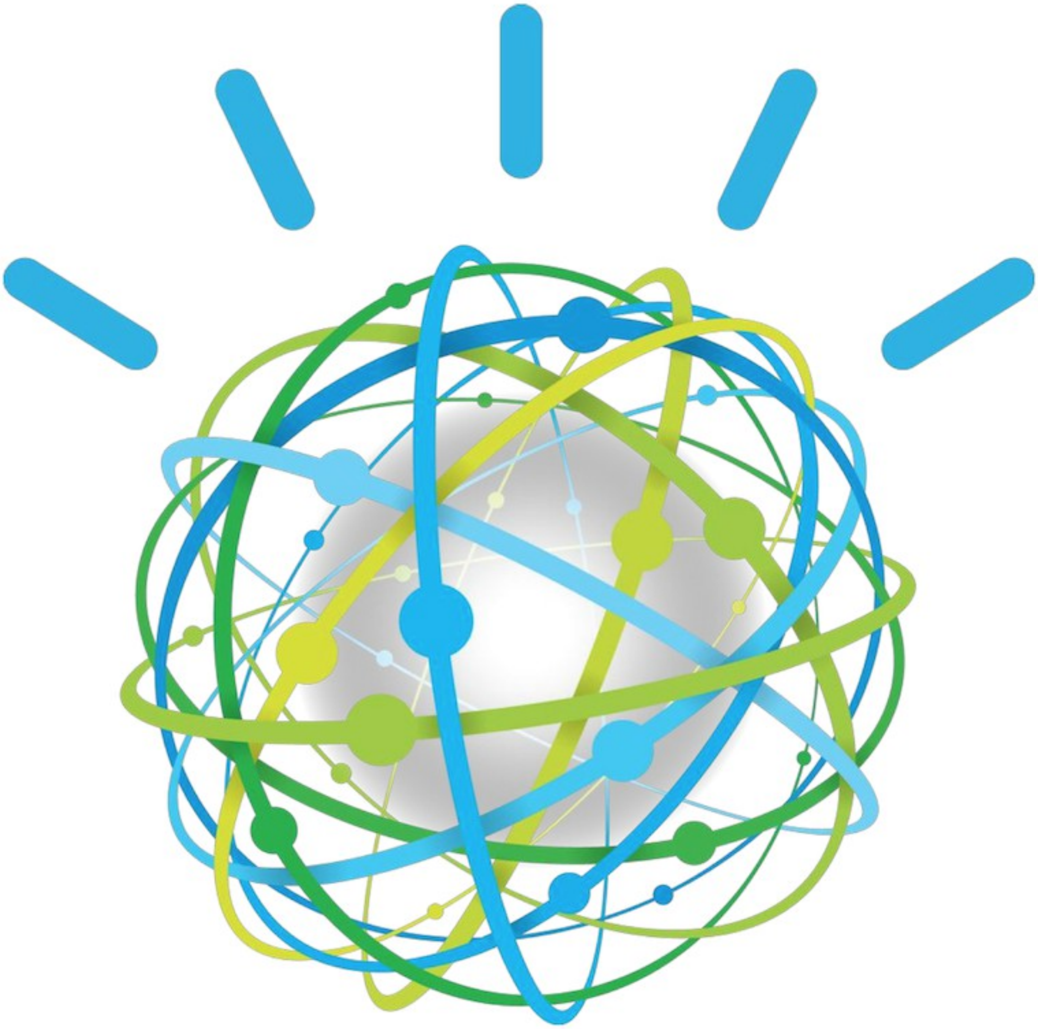
IBM Watson Ecosystem Academic Partners

Getting Started Guide



Version 2.1 – June 2015

IBM Watson

Overview

As a partner in the IBM Watson Ecosystem, you have a unique opportunity to build a one of a kind application using the cognitive computing power of IBM Watson.

This document introduces you to the tools and concepts you need to build your *powered-by-Watson* application.

You begin by learning about The IBM Watson Experience Manager web portal. This portal allows you to upload and test your data on the IBM Watson instance assigned to your team.

