section of your Discovery instance, you'll see the two corresponding buttons.



Collections of your private data and pre-enriched data to configure and query against. Learn more.

In case you're curious about which data sources are available, you can click on that second button and you see a list of services similar to the one below:



Select a data source to connect to your collection

Don't see the data source you need?

The most relevant source here would be Web Crawl, which allows us to specify a URL and the collection will store each page connected to that URL as a document. You can even specify how deep the crawler should go (e.g., how many links it should follow before stopping).

For the sake of showing how to import documents, and to spare Coursera's servers from thousands of people all crawling their course catalog, I prepared a collection of 500 courses of theirs. The reason why I limited it is that your Lite free account is limited to 1000 documents a month, and you'll need more documents for a second collection during your project in Module 6.

Let's see how to upload them.

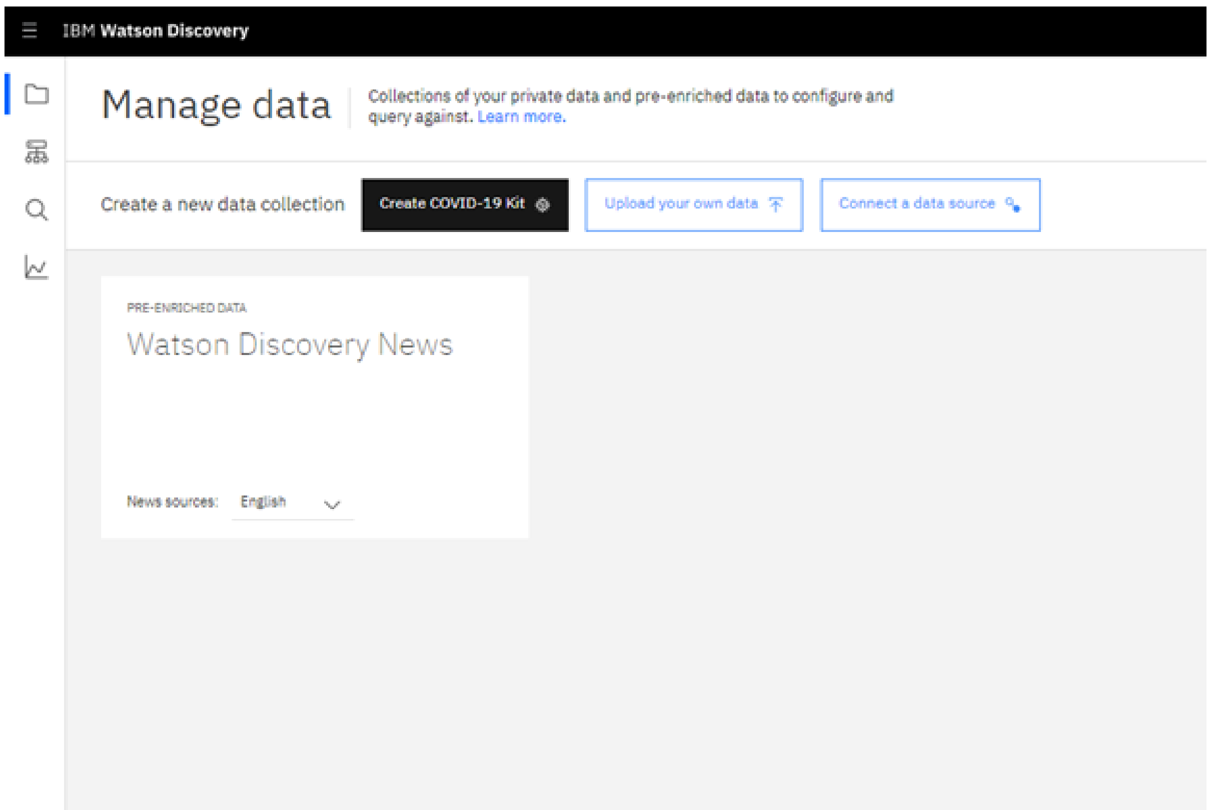
Upload Coursera course documents

> **NOTE:** In order to save your upload time, we have divided the data into two parts, one with 5 documents and other with remaining 495 documents.

1. **Download the following zip file** containing a list of courses as JSON files and **unzip it**.

<a href=https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBMDeveloperSkillsNetwork-CB0106EN-SkillsNetwork/labs/Module%202-coursera/data/5-Coursera-courses.zip?utm_medium=Exinfluencer&utm_source=Exinfluencer&utm_content=000026UJ&utm_term=10006555&utm_id=NA-SkillsNetwork-Channel-SkillsNetworkCoursesIBMDeveloperSkillsNetworkCB0106ENSkillsNetwork20719128-2021-01-01>5-coursera-courses.zip

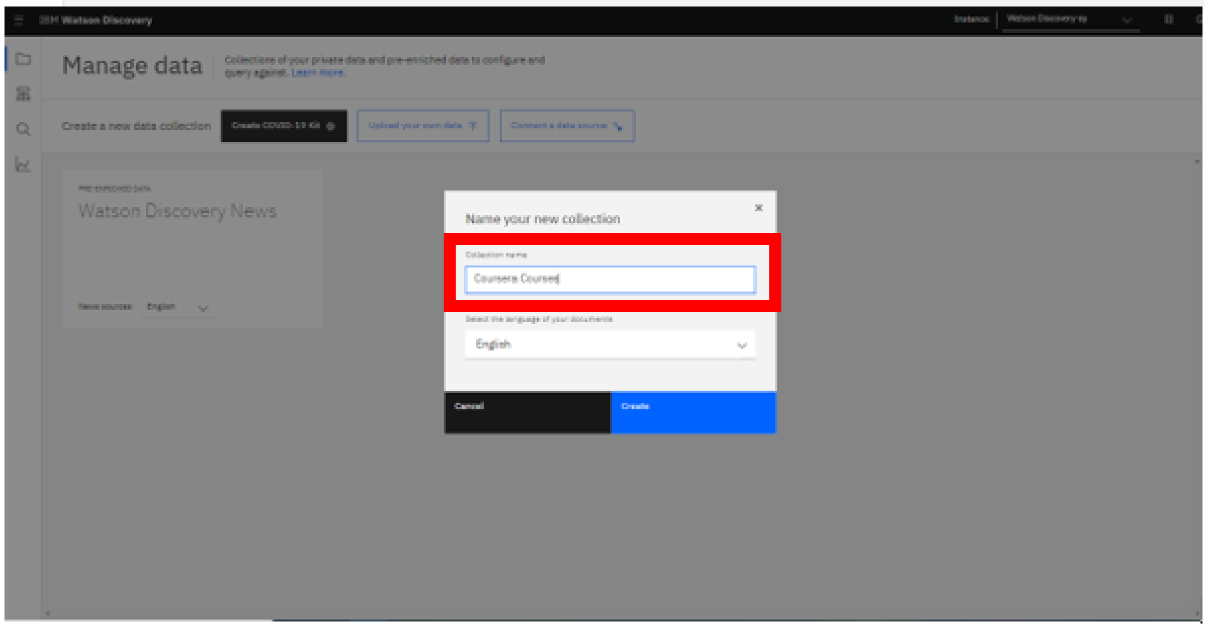
2. From the *Manage Data* section of your Discovery instance, **select *Upload your own data***. You'll be asked to give a name to your collection. **Call it Coursera Courses** or something similar and **click Create**.



