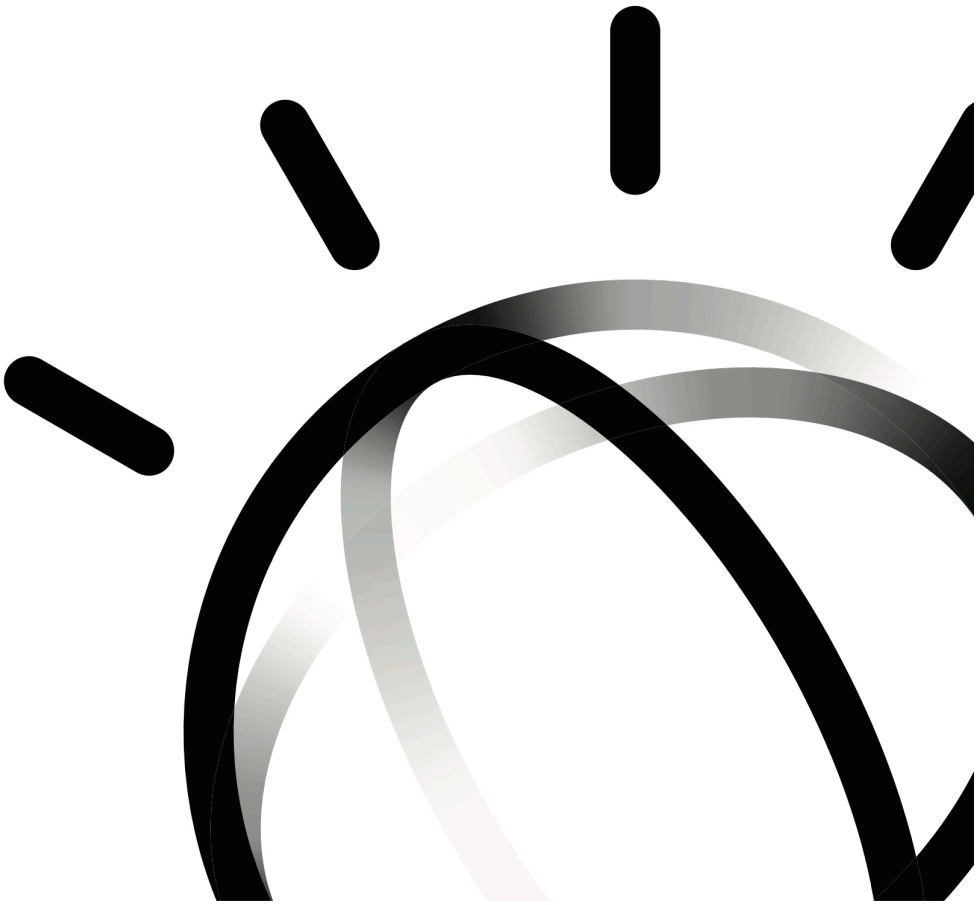
IBM Watson Solutions

Business and Academic Partners



Ingest, Convert, Enrich and Query with Watson Discovery Service

Prepared by Armen Pischdotchian

Updated by Hélène Fourot Quillaud

Version 3.1 August 2018

Overview

With the IBM Watson™ Discovery service you build cognitive, cloud-based exploration applications that unlock actionable insights hidden in unstructured data - including your own proprietary data, as well as public and third party data. Creating your first discovery journey using the IBM Watson™ Discovery service entails the following steps:

- Convert, enrich and normalize data.
- Securely explore your proprietary content as well as free and licensed public content.
- Apply additional enrichments such as concepts, relations, and sentiment through natural language processing.
- Query and analyze your results.
- Simplify development while providing direct access to APIs.

IBM Watson Discovery service architecture is depicted below.

You can upload content and begin finding insights with the Discovery service by using either the Discovery Tooling or the Discovery API. This document shows you how to use both.

You might want to watch this and related videos from the playlist:

https://www.youtube.com/watch?v=fmIPeopG-ys&t=1s

This workshop will start by guiding you on how to upload and query documents. Then, it will show you how to configure your sample document with conversions and enrichments, so that you can leverage custom configurations for better insights