Then, you use the RESTful service API to connect to your data using sample source code and applications for a hybrid mobile, Android native and iOS native reference application.

Watson Experience Manager (WEM) is a browser-based tool that allows you to interact with your Watson instance (delivered via the Watson Developer Cloud) while developing your powered-by-Watson app.

This guide comprises two main sections:

- Working with the Watson Experience Manager (WEM)

  The first link in your on boarding email points to the Watson instance. By now you already have obtained your username and password from the on boarding emails that you received. Keep them handy, as later on in this guide you will find instructions on using them. It ends with **WatsonExperienceManager** context, where <nnn> is your particular instance number.

  https://watson-wdc01.ihost.com/instance/<*nnn*>/predeploy/WatsonExperienceManager

  or

  https://watson.ihost.com/instance/<*nnn*>/predeploy/WatsonExperienceManager

- Working with the QAAPI

  The second URL in your on boarding email, is the RESTful API that points you to the instance as an end-point where you can embed that in an application that you build. It ends with **deepqa/v1/question** context. Clicking this second link will give you an HTTP 500 Error message. Remember, this is just an end point API, and you must specify, the POST method along with Basic Auth parameters within the body of the API. More on that later.

  https://watson-wdc01.ihost.com/instance/<*nnn*>/deepqa/v1/question

Immediately following the Overview section is the **Prerequisite** section. You are presented with a set of prerequisite videos that will rapidly help you gain a deeper appreciation in the work that has gone into creating this cognitive system, plus, a glimpse to the direction it is heading.

Following the prerequisites, pay close attention to the **Guidelines** section and bear in mind some of the strongly recommended approaches as you begin working with your instance of Watson.

Working with the Watson Experience Manager (WEM)

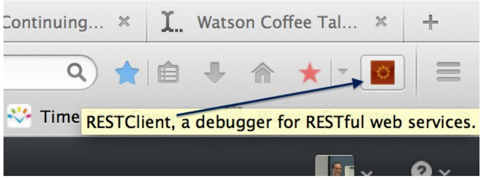
The table below outlines, at a high level, the sequence of task that you will perform in WEM:

Order of Steps	Task	Description
1	Assessing and configuring Watson	The first order of task is to set the proper user roles and adjust certain look and feel to your private instance.
2	Manage Projects	The discussions you have had surrounding having one or more corpora, will make impact at this stage.
2	Curate the documents prior to upload	Once you have found relevant documents that support your chosen domain, you must spend considerable time sanitizing, massaging or curating your Word, HTML or to a more limited extent, your PDF documents.
3	Upload and view TREC files	Uploading documents engages the preprocessor engine, which in turn, creates the TREC files, more on that later, which provide a structured format for downstream processes.
4	Identify answer units	Unlike a typical engagement where questions drive the content, for the Academic purposes, select headings or passages from your uploaded documents that would serve as answers to questions that you ask the system next.
5	Enter domain relevant questions	At this point you are ready to enter questions into Watson. Watson will return passages that support your question in descending confidence level.
6	Match question answer pairs	In this step, you begin answering the questions that you had entered per Watson's recommendation.
7	Cluster questions	Enter paraphrases of the original (parent) question. Watson's accuracy is markedly improved by clustering similar questions. Pairing questions with answers and clustering questions is known as ground truth development.
8	Create corpus	You ingest or create corpus only after having created your ground truth which comes after uploading the curated data. This delayed ingestion also indexes your question and answer units, making them more readily available to the Lucene search engines during the training phase.
9	Deploy and Test	At this time, the only event remaining is deploying your instance and testing the system for quality of answers.

Working with the QAAPI

The second URL that appears in your welcoming email provides the appropriate endpoint to push your question and get your response using APIs that you include in your applications that are powered by Watson.

Note that clicking that second link will result in a Watson Error pointing to missing Basic Auth parameters. Use a browser-specific RestClient plugin to explore as your REST API endpoint. More on that in the second half of this guide. The image below depicts the plugin for Firefox browser.