Manage data

Collections of your private data and pre-enriched data to configure and query against. Learn more.

Create a new data collection

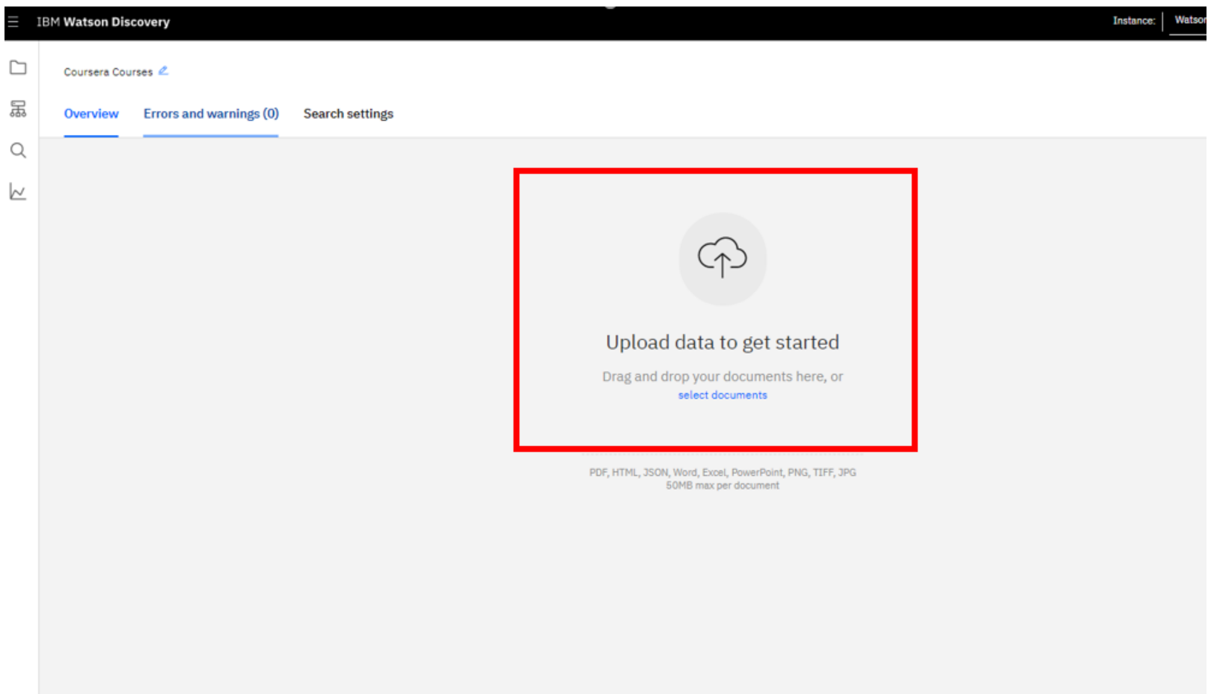


Name your new collection

Collection name: Coursera Courses

Select the language of your documents: English

3. You'll be prompted to upload documents. **Click on the upload icon and add the files you extracted**. You can use CMD+a or CTRL+a to select all of them at once in the upload dialog. Alternatively, you can simply drag and drop the files on the page.



Upload data to get started

Drag and drop your documents here, or select documents

PDF, HTML, JSON, Word, Excel, PowerPoint, PNG, TIFF, JPG
50MB max per document

> NOTE: Processing the data will take a few minutes. If you see Errors and warnings during the process, simply ignore the them. It gives a warning because the course document contains a reserved field of id.

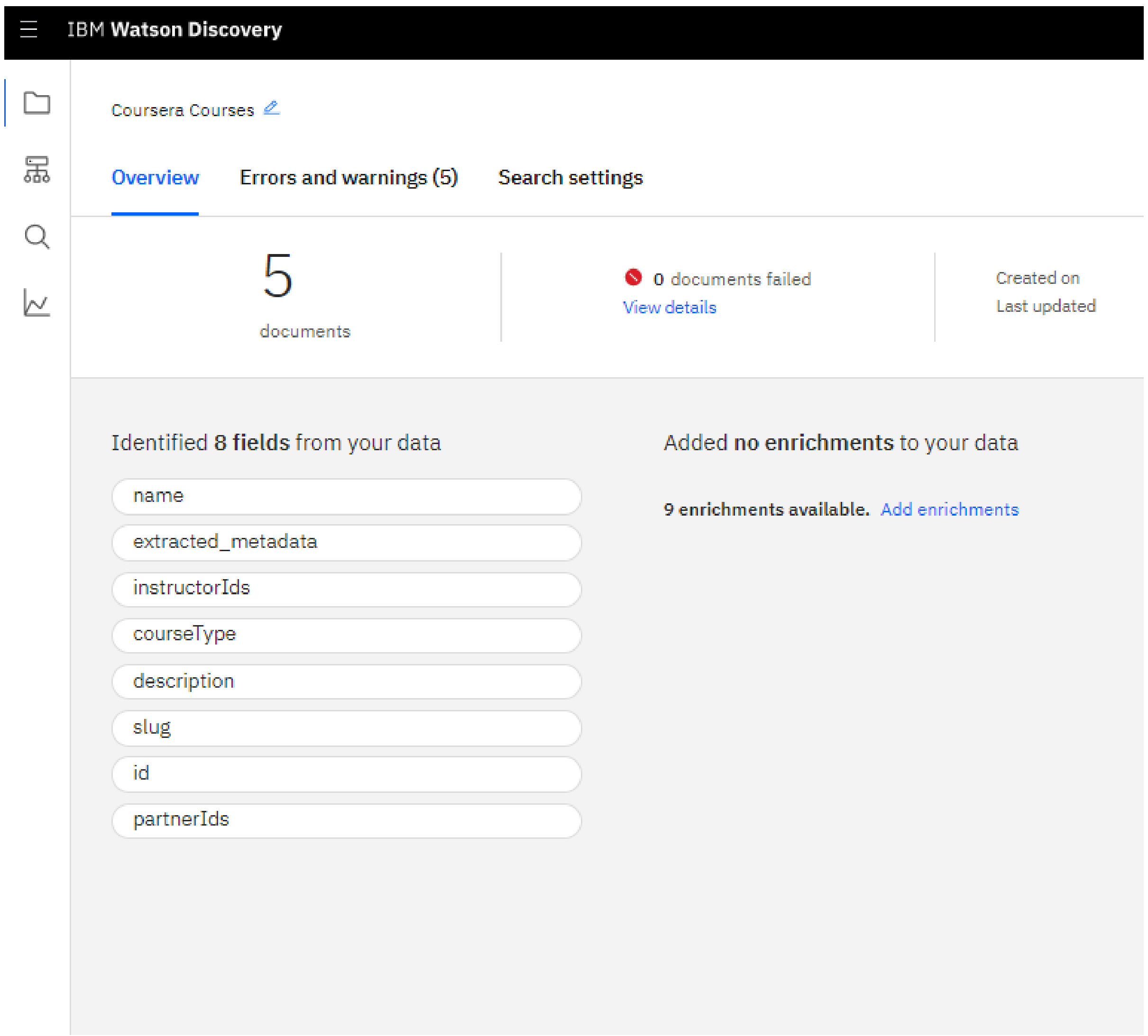
```
{
  "courseType":"v2.ondemand",
  "description":"Gamification is the application of game elements and digital game design techniques to non-game problems, such as business and social impact challenges. This course will teach you the mechanisms of gamification, why it has such tremendous potential, and how to use it effectively. For additional information on the concepts described in the course, you can purchase Professor Werbach's book For the Win: How Game Thinking Can Revolutionize Your Business in print or ebook format in several languages.",
  "id":"69Bku0KoEeWZtA4u62x6lQ",
  "slug":"gamification",
  "instructorIds":[
     "226710"
  ],
  "specializations":[

  ],
  "partnerIds":[
     "6"
  ],
  "name":"Gamification"
}
```