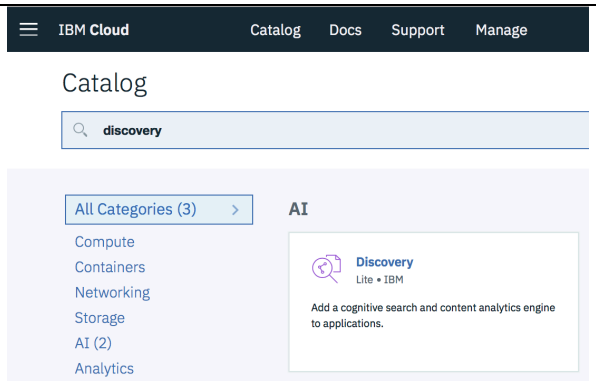
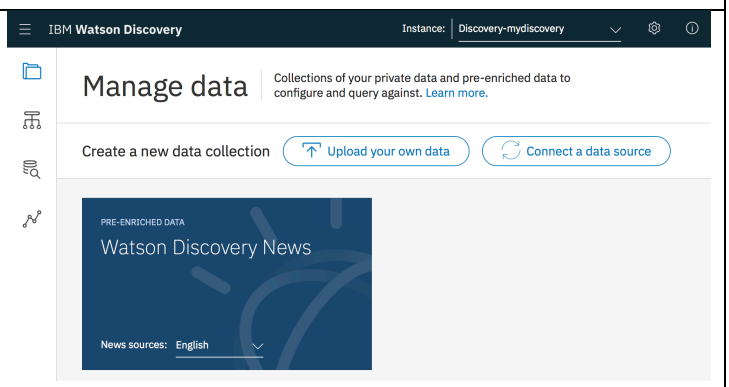
At the end of this workshop, you may want to spend some time and consider coupling the Discovery service with the Watson Assistant service and think about how you could use the Watson Knowledge Studio (a SaaS offering, not on IBM Cloud) to further edit the annotators used with some of the cognitive language microservices used within the Discovery service.

Prerequisite

Prior to starting this lab, you must obtain IBM Cloud credentials.

Working with the Discovery Tooling

Uploading and querying documents

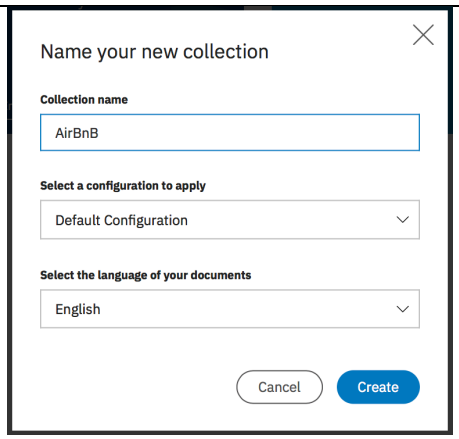
Steps	Example screen capture
1. Login into IBM Cloud at https://console.ng.bluemix.net 2. Click the **Catalog** tab. 3. Search for the **Discovery** service and click that tile. 4. Edit the Service name to something meaningful to you (for example: Discovery-**mydiscovery**) and click **Create** (If you have just created your account and accessed it from the confirmation email, you may need to log into IBM Cloud once again, then you can see the Create button in the bottom right corner). 5. Click the **Launch Tool** button.	
Watson Discovery News, a public data set that has been pre-enriched with cognitive insights, is included with Discovery. You can also use this public, unstructured data set to query for insights that you can integrate into your applications. Watson News is a dataset of primarily English language news sources that is updated continuously, with approximately 300,000 new articles and blogs added daily. In order to experiment with the robust querying capabilities of the Discovery service, you will create a new Collection and upload the AirBnB reviews that are already in JSON format as final documents to your collection. 6. Click the folders icon in the top right corner of the page. 7. Create a new collection by clicking **Upload your own data**.	

8. Name your collection; for example: **AirBnB**.
9. Click **Create**.

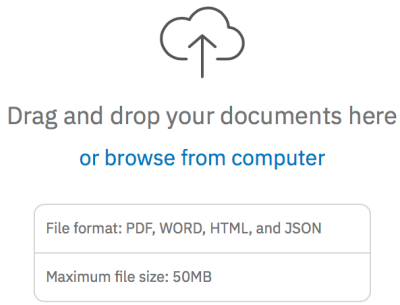
   Each collection you create is a logical division of your data in the environment. Each collection will be queried independently when you get to the point of delivering results.

   Why would I want more than one collection? There are a few reasons, including:

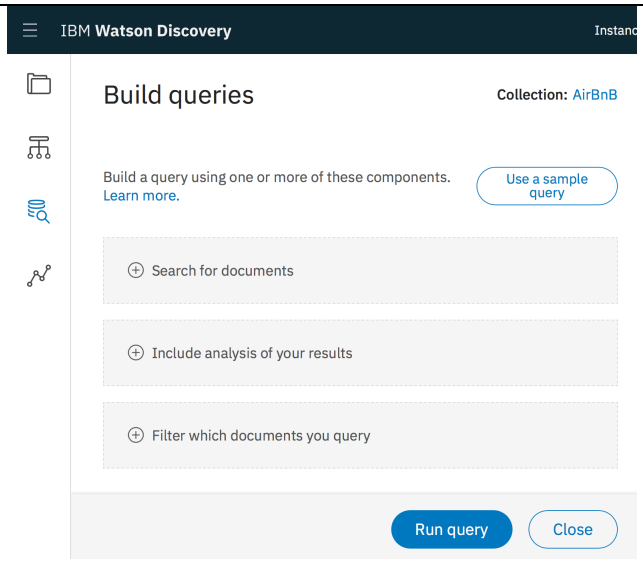
- You may want multiple collections in order to separate results for different audiences
- The data may be so different that it doesn't make sense for it all to be queried at once