RESTClient, a debugger for RESTful web services.

Prerequisite

It is highly recommended that you view the following videos before starting to use the Watson Experience Manager.

- Watch the *IBM Watson Experience Manager Overview* tutorial which illustrates how to use IBM Watson Experience Manager for: (i) uploading content to create a Corpus, (ii) testing Watson by asking questions and exploring answers, (iii) configuring Watson to change the look and feel of Watson, and (iv) training Watson by providing a set of question-answer pairs.

  https://www.youtube.com/watch?v=aBYSEJ-cu1w

- Watson tutorial videos:
  - Module #1: What is IBM Watson? This video presents an overview of cognitive computing, IBM Watson and IBM’s strategy to proliferate cognitive computing in the marketplace with Watson.

    https://ibm.biz/watsontutorial-overview

  - Module #2: What is the IBM Watson Ecosystem? This video describes the IBM Watson Ecosystem, how it enables partners to develop cognitive computing applications, and the different types of partners within the ecosystem (Application Partners, Content Partners, Talent Partners). It also discusses how to get started, and what partner companies get when they become part of the IBM Watson Ecosystem.

    https://ibm.biz/watsontutorial-ecosystem

- Module #3: Roadmap for Building Your “Powered by Watson" Cognitive Application. This video outlines the roadmap (including key activities and milestones) for planning your IBM Watson application including the Engage, Prototype, Build and Deploy phases.

  https://ibm.biz/watsontutorial-plan

- Module #4-1: Designing and Developing Your "Powered by Watson" Cognitive Application. This video covers the process of building a “Powered by Watson” cognitive application, describes the Watson Experience Manager, (which is a browser based tool to access Watson) and touches on the kinds of content that are accepted by Watson.

  https://ibm.biz/watsontutorial-build

- Module #4-2: Enrich Your "Powered by Watson" Cognitive Application With Content. This video focuses on how to enrich Watson applications with most relevant content and describes how to upload and manage content using Watson Experience Manager.

  https://ibm.biz/watsontutorial-content

- Module #4-3: Train your "Powered by Watson" Cognitive Application. This video discusses training of Watson and illustrates one way to assist in training Watson by providing a set of question-answer pairs.

  https://ibm.biz/watsontutorial-train

- Module #4-4: Test and Deploy your "Powered by Watson" Cognitive Application. This video describes how to test the Watson application using Watson Experience Manager and highlights various considerations to think about when deploying the application including the user base, rollout plan, security provisions, and promoting the application to the production zone in the Watson Developer Cloud.

  https://ibm.biz/watsontutorial-testanddeploy

More videos will be posted on the IBM Watson Ecosystem YouTube channel as they become available: https://ibm.biz/watsonecosystem-youtube

For more background material, please consult the IBM developerWorks Watson community: https://developer.ibm.com/watson/

Guidelines

Follow these guidelines closely as you begin training the system in order to achieve higher quality of responses from Watson.

- Focus your team’s initial efforts in identifying one (and up to five) domains that are well focused and narrow in scope. For example, if your domain is travel, narrow it down to a particular city (Boston) and better yet specific attractions (light houses).
- Spend considerable time curating your raw documents; that is, ensuring that your PDFs are not just scanned OCR text, that your html files are free of busy advertisement or a java script running in the background, and that your Word documents are well segmented with headings tagged from the authoring tool itself.
- Resist clicking create corpus after uploading your documents. The Lucene search engine allows you to match answers to questions without having to ingest (or create corpus). After you have a robust number of question-answer pairs (around 400 or so), then click Create Corpus. This way, the uploaded documents and your ground truth (question-answer pairs) are all indexed.
- If you are using multiple projects, keep the contents of each project around 1 GB. This will give you a total of 5 GB for your instance. The reason being, during index creation (ingest), the directory expands tenfold before settling down to original size.
- Make an effort to attend the office hours that are held for each region as it pertains to you.
- Please note, that even though you can have up to 5 corpora, it is a single instance of Watson, this means **you can deploy one project at a time**. For example, students can work simultaneously on their projects, uploading documents, mapping question to answers, but they have to take turns when it comes to deploying their project. Each deployment undeploys the previously deployed project. Nothing is deleted, just the binding between the corpus and the pipeline is re-established with the new next project.

