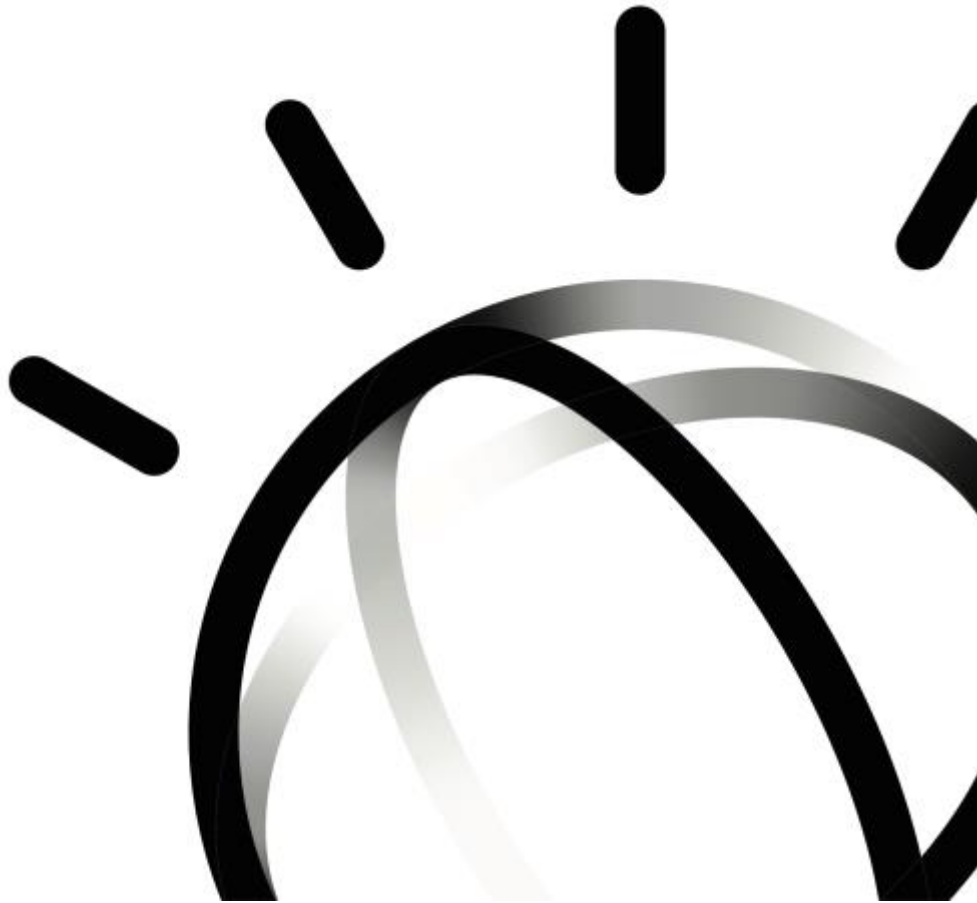


IBM Watson Solutions
Business and Academic Partners



**Ingest, Convert, Enrich and Query with
Watson Discovery Service**

Prepared by Armen Pischdotchian

Version 2.0 August 2017

Overview

What is Bluemix you ask? [Bluemix](#) 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>

Additionally, for our academic partners, there are no-charge 12-month licenses for faculty and no-charge 6-month licenses for students - all renewable and NO CREDIT CARD required!

To get started, you will need to become an Academic Initiative member. Refer to this document for details: http://it.husc.edu.vn/Media/TaiLieu/IBM_Academic_Initiative_for_Cloud_Process.pdf

The purpose of this guide is not to introduce you to Bluemix, that foundational knowledge is a prerequisite study on your part and you can obtain it from the links mentioned above. This guide is more of an instructional approach to working with the IBM Watson™ Discovery service where 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:

1. Crawl, convert, enrich and normalize data.
2. Securely explore your proprietary content as well as free and licensed public content.
3. Apply additional enrichments such as concepts, relations, and sentiment through natural language processing.
4. Query and analyze your results.
5. Simplify development while still 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.

It is strongly recommended that you 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 configure your sample document with conversions and enrichments, you then upload your own documents and begin querying the insights that the robust Natural Language Understand (NLU) stack provides.