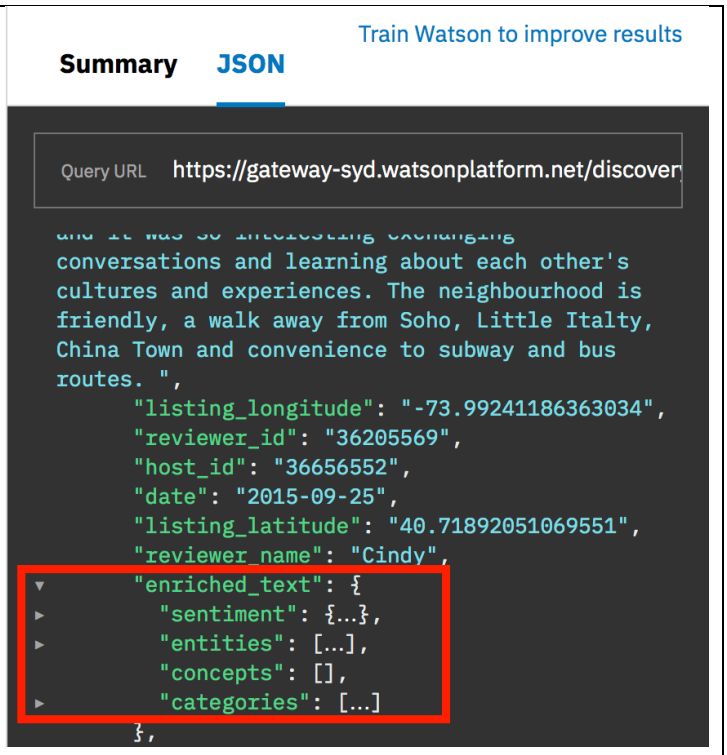
10. Extract the contents of the rental_reviews_easy_ingest_handpicked.zip file that you can download from Github at https://github.com/apischdo/Bluemix-workshop-assets/blob/master/Discovery.zip.
11. Select all JSON files and upload them to the Discovery Service. This may take a few minutes.



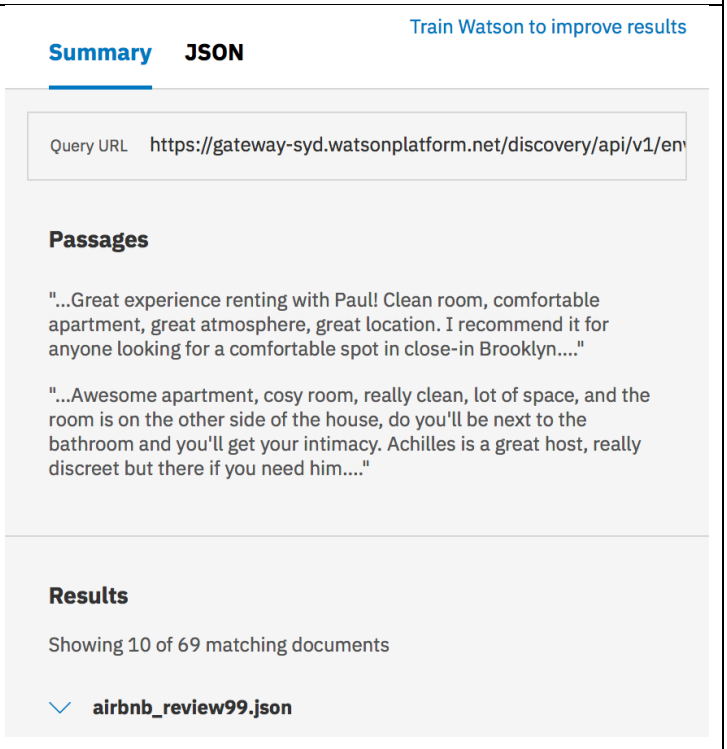
12. Now that you have uploaded documents in the AirBnB collection, you are ready to query the data. Click the magnifying glass icon on the left.
13. Without running any specific queries, just click **Run query**.
14. Take a moment and view the results. There are two ways to look at results, a **Summary** view and a **JSON** view. First, take a look at the Summary view.