The most important fields for us are name, slug, and description. Name is the title of the course, slug the path to the course that we'll append to *https://www.coursera.org/learn/* to generate the course URL, and description gives Watson Discovery enough info to determine the relevance of the user query to the course content.

Remember, Discovery is only as good as the data you feed it.

4. Once the import is completed, you should see something similar to the image below. The number of imported documents is reported, as well as the fields that were identified by Watson. Name, slug, and description were identified correctly, so we are good as far as we are concerned.

> *Don't worry if you see a warning about the data containing an id. It just means that Discovery has already a protected id field, so the one we uploaded in our JSON will be ignored.*



5 documents

0 documents failed

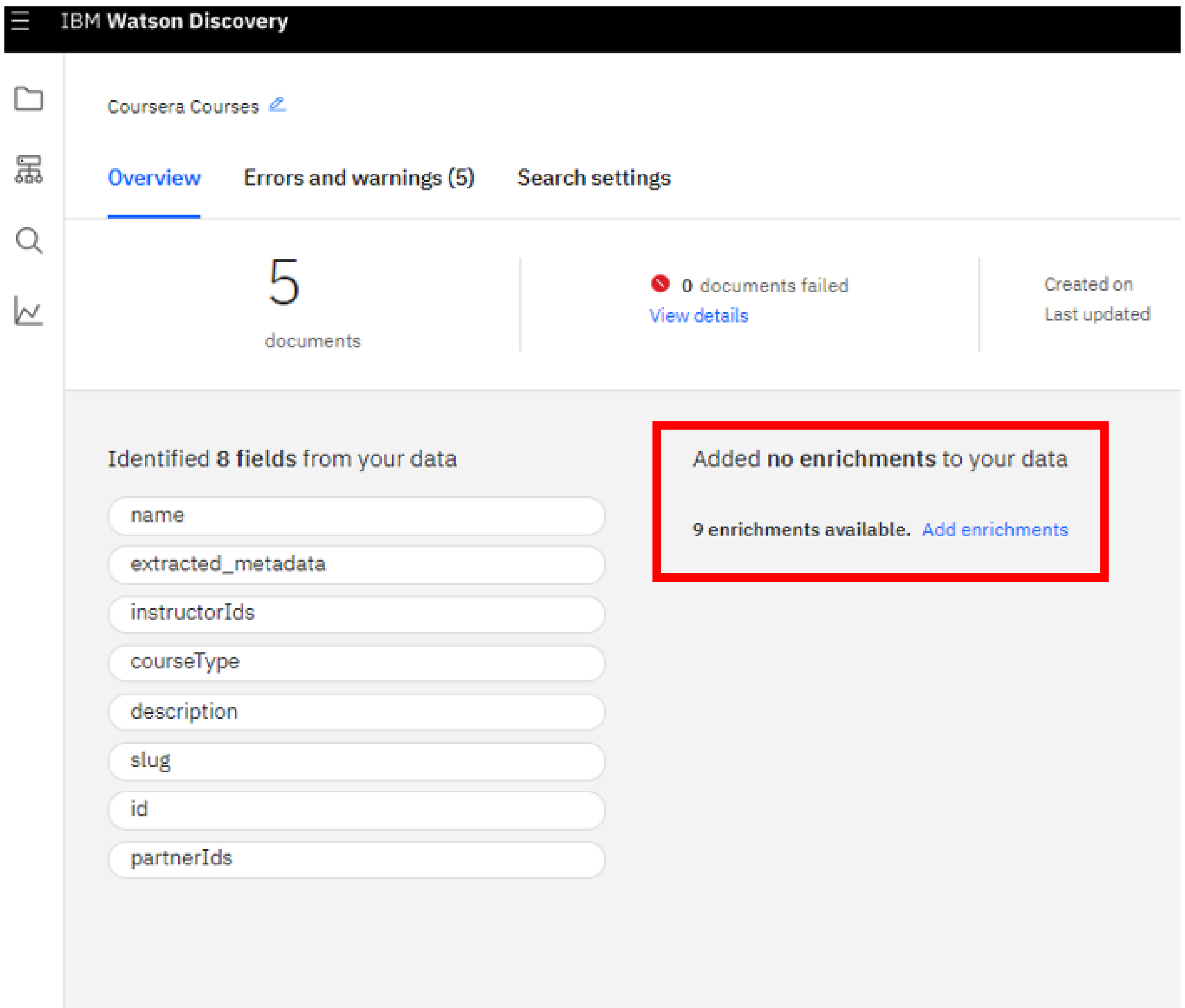
Identified 8 fields from your data

name
extracted_metadata
instructorIds
courseType
description
slug
id
partnerIds

Added no enrichments to your data

9 enrichments available.