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

Prerequisite

For this workshop all you need is a Bluemix account; you will not be building an app, but working within the conversation service in Bluemix to build a detailed dialog tree.

- **Obtain Bluemix credentials:**

1. Direct your browser to the Bluemix home page: <https://console.ng.bluemix.net/home/>
2. Click **Sign Up** on the top right. If you are affiliated with a university, use your edu email.
3. Enter requested information: for **Org**, select the suggestions that are provided for you, for example, your email address; for **Space** name, you can also select the provided suggestions, such as dev; alternatively, you can specify your own values and click **Create Account**.
4. Look for the email confirmation. You will have to login once again into Bluemix after clicking the link from the email.

- **Download the Discovery_v2.pdf document from Github**

Complete the following steps:

1. Direct your browser to GitHub (no need to sign up): <https://github.com/>
2. Search for **bluemix-workshop**.
3. Scroll down and select: *apischdo/Bluemix-workshop-assets*
4. Download just the **Discovery_v2.pdf** document.
5. Follow the instructions in that document.

- **Download the Postman application**

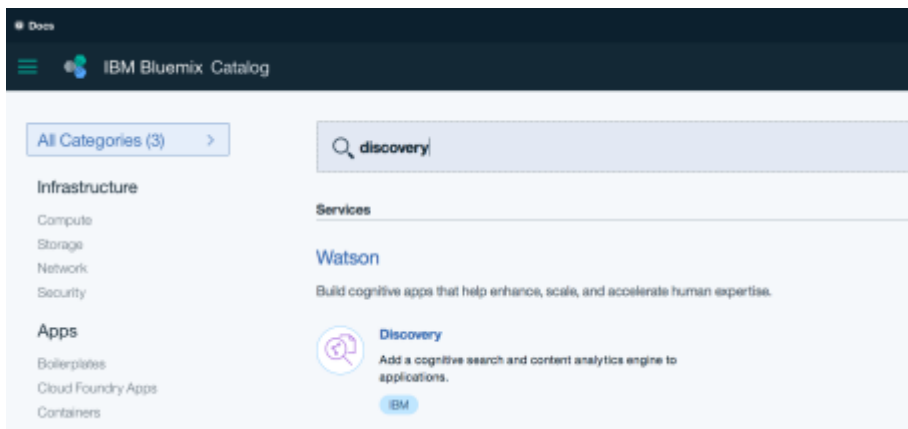
Postman is a recommended tool for developers who work with APIs.

- 1) Direct your browser to this link: <https://www.getpostman.com/>
- 2) Download the free Postman app appropriate for your operating system.

Working with the Discovery Tooling

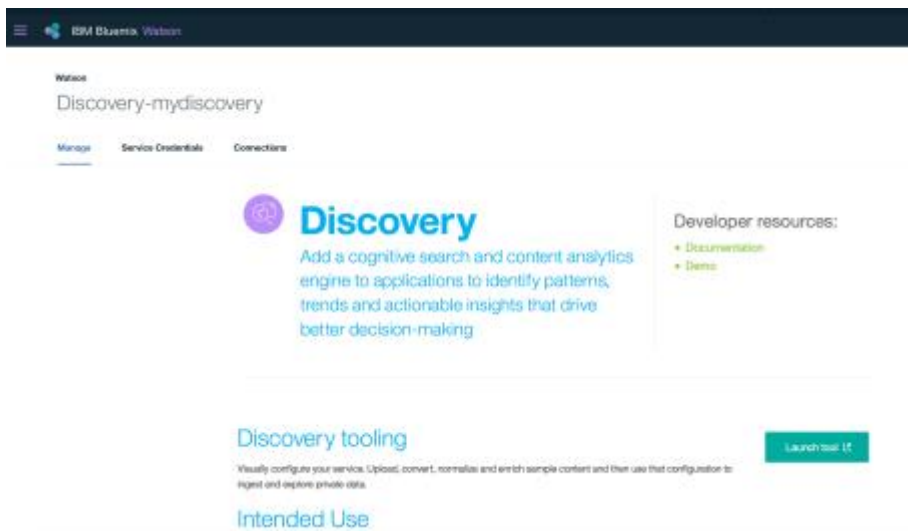
Let's begin our journey:

1. Login into Bluemix: <https://console.ng.bluemix.net>
2. Click the **Catalog** tab.
3. Search for the **Discovery** service and click that tile.