15. Select **JSON** and look at the **enriched_text** section below the first passage.
16. Notice the hierarchy: sentiment, entities, concepts. You will note that these are the enrichments that are applied by Watson Discovery to the ingested documents. In essence, they are information extracted from the unstructured text and now available as structured meta-data that you can query.
17. You are now ready to run some queries.



18. Let's build a basic query using natural language. In the *Search for Documents*, type: **room with a great view.** On the right-hand panel, notice that results are presented in two sections.
    a. **Passages** are relevant paragraphs extracted from documents
    b. **Results** are enriched documents
19. In the **More options** section, you may set **Include relevant passages** to **No** and run your query again. You now only see the **Results** section.
20. You may click **Learn more** and take a moment and read the guidelines for querying. Basic queries can also be run using query language.
21. Clear all fields, and click **Run query** again.



22. Now let's build an aggregation query using query language. Open the *Include analysis of your results* section, select **Edit in query language** and then click the question mark next to the input field.

23. Try out some aggregation formats replacing the example text to key words relevant to the AirBnB documents. Ensure that the hierarchy follows your classification scheme, not necessarily the example.

    `term(enriched_text.categories.label,count:5)`

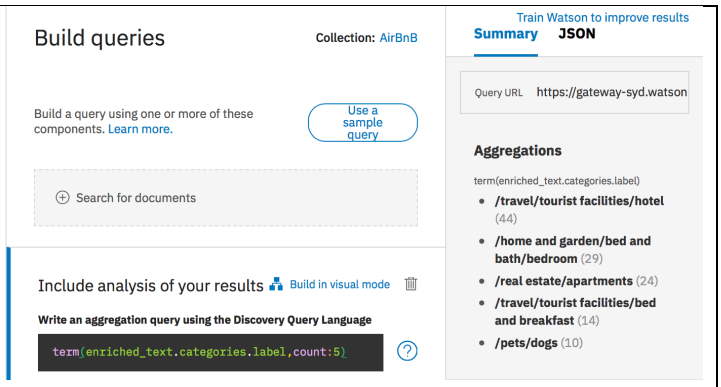
    Returns the five most frequent categories in the data set and the number of documents for each category

    `term(host_name,count:3)`

    Returns the three most frequent host names for this collection.

24. Click the **JSON** tab for more details.

25. Clear all fields, and click **Run query** again.



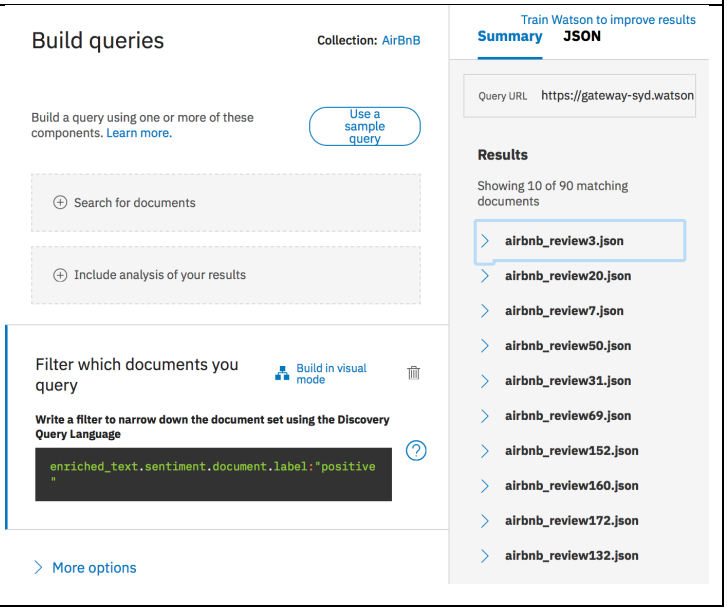
26. And finally, let's build a filter query. Repeat the above action with the *Filter which documents you query* section by clicking the question mark and following the example format.

    `enriched_text.sentiment.document.label:"positive"`

    `enriched_text.sentiment.document.score>0`

    Results include documents with a positive sentiment.

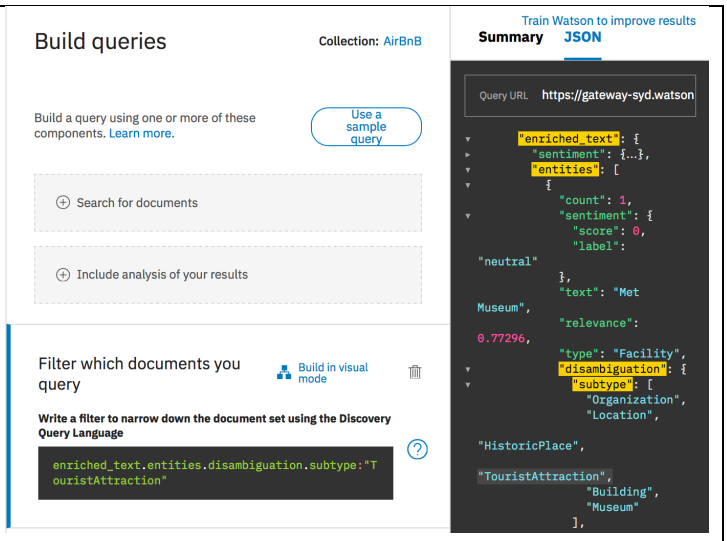
    Filter queries run faster than basic queries because they are cached, but results are not sorted.