5. You'll notice how there are several enrichments available. **Click on Add enrichments**.



5 documents

0 documents failed

Created on

Last updated

Identified **8 fields** from your data

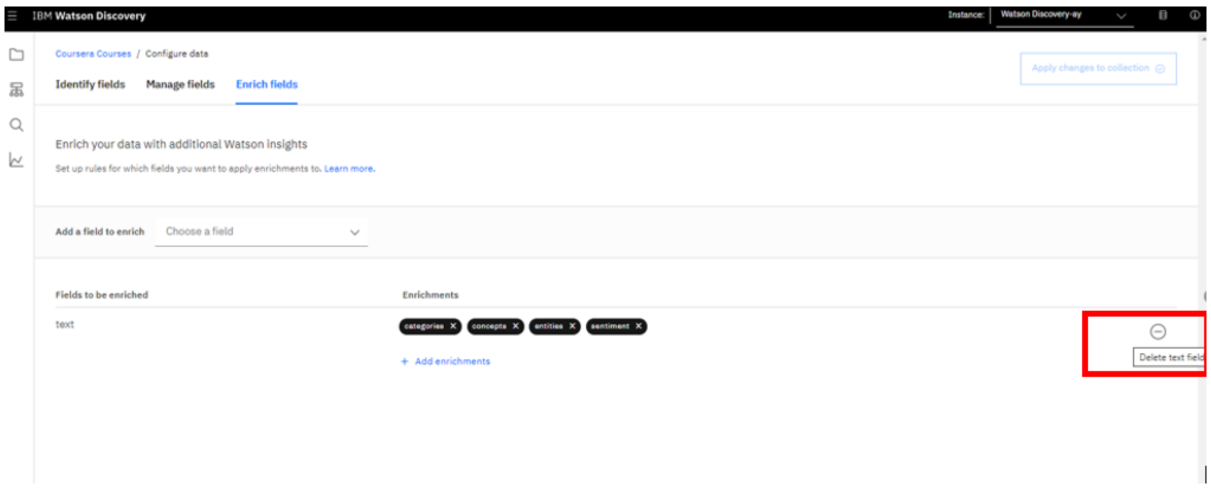
- name
- extracted_metadata
- instructorIds
- courseType
- description
- slug
- id
- partnerIds

Added **no enrichments** to your data

9 enrichments available.

6. Enrichments tell Watson what kind of analysis we want to run on the data we uploaded, and then store that information as metadata attached to our raw content.

You'll see a enrichment added to the text field already. Delete the previosly added enrichment field by clicking on the '-' sign on the right side of the enrichment field.



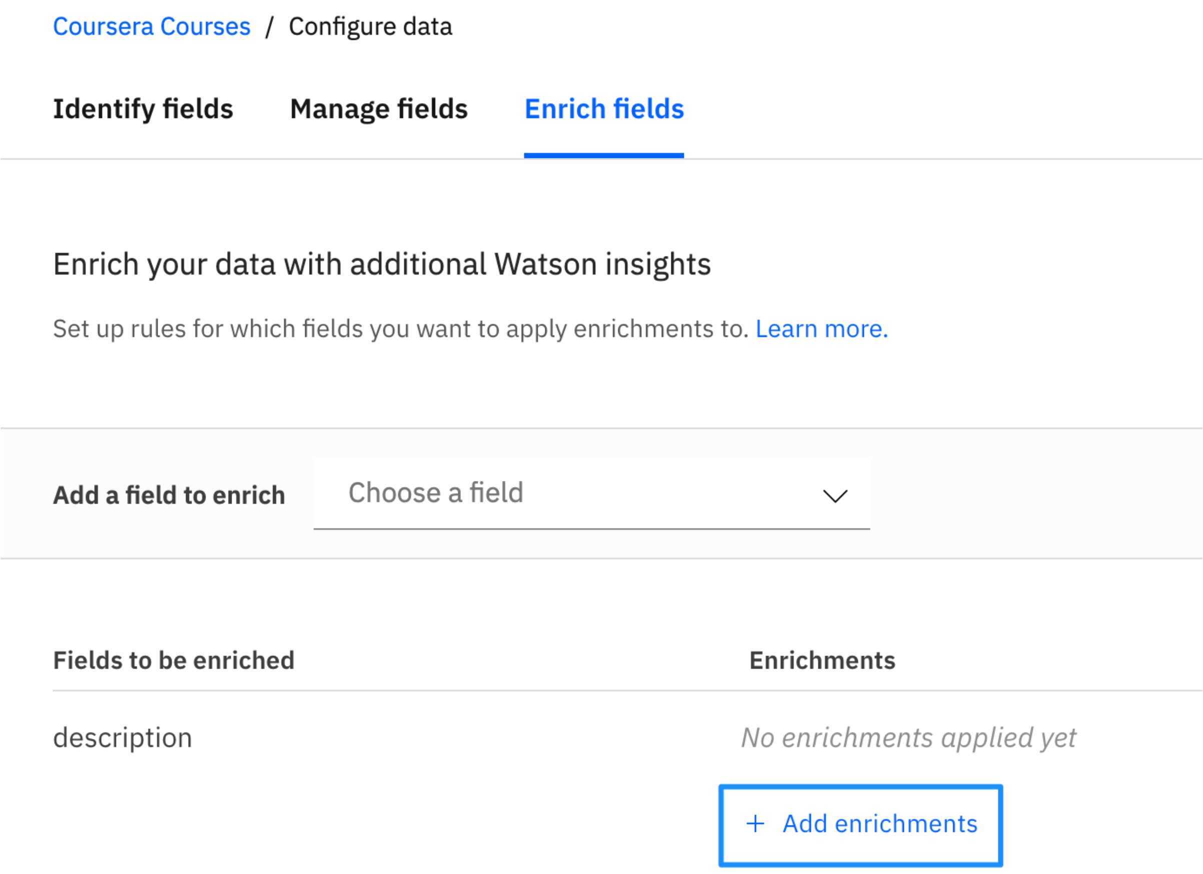
Enrich your data with additional Watson insights

Set up rules for which fields you want to apply enrichments to.

Add a field to enrich: Choose a field

Fields to be enriched	Enrichments
text	categories, concepts, entities, sentiment

Now, **Select description** as the field to enrich, and **then click on + Add enrichments** as shown in figure.



Enrich your data with additional Watson insights

Set up rules for which fields you want to apply enrichments to.

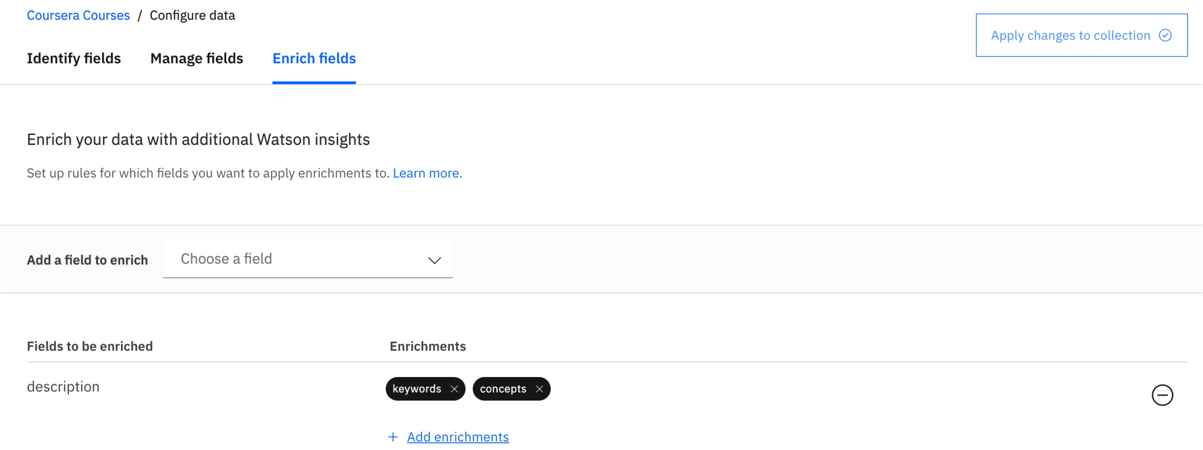
Add a field to enrich: Choose a field

Fields to be enriched	Enrichments
description	*No enrichments applied yet*

7. You'll see a series of possible enrichments, including keyword extraction, sentiment analysis, concept tagging, and more. Spend some time familiarizing yourself with what these enrichments do by clicking on the Learn more under each enrichment.