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 Bluemix once again, then you can see the Create button in the bottom right corner).

4. Click the **Launch Tool** button.



5. Click **Create a data collection** to create a new collection space

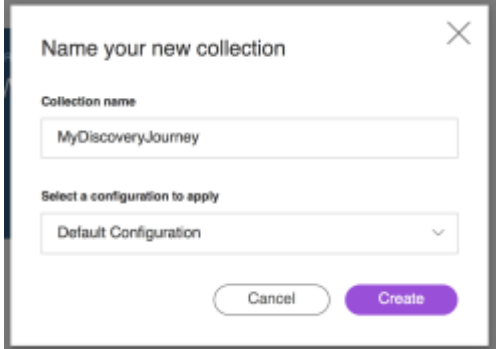
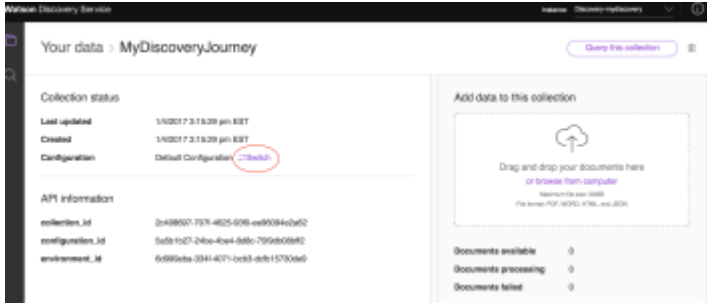
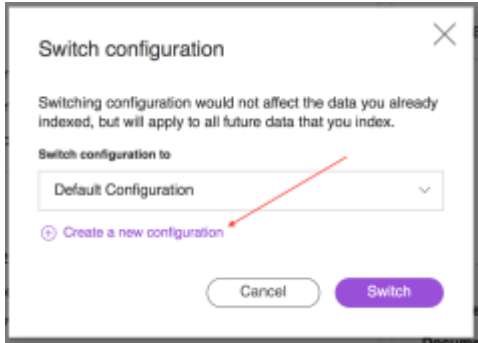
Watson News, a public data set that has been pre-enriched with cognitive insights, is also included with Discovery. You can also use this public, unstructured data set to query for insights that you can integrate into your applications. Watson News is a dataset of primarily English language news sources that is updated continuously, with approximately 300,000 new articles and blogs added daily. See a demo of what you can build with Watson News [here](#).

6. You must now setup your storage environment, click **Continue**.

7. Allow enough time for the environment to create and click **Continue** once you are all set up. The free 30 day trial gets:

- 1GB RAM, 2GB storage
- Unlimited enrichments
- 1000 news queries
- Custom domain model

Complete the following steps:

Steps	Example screen capture
<p>8. Name your collection, in this example: MyDiscoveryJourney.</p> <p>9. Click Create.</p> <p>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.</p> <p><i>Why would I want more than one collection?</i> There are a few reasons, including:</p> <ul style="list-style-type: none"> • 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 	
<p>After your collection has been created, you can immediately start uploading content using the upload area at the right of the screen. However, before you add your own content to the Discovery service, best practice is to configure the service to process the content the way that you want.</p> <p>10. Click Switch.</p> <p>11. Click Create a new configuration.</p> <p>12. Name your configuration, for example Config01.</p> <p>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; however, you will most likely want to specify one or more custom configurations. If this is the case, you will need to complete the following tasks before uploading your actual documents:</p> <ul style="list-style-type: none"> • Identify some sample content (documents that are representative of your files) • Upload the content (Uploading Sample Documents) • Adjust the conversion • Define enrichments • Normalize the results 	 

To make the configuration process more efficient, you can upload up to ten 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.