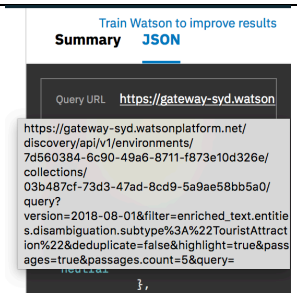
27. How about finding out apartments that are near tourist attractions? Clear all fields and run the following query in the *Filter which documents you query*…it's one line.

    `enriched_text.entities.disambiguation.subtype:"TouristAttraction"`

    The result set contains documents that have at least one entity with « TouristAttraction » in its list of disambiguation subtypes.



28. Notice the **Query URL** built based on your query. This URL will be used when you build an application that needs to query the document collection using Discovery. You may copy the URL into a temporary location, it will be used in the last section of this lab, **Working with the Discovery API**.

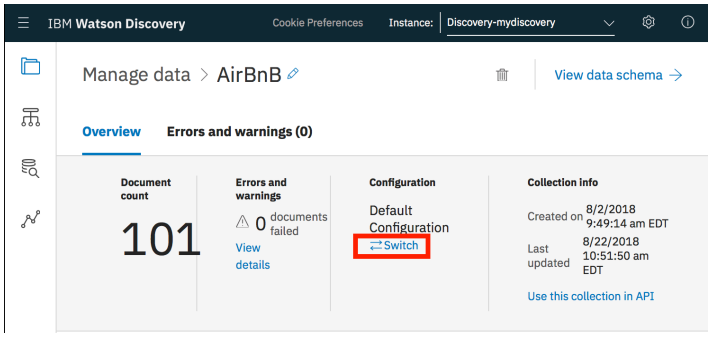


Building a custom configuration

You have been adding content to your collection just after it was created. When a collection is created, a set of default configurations are automatically provided. If you are happy with these defaults, you can proceed to uploading your content. This is what we have done so far. If you need to configure the service to process the content the way that you want, then the best practice is to define a **custom configuration** before uploading your content.

29. From the "Manage Data" page, in the Configuration section, Click **Switch**.
30. Click **Create a new configuration**.
31. Name your configuration, for example **Config01**. In order to set up this configuration, you will:

- Identify some sample content (documents that are representative of your files)
- Upload the sample content
- Adjust the conversion
- Define enrichments
- Normalize the results



To make the configuration process more efficient, you can upload of Microsoft Word, HTML, JSON, or PDF files that are representative of your document set. These are called sample documents. Sample documents are not added to your collection, they are only used to identify fields that are common to your documents and customize those fields to your requirements.

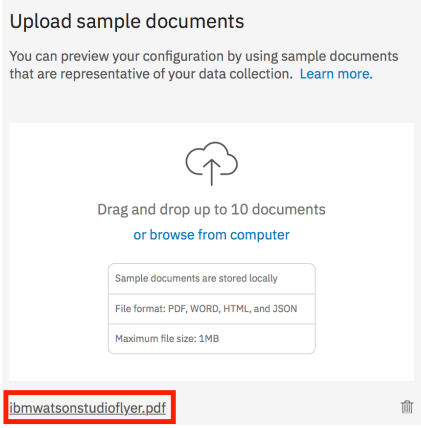
When creating a new configuration file in the Discovery tooling, you can upload sample documents via drag and drop or browse. Click on the file name in the Upload Sample Documents pane to preview each file.

Remember the following items when uploading sample documents:

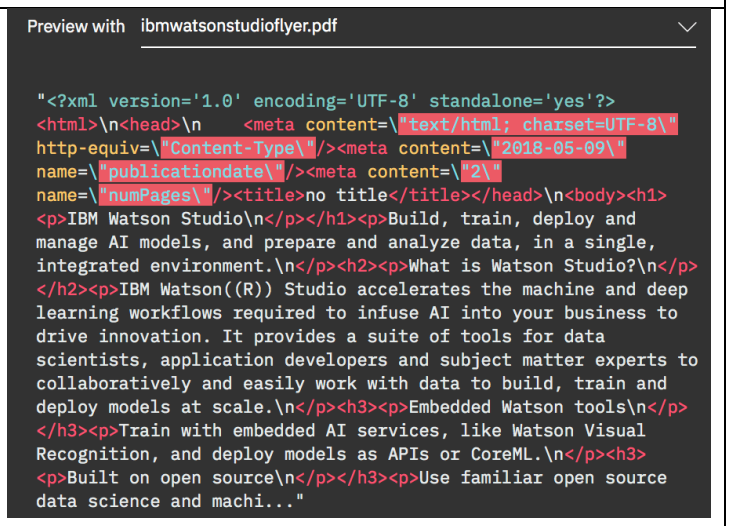
- All of your documents are converted to JSON before they are enriched and indexed.
- Microsoft Word and PDF documents are converted to HTML first, then JSON.
- HTML documents are converted directly to JSON.

Note: Sample documents are automatically deleted after 1 month, but you can upload the same documents again if you would like to make additional changes to your configuration.

32. For this exercise, upload a PDF sample document from your local drive. Here we are using the following file: https://public.dhe.ibm.com/common/ssi/ecm/41/en/41015641usen/ibmwatsonstudioflyer.pdf
33. Once the document is uploaded it will appear in the right pane. Click the document name.



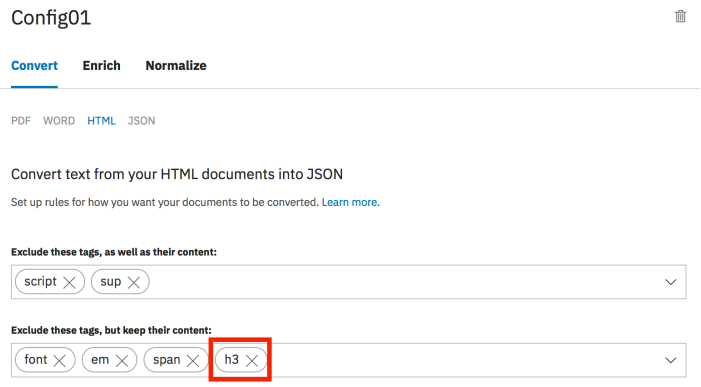
34. Click the **Convert** link.
35. Select **PDF**.
36. For example, this particular document uses font size **24** and **bold** for **h1,** font size **14** and **bold** for **h2** and font size **12** and **bold** for **h3**. On the left panel, you can specify html text conversion features that are relevant to the document.
37. Click **Apply and Save**. Notice the updates in the right panel.
38. Click the **Learn more** link and take a moment to read the details behind text conversion.



39. Click the **HTML** link.

40. For the purposes of this exercise, type **h3** in the *Exclude these tags, but keep their content* box. Click **Apply & Save** and look at the right panel to see that your <h3> tags are now gone.

Important to note, that the conversions you specify for the html conversion will apply to both your PDF and MS Word uploads. If you want your PDF documents to have a different configuration from your MS Word documents, then you have to place them in separate collections with separate custom configurations.



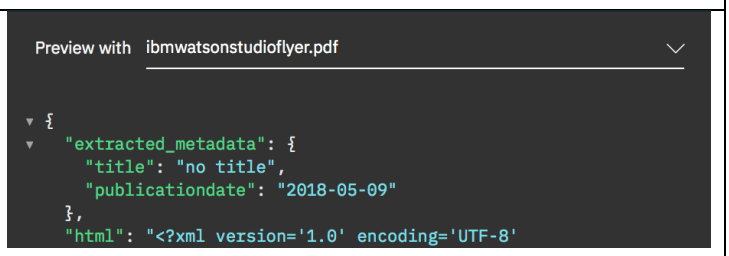
41. Click the **JSON** link.

    You will now have an opportunity to do some housekeeping and clean the output of your documents before they are indexed.

42. For the purpose of this exercise, select **Add field**, then **remove** and *extracted_metadata.publicationdate* (or any other field applicable to your document). Turn on the **Remove empty fields** option

43. Click **Apply and save**.

44. Look at the right panel and notice that the removed field is now gone.



45. Now that we have configured the Conversion, let's move on to configuring the enrichment step. Click the **Enrich** link.

46. Take a moment and observe the enrichments that are applied to the document. In the right panel, "enriched_text".

    Let's assume that you do not need the **sentiment** enrichment based on this sample document.