8. Once you have a general understanding of what these enrichments do, **click the *Add* button for the *Keyword Extraction* and *Concept Tagging* enrichments.**

9. Once done, close the *Add enrichments* window to return to the main screen, as shown below.



Coursera Courses / Configure data

Enrich your data with additional Watson insights

Set up rules for which fields you want to apply enrichments to. Learn more.

Add a field to enrich: Choose a field

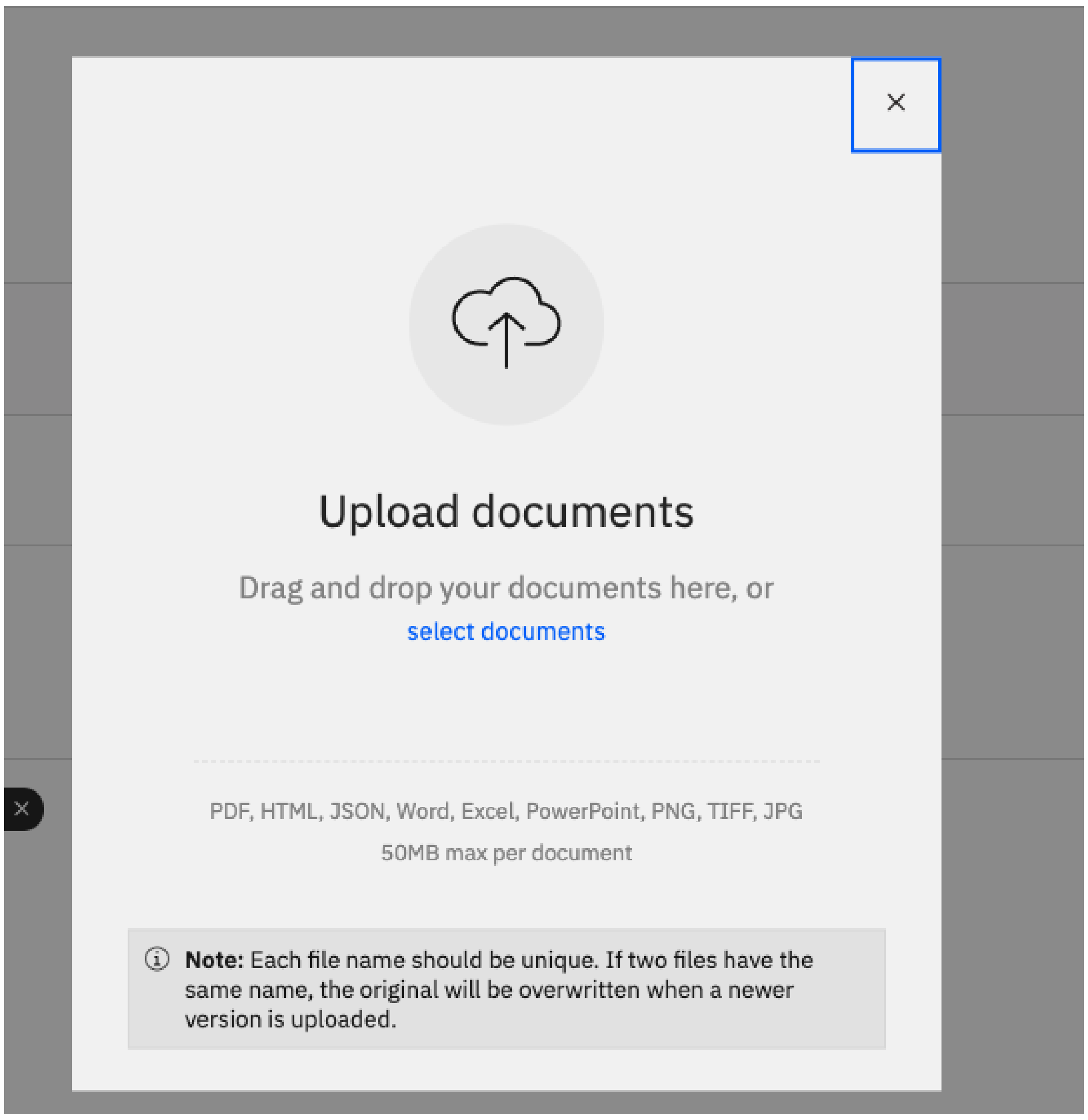
Fields to be enriched	Enrichments
description	keywords ✕ concepts ✕

By the way, aside from enriching fields, you can also identify/map fields to the data if the document you uploaded was unstructured. In our case, we used JSON which is already structured, so the *Identify fields* tab is not useful to us here.

The *Manage fields* tab allows us to turn off fields that we don't want indexed. This is more important when you have lots of irrelevant fields. In our case, we can leave all fields on by default or feel free to turn off some as an exercise, but preserve name, slug, and description.

When you're done exploring the data configuration section, **click on Apply changes to collection**.

If you get this popped up. Please upload the 5 documents again. Re-uploading the duplicate documents will simply add enrichments.