Accessing and configuring your Watson instance

To start, you use the **Administration** link in WEM to assign user roles to users. Next, you use the **Configuration** link to customize a desktop versus a mobile look and feel plus adjust the confidence threshold levels. Finally, you use WEM to start a project. Once a project is defined, you can navigate to that project and upload your application specific content and create a corpus. After you create a corpus, you can deploy it to the test pipeline and then, you can use the WEM UI test window to pass test questions to Watson and retrieve answers. Please note that this is not a programming model, where you explicitly select an answer and expect to get that same answer back. This statistical product is an advisor and it will return it's highest garnered confidence ranked answers based on the robustness of your ground truth, minimum of 400 approved question-answer pairs; clustered questions, up to 5 levels deep; well segmented documents and other consideration described later in this guide.

Accessing your private instance

To access your unique instance, replace the <*nnn*> in the URL links below with the Watson instance number that was provided to you in the Getting Started email.

The Watson Experience Manager has a comprehensive set of documentation that is hosted on your sandbox environment:

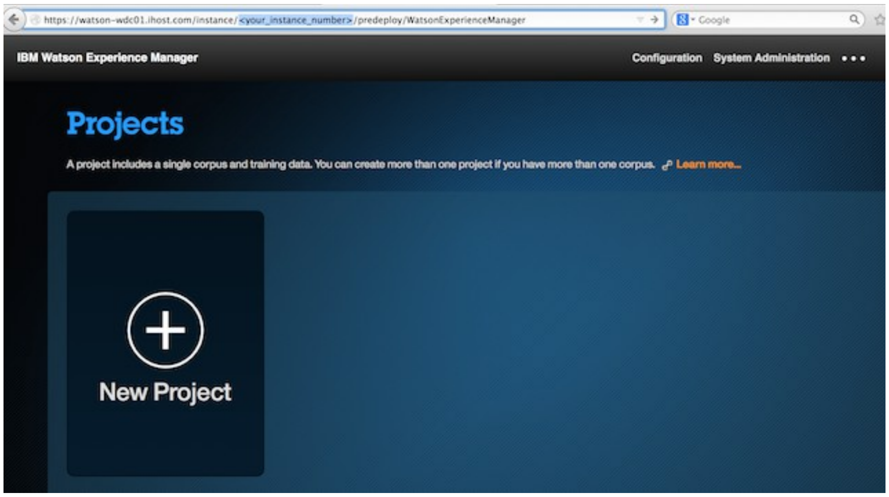
`https://watson-wdc01.ihost.com/instance/<nnn>/predeploy/index.jsp`

Complete the following steps to access the WEM homepage:

1. To access your Watson instance, enter the first URL from your on boarding email in a browser:

   `https://watson-wdc01.ihost.com/instance/<nnn>/predeploy/WatsonExperienceManager`

2. When prompted, enter your assigned username and password. Once authenticated into WEM, you will be redirected to the homepage.



https://watson-wdc01.ihost.com/instance/<your_instance_number>/predeploy/WatsonExperienceManager

A project includes a single corpus and training data. You can create more than one project if you have more than one corpus.

Administrating user roles

As an administrator (CUSTOMER_ADMIN), you can edit role assignments for other users of this system. The steps outlined in this guide depict the CUSTOMER_ADMIN role. Before we delve deeper in using the system, it is important, as an administrator, to assign appropriate roles to various personas in your group.

There are several well defined personas which represent different user roles and contexts in a Watson solution. Start by first understanding the different roles and identifying the people who can fill those roles in developing the powered-by-Watson application.

1. From the home drop down menu select **System Administration**.
2. Click the **User Role Administration** tab.
3. Select a user and click the edit (pencil) icon to the right.