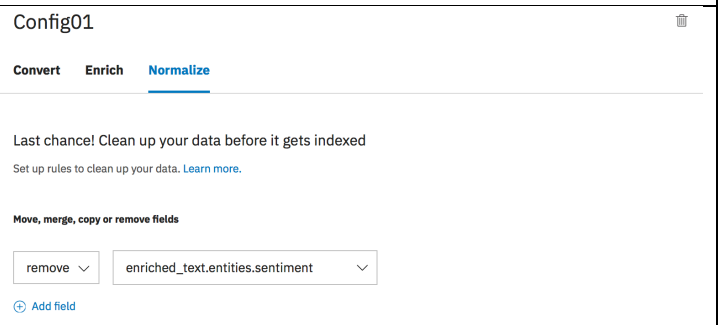
47. Remove the **sentiment** enrichment, add the **Keywords** enrichment and click **Apply and Save**. Check changes in the enriched sample documents in the right panel.

48. Click **Normalize**. This is where you get an opportunity to clean and normalize the JSON output.

49. In this example **remove** the field **enriched_text.entities.sentiment**.

50. Click **Apply and Save**.

51. Congratulations! Now that you have created your custom configuration based on sample data, you are ready to upload and query your actual (not sample) data. The query does not run on sample documents. The sample documents reside in a temporary repository and are not indexed for query.



Working with the Discovery API

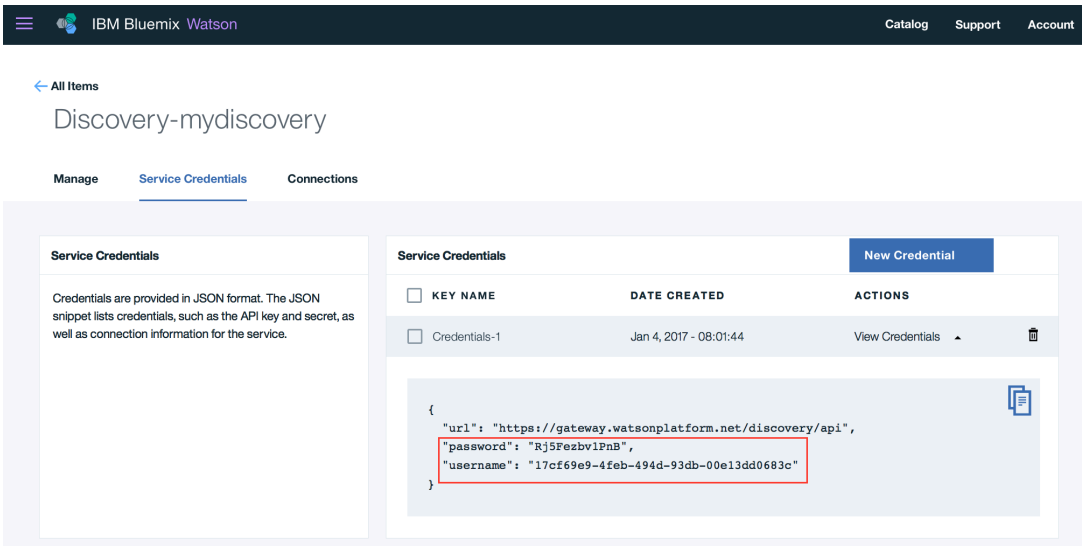
So you had some fun with the tooling. But to do bulk uploads, crawling (static) and including custom models and annotations from Watson Knowledge Studio (not in the scope of this document), you would use the API approach. Let's begin by installing Postman, which is a recommended tool for developers who work with APIs.

1) Direct your browser to this link: https://www.getpostman.com/

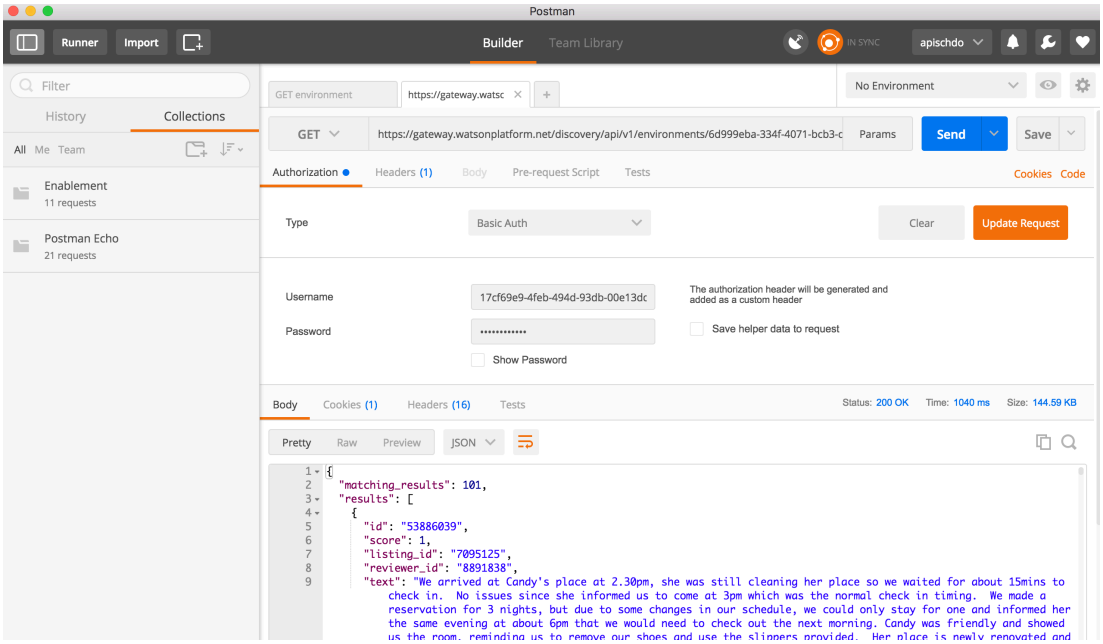
2) Download the free Postman app appropriate for your operating system, and install it.