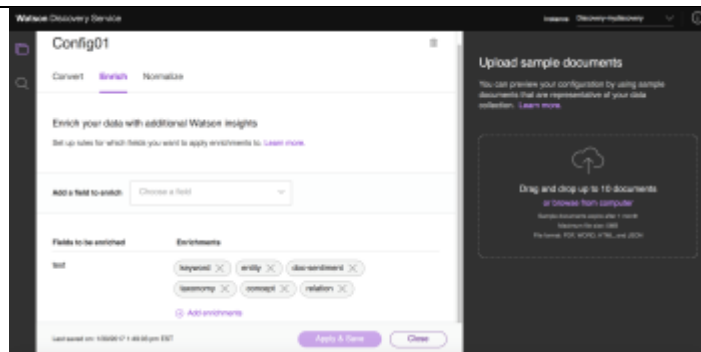
When creating a new configuration file in the Discovery tooling, you can upload sample documents via drag and drop or browse. Click on the file name in the Upload Sample Documents pane to preview each file.

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: The maximum file size for a sample document is 5MB. 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.

- For this exercise, upload a sample document from your local drive (5MB maximum). Later in this doc you will upload AirBnB documents that pertain to customer reviews on Manhattan apartments.
- Once the document is uploaded it will appear in the right pane. Click the document name.

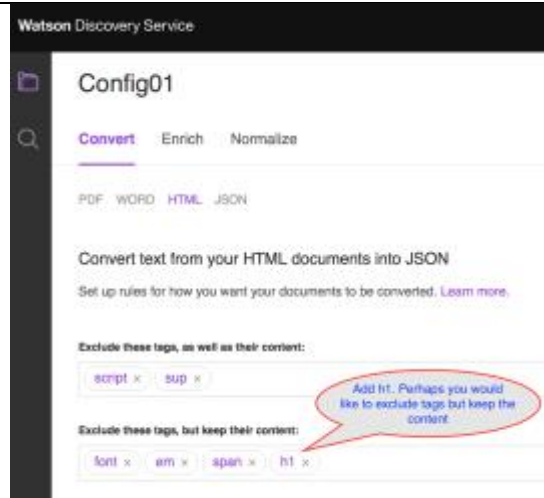


- Click the **Convert** link.
- Select **PDF**.
- For example, this particular PDF uses font size range of **16 to 16** for H1; it is **bold** and it depicts **Times New Roman**. Specify html text conversion features at this time that is relevant to your document.
- Click the **Learn more** link and take a moment to read the details behind text conversion.
- Click **Apply and save**.

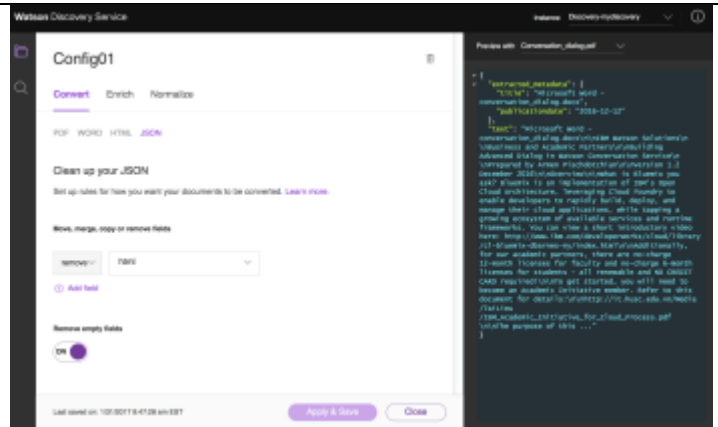


- 20. Click the **HTML** link.
- 21. For the purposes of this exercise, type **h1** in the *Exclude these tags, but keep their content* box.

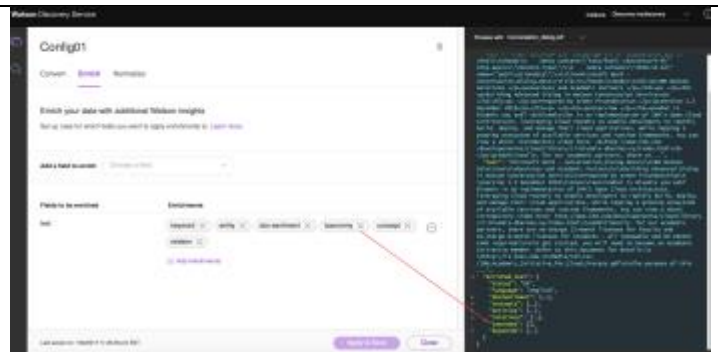
Important to note, that the conversions you specify for the html conversion will apply to both your PDF and MS Word uploads. If you want your PDF documents to have a different configuration from your MS Word documents, then you have to place them in separate containers.