Upload documents

Drag and drop your documents here, or select documents

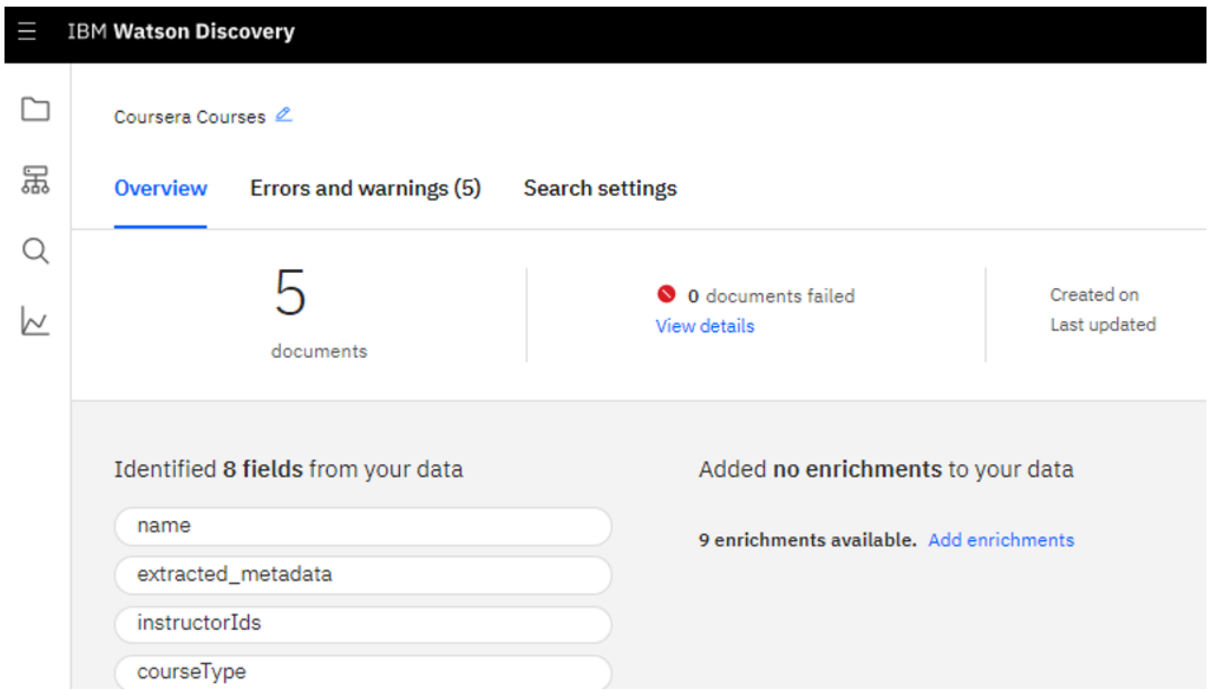
PDF, HTML, JSON, Word, Excel, PowerPoint, PNG, TIFF, JPG

50MB max per document

Note: Each file name should be unique. If two files have the same name, the original will be overwritten when a newer version is uploaded.

Please navigate to Overview tab of Courses collection to check whether enrichments are properly added to the collection.

This screenshot below is an example of missing enrichments in the data.



5 documents

0 documents failed View details

Created on

Last updated

Identified 8 fields from your data

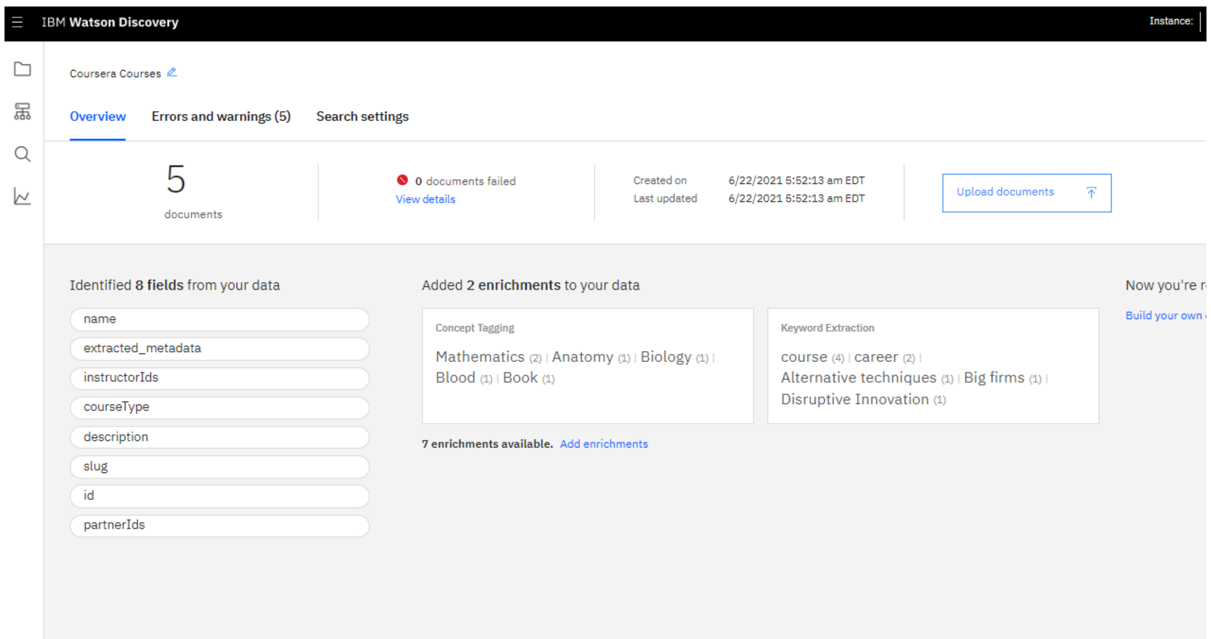
- name
- extracted_metadata
- instructorIds
- courseType

Added no enrichments to your data

9 enrichments available. Add enrichments

If enrichment is not added, please start the process of adding enrichment again.

It should show **added 2 enrichments to your data** like below screenshot if enrichments are properly added to the collection.



5 documents

0 documents failed View details

Created on 6/22/2021 5:52:13 am EDT

Last updated 6/22/2021 5:52:13 am EDT

Identified 8 fields from your data

- name
- extracted_metadata
- instructorIds
- courseType
- description
- slug
- id
- partnerIds

Added 2 enrichments to your data

Concept Tagging: Mathematics (2) | Anatomy (1) | Biology (1) | Blood (1) | Book (1)

Keyword Extraction: course (4) | career (2) | Alternative techniques (1) | Big firms (1) | Disruptive Innovation (1)