With that installation out of the way, you are now ready to begin using the tool.

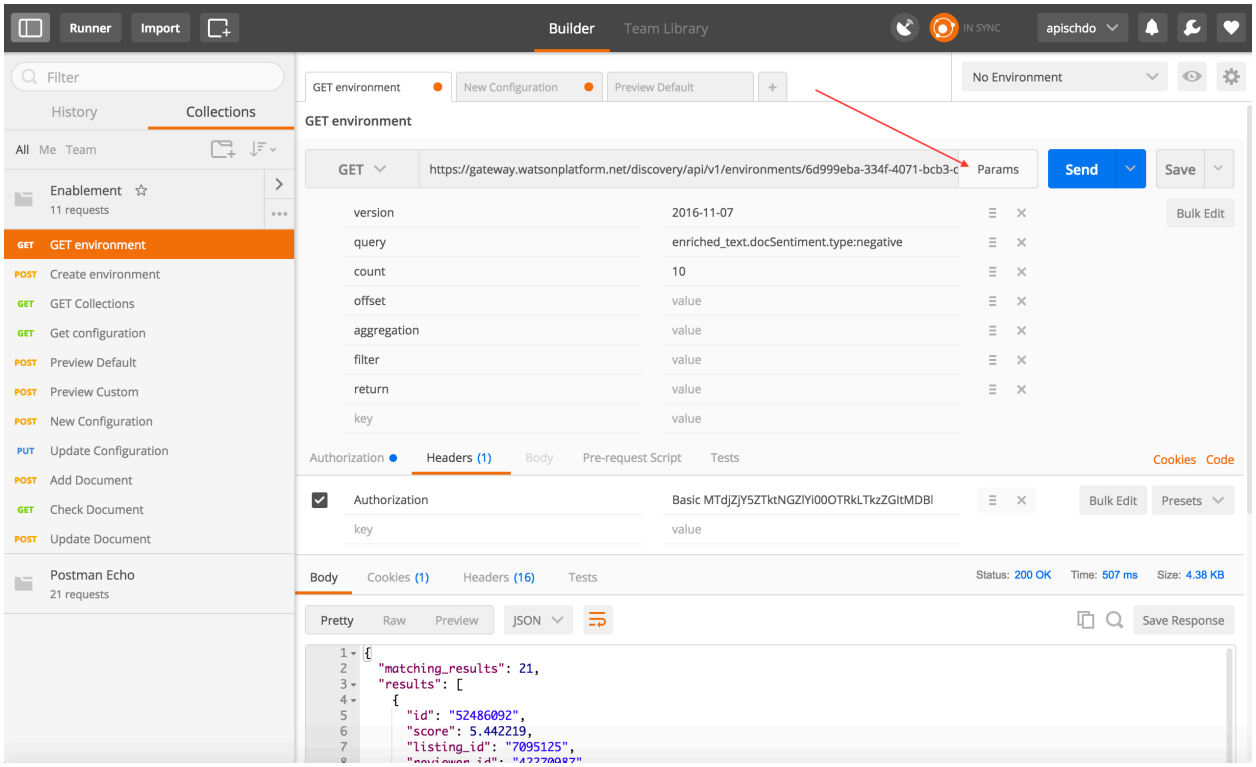
1. Go to IBM Cloud and access the Service Credentials panel of the Discovery service that you created in the beginning of this document.
2. Click the Service credentials link and note the username and password. You will be entering these in Postman along with the URL that copied earlier.



1. Open the Postman application: paste the URL as a Get method; select **Basic Auth** and enter the service credentials.
2. Click **Send**.



3. Click **Params**, enter a count of 10 as default and run some of the queries that you performed earlier. When you build your own application, you would not be using the Tooling we saw earlier, but APIs to integrate your query results within the frames of your application.



Use the documentation frequently by clicking Learn more from the tooling interface and explore the contents of the doc from the left panel. Enjoy your discovery journey.