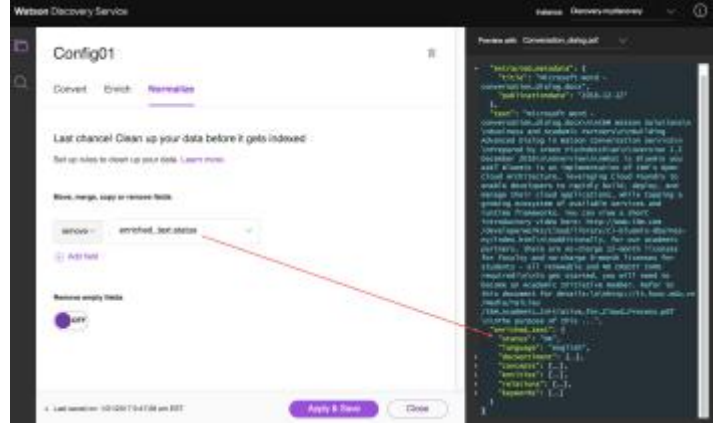
- 22. Click the **JSON** link.
- You will now have an opportunity to do some housekeeping and clean the output of your documents before they are indexed.
- 23. Type or select **html** from the drop-down list next to the remove option.
- 24. Turn on the **Remove empty fields** option
- 25. Click **Apply and save**.
- 26. Allow a few minutes for the output to appear and notice that the html tags and its innards are removed.



- 27. Click the **Enrich** link.
- 28. Take a moment and observe the enrichments that are applied to the document. In the right panel.
- Notice that you may not need the **Taxonomy** enrichment based on this sample document.
- 29. Remove the **Taxonomy** enrichment and click **Apply and Save**.



30. Click **Normalize**. This is where you get an opportunity to clean and normalize the JSON output.
31. In this example **remove the enriched_text.status** of OK.
32. Click **Save and apply**.
33. Now that you have configured your sample data, you are ready to start querying you uploaded 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.

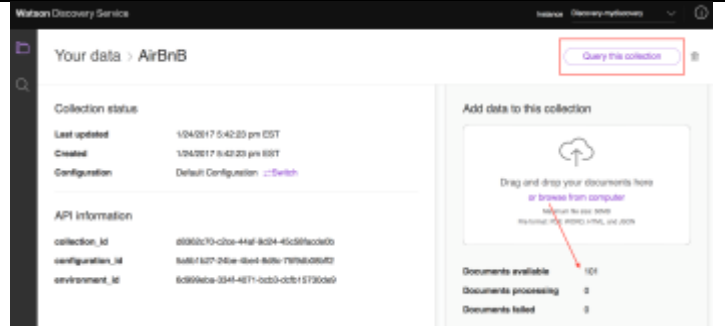


In order to experiment with the robust querying capabilities of the Discovery service, it would help to have uploaded more than one document; that way, you can use the Aggregation feature of the query builder to gain insights from numerous separate documents. To accomplish this goal, in the following steps, you will create a new Collection and upload the AirBnB reviews that are already in JSON format as final documents to your collection.

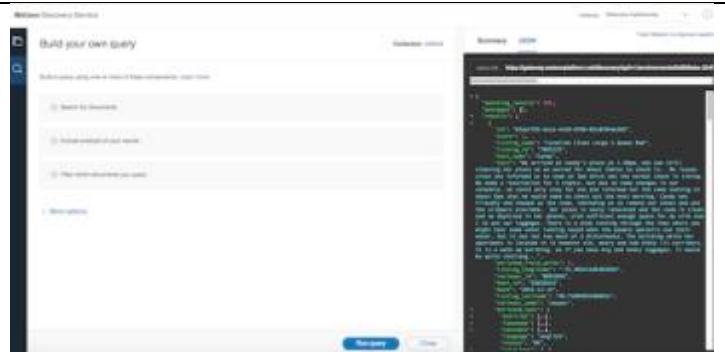
34. Click the folders icon in the top right corner of the page.
35. Create a new collection and name it; in this example, it is AirBnB