7 enrichments available. Add enrichments

Now you're r

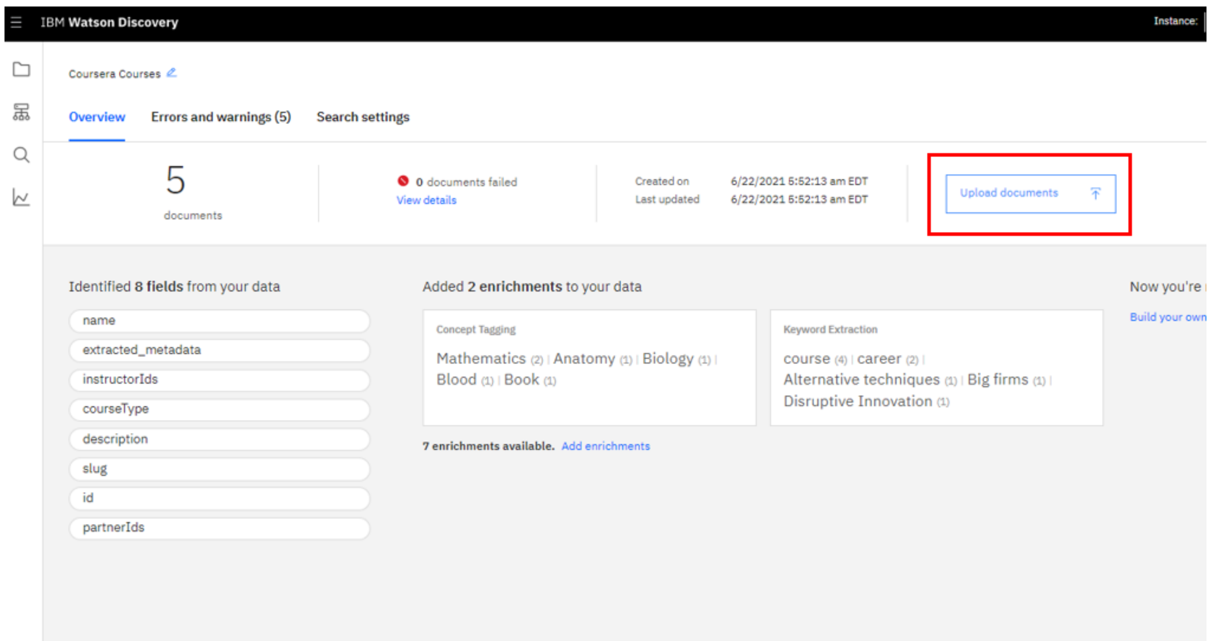
Build your own

If you have successfully uploaded 5 documents and enrichment is added to them properly then you are all set to upload the remaining 495 documents.

Uploading the remaining documents

1. You need to click on the upload button in the same collection to upload the remaining documents as you uploaded 5 documents earlier. You can download the remaining date from the below link:

<a href=https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBMDeveloperSkillsNetwork-CB0106EN-SkillsNetwork/labs/Module%202-coursera/data/495-Coursera%20Courses.zip?utm_medium=Exinfluencer&utm_source=Exinfluencer&utm_content=000026UJ&utm_term=10006555&utm_id=NA-SkillsNetwork-Channel-SkillsNetworkCoursesIBMDeveloperSkillsNetworkCB0106ENSkillsNetwork20719128-2021-01-01>495-coursera-courses.zip



5 documents

0 documents failed View details

Created on 6/22/2021 5:52:13 am EDT

Last updated 6/22/2021 5:52:13 am EDT

Identified 8 fields from your data

- name
- extracted_metadata
- instructorIds
- courseType
- description
- slug
- id
- partnerIds

Added 2 enrichments to your data

Concept Tagging: Mathematics (2) | Anatomy (1) | Biology (1) | Blood (1) | Book (1)

Keyword Extraction: course (4) | career (2) | Alternative techniques (1) | Big firms (1) | Disruptive Innovation (1)

7 enrichments available. Add enrichments

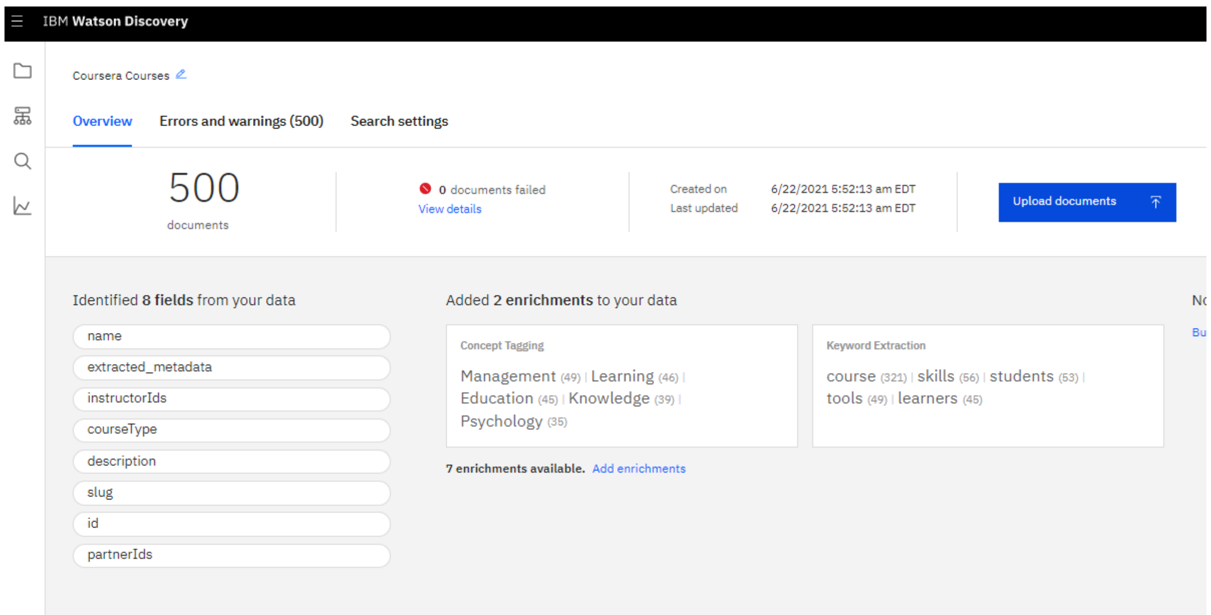
Now you're

Build your own

2. Once you click on upload you'll be prompted to upload documents. **Click on the upload icon and add the files you extracted**. You can use CMD+a or CTRL+a to select all of them at once in the upload dialog. Alternatively, you can simply drag and drop the files on the page.

> NOTE: Processing the data will take a few minutes. If you see Errors and warnings during the process, simply ignore the them. It gives a warning because the course document contains a reserved field of id.

After the documents are uploaded, you do not need to add enrichments again to these documents as they will be auto enriched and you should see something like this on your screen:



500 documents

0 documents failed

Created on 6/22/2021 5:52:13 am EDT
Last updated 6/22/2021 5:52:13 am EDT

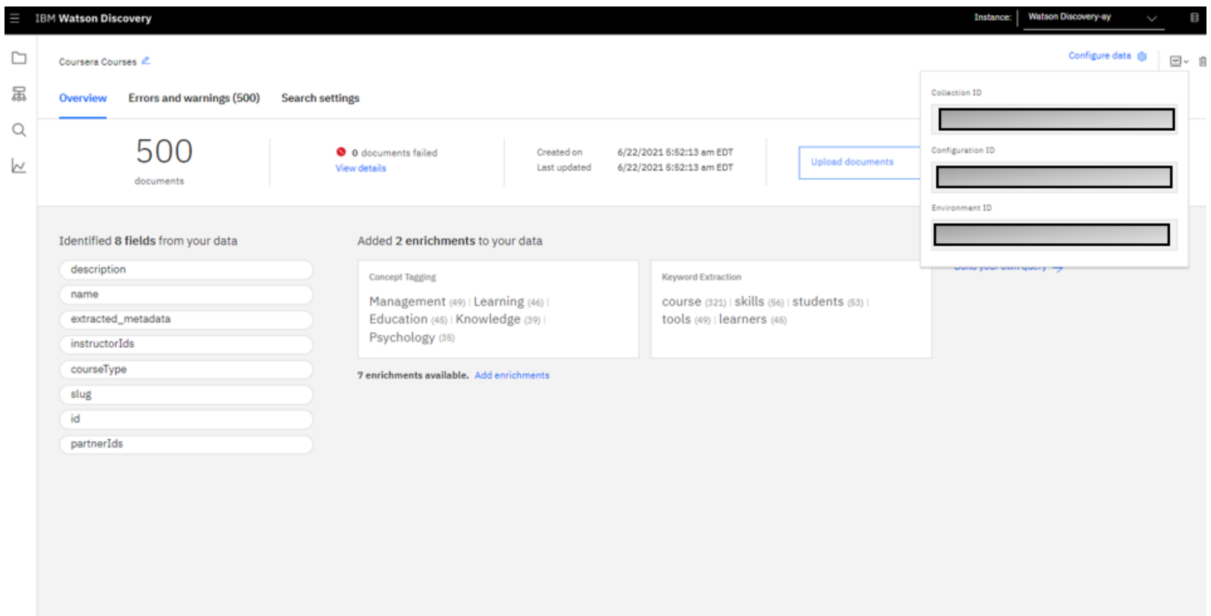
Identified **8 fields** from your data: name, extracted_metadata, instructorIds, courseType, description, slug, id, partnerIds