4. You can allow users to have as many or less roles as the CUSTOMER_ADMIN. Include each on a separate line, followed by a comma (all upper case):

   ```
   CUSTOMER_ADMIN, GTT_APPROVER, GTT_APPROVER_EXT, GTT_AUTHOR,
   CORPUS_UPLOADER, CORPUS_DEVELOPER, CORPUS_MANAGER
   ```

5. Edit each role as it adheres to your requirements and click **OK**.

A user can have multiple roles and can access only the tools and areas that apply to the roles they are assigned. For example, you might have a subject matter expert who identifies and uploads documents from which Watson finds answers. This same person might also help create training questions and answers but not approve the questions and answers. In this example, assign the user the roles of GTT_AUTHOR and CORPUS_UPLOADER.

Table I: User Roles and Actions

User Role	Actions
CORPUS_MANAGER	✓ Create a corpus and deploy it to Watson ✓ View information and reports about a version of the corpus ✓ Upload and delete content ✓ Test Watson
CORPUS_UPLOADER	✓ Upload and Delete Content in Corpus Management tool ✓ View TREC files
CUSTOMER_ADMIN	✓ Access Configuration tool and Customize Watson ✓ Assign User Roles in System Administration tool ✓ Generate usage reports in System Administration tool ✓ Test Watson ✓ View TREC files
GTT_APPROVER	✓ Review, approve, and reject questions and associated answers in Expert Training tool
GTT_APPROVER_EXT	✓ Reassign questions to different users in Expert Training tool
GTT_AUTHOR	✓ Create and edit questions in Expert Training tool

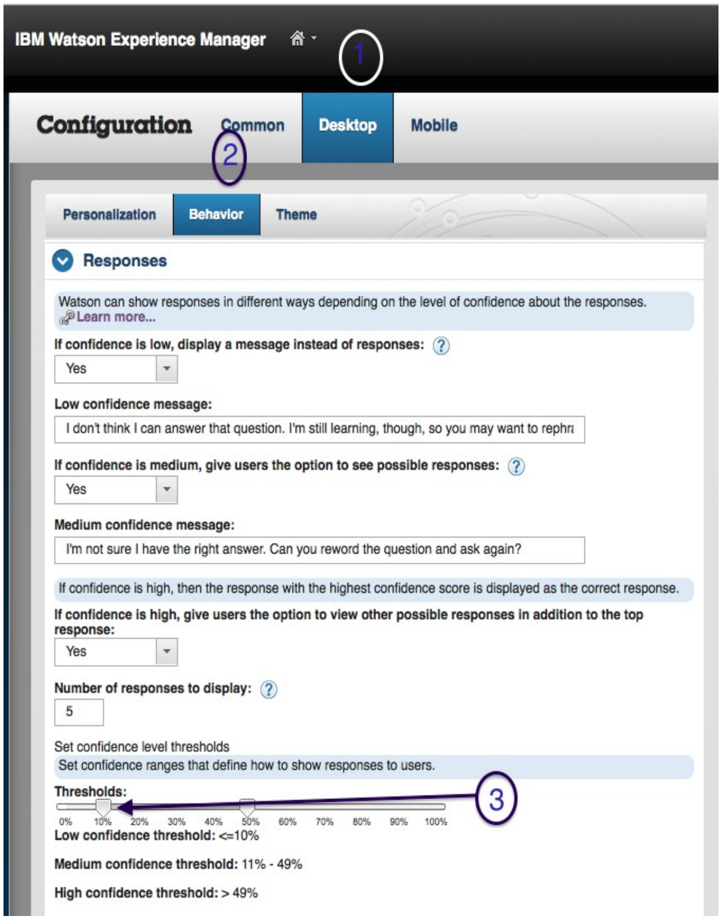
Configuring the UI for desktop and Mobile displays

Since this is a shared environment, any changes that you make to the configuration will override changes made by other users. In another words, the changes are not just per project but affect all projects since they are system wide.