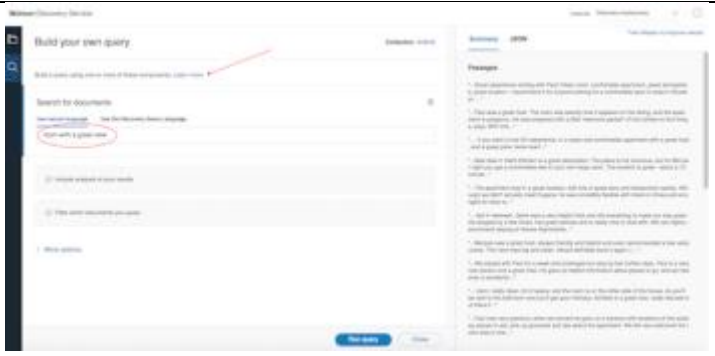
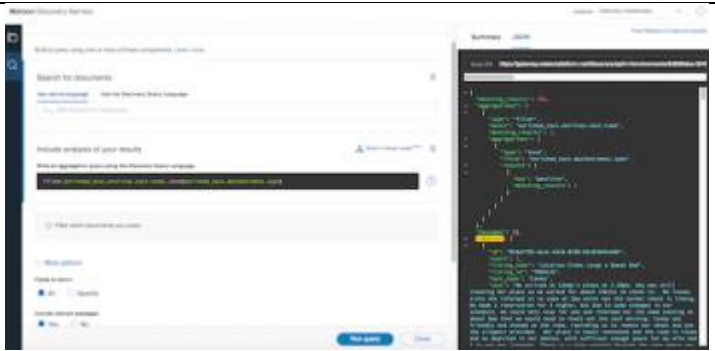
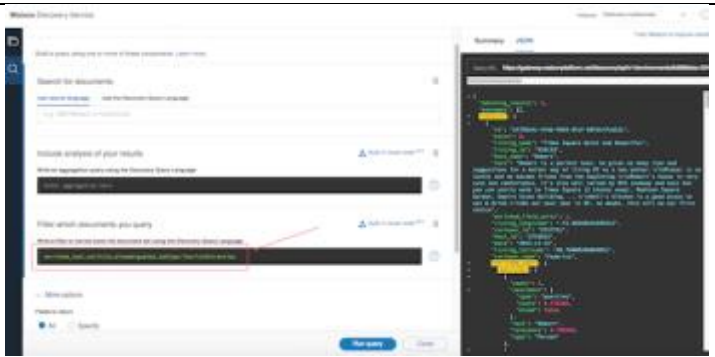



36. Extract the contents of the zip file that you downloaded from Github, select all and upload it to the Discovery Service. This may take a few minutes.
37. Once the upload is complete, click **Query this collection**. You will use the default configuration for this exercise.



38. Click the magnifying glass icon in the top left corner (below the folders).
39. Choose the AirBnB collection and click **Get Started**.
40. Take a moment and view the generated queries.
41. Without running any specific queries, just click **Build your own query** and click **Run query**.
42. Take a moment and view the results.
43. View the **enriched_text** section below the first passage.
44. Notice the hierarchy: docSentiment, entities, concepts. You will note that these are the enrichments that are applied per default configuration.