Added **2 enrichments** to your data

Concept Tagging: Management (49) | Learning (46) | Education (45) | Knowledge (39) | Psychology (35)

Keyword Extraction: course (321) | skills (56) | students (53) | tools (49) | learners (45)

Please make a note of its collection ID and environment ID as it will be used in the coming labs.



Collection ID [redacted]

Configuration ID [redacted]

Environment ID [redacted]

500 documents

0 documents failed

Created on 6/22/2021 5:52:13 am EDT
Last updated 6/22/2021 5:52:13 am EDT

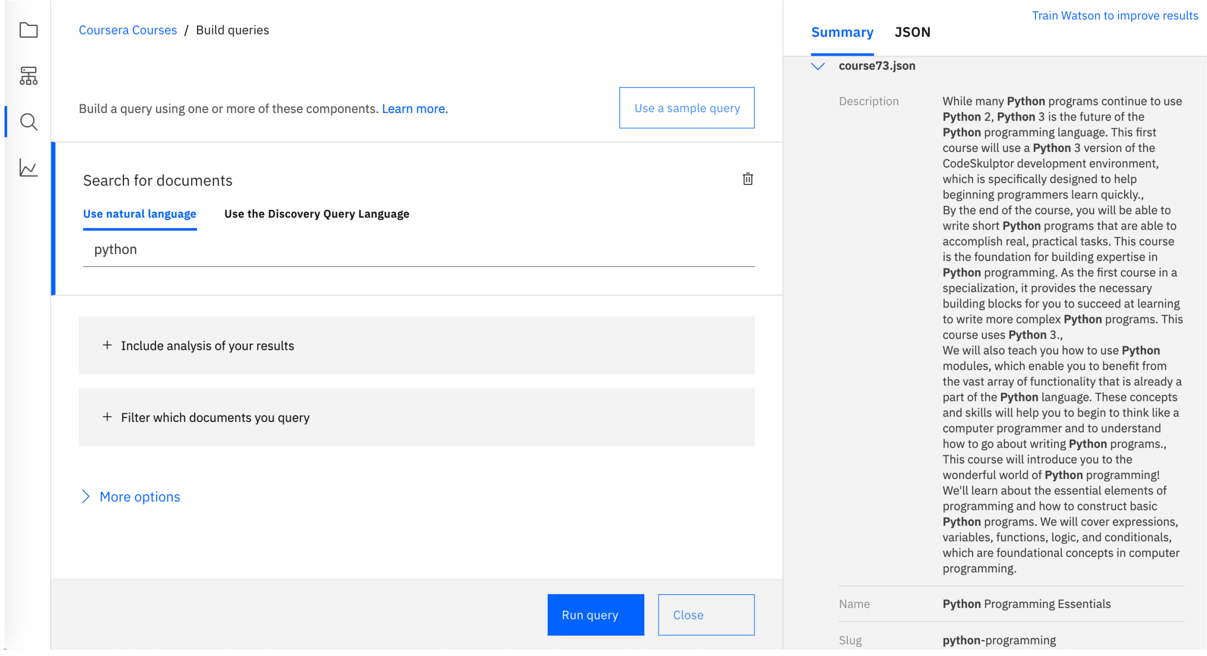
Identified **8 fields** from your data: description, name, extracted_metadata, instructorIds, courseType, slug, id, partnerIds

Added **2 enrichments** to your data

Concept Tagging: Management (49) | Learning (46) | Education (45) | Knowledge (39) | Psychology (35)

Keyword Extraction: course (321) | skills (56) | students (53) | tools (49) | learners (45)

1. At this point we are all set. As a sanity test, click on the **query icon** on the very left and then **click on *Search for documents***.
2. You'll be prompted to enter a natural language query. **Try "*python*" and click *Run* query**. The first 10 results will be displayed in the Summary panel. These are not randomly selected out of the list of matching results. These are ranked based on relevance to the query. In fact, when I ran it, all top 10 results where highly relevant Python courses.


Coursera Courses / Build queries
Use a sample query
Use natural language
Use the Discovery Query Language
+ Include analysis of your results
+ Filter which documents you query
More options
Run query
Close
Train Watson to improve results
Summary
JSON

Build a query using one or more of these components. Learn more.

Search for documents

python

course73.json

Description: While many **Python** programs continue to use **Python** 2, **Python** 3 is the future of the **Python** programming language. This first course will use a **Python** 3 version of the CodeSkulptor development environment, which is specifically designed to help beginning programmers learn quickly., By the end of the course, you will be able to write short **Python** programs that are able to accomplish real, practical tasks. This course is the foundation for building expertise in **Python** programming. As the first course in a specialization, it provides the necessary building blocks for you to succeed at learning to write more complex **Python** programs. This course uses **Python** 3., We will also teach you how to use **Python** modules, which enable you to benefit from the vast array of functionality that is already a part of the **Python** language. These concepts and skills will help you to begin to think like a computer programmer and to understand how to go about writing **Python** programs., This course will introduce you to the wonderful world of **Python** programming! We'll learn about the essential elements of programming and how to construct basic **Python** programs. We will cover expressions, variables, functions, logic, and conditionals, which are foundational concepts in computer programming.

Name: **Python** Programming Essentials

Slug: **python**-programming

This is just a simple query of course, but Watson Discovery is capable of producing highly relevant results for more elaborate queries as well. And when you need really complex, precise queries in your applications, you can always rely on programming the queries with the Discovery Query Language.

Author(s)

[Antonio Cangiano]

Changelog

Date	Version	Changed by	Change Description
2020-09-16	2.0	Shubham	Migrated Lab to Markdown and added to course repo in GitLab
2021-06-22	2.1	Anamika	Updated Instructions