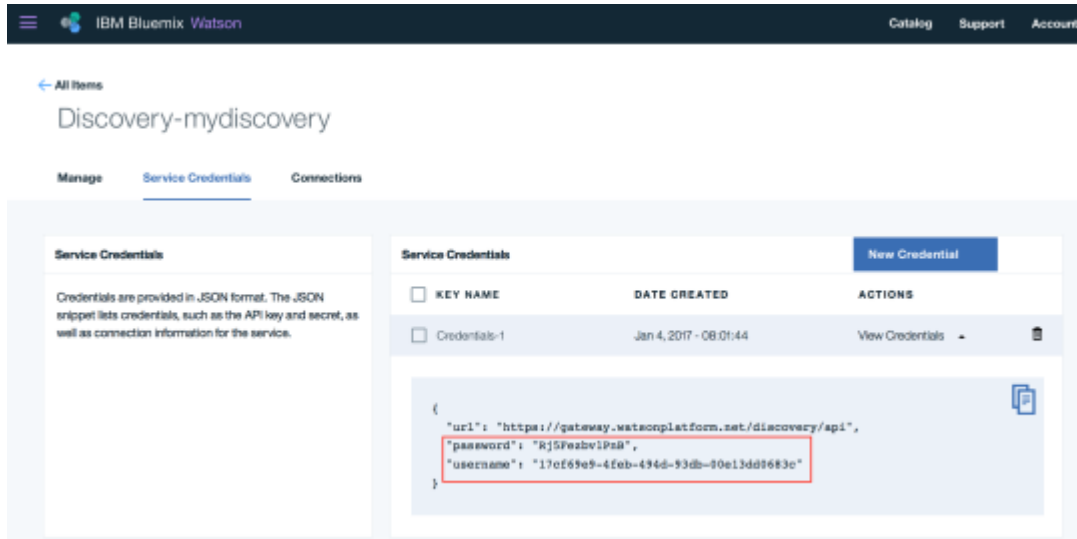


<p>You are now ready to run some queries.</p>	
<p>45. In the <i>Search for Documents</i>, type: room with a great view.</p> <p>46. Try another query in the same box, for example: quiet place.</p> <p>47. Now type this: bad.</p> <p>48. Click Learn more and take a moment and read the guidelines for querying.</p> <p>49. Clear all fields, and click Run query again.</p>	
<p>50. Open the <i>Include analysis of your results</i> field and then click the question mark next to the input field.</p> <p>51. 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.</p> <p>52. Click the JSON tab for more details.</p> <p>53. Repeat the above action with the Filter which documents you query section by clicking the question mark and following the example format.</p> <p><code>term(host_id)</code></p>	
<p>54. How about finding out apartments that are near tourist attractions? Clear all fields and run the following query in the <i>Filter which documents you query</i>...it's one line.</p> <p><code>enriched_text.entities.distance.subType:TouristAttraction</code></p> <p>55. Click Learn more from the provided link and take a moment to study the various queries that you can run.</p> <p>56. Explore the More options section and always refer to the Learn more link.</p>	
<p>57. Clear all fields.</p> <p>58. Click Run Query, for a fresh start.</p> <p>59. Copy the service URL in a temporary location.</p>	

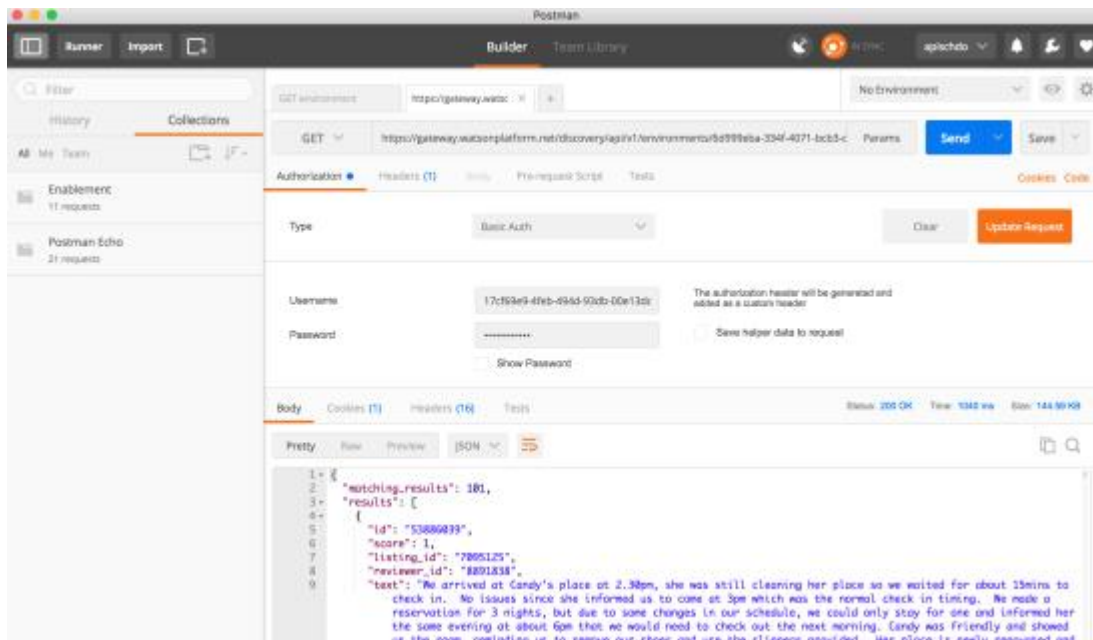
Working with the Discovery API (Optional)

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. This is why you downloaded the Postman API developer app noted earlier in this document. Let's begin.

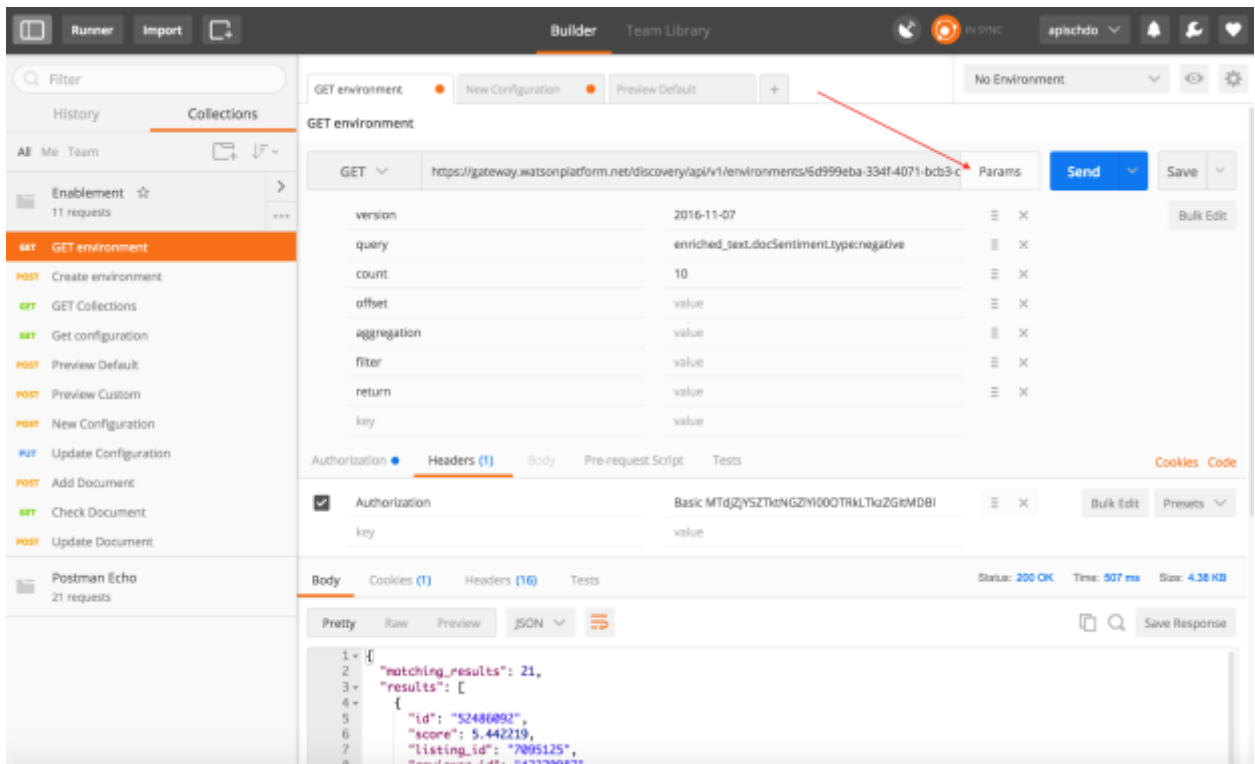
1. Go to Bluemix and access the Service Credentials panel of the Discovery service that you created in the beginning of this document.
2. Click the Service credentials link and note the username and password. You will be entering these in Postman along with the URL that copied earlier.



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.