One note worthy test that you can run, is by changing the confidence thresholds and observing if Watson is answering more questions, albeit at a lower confidence threshold.

1. Click **Configure Watson** from the home page.
2. Click the **Desktop** tab.
3. Click **Behavior**.

4. Adjust the low threshold to any value less than default of 10 or 20. In this example, 10. It is recommended that you lower the threshold down to a **value of 1 or 2** percent.



Responses

Watson can show responses in different ways depending on the level of confidence about the responses.
Learn more...

If confidence is low, display a message instead of responses: Yes

Low confidence message: I don't think I can answer that question. I'm still learning, though, so you may want to rephra

If confidence is medium, give users the option to see possible responses: Yes

Medium confidence message: I'm not sure I have the right answer. Can you reword the question and ask again?

If confidence is high, then the response with the highest confidence score is displayed as the correct response.

If confidence is high, give users the option to view other possible responses in addition to the top response: Yes

Number of responses to display: 5

Set confidence level thresholds
Set confidence ranges that define how to show responses to users.

Thresholds: 0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%

Low confidence threshold: <=10%

Medium confidence threshold: 11% - 49%

High confidence threshold: > 49%

5. Click **Apply Settings to Test**.

Managing Projects

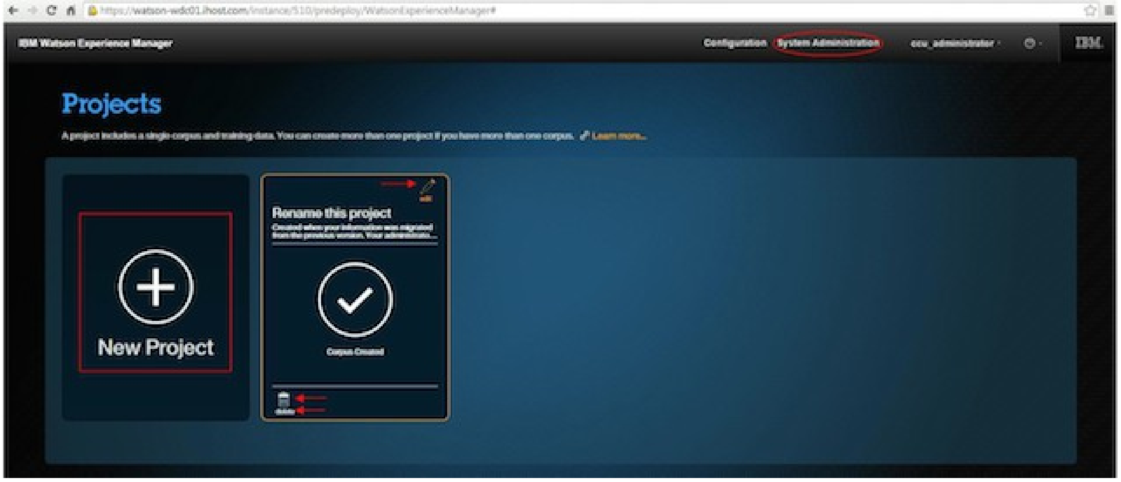
Watson Experience Manager allows you to create multiple projects where each project has a corpus and expert training data that is separate from other projects. This separation offers flexibility in how you organize your documents into a corpus and how you organize your expert training work.

Any user with CUSTOMER_ADMIN role can create, delete, rename, and change the description of a project. Projects have several features:

[Each project supports a single corpus that is separate from a corpus in any other project.

[Each project has its own expert training data which is separate from other projects. The answers you match to questions in Expert Training tool in project ABC are specific to the corpus created in project ABC.

You might create different projects to help compare how different documents affect how Watson answers expert training questions. You can also use projects to organize a large collection of documents into sub-collections. **However, although you can work on more than one project, you can deploy only one project to the pipeline** (meaning Watson can answer questions from the corpus of only one project).


https://watson-wdc01.ihost.com/instance/510/predeploy/WatsonExperienceManager#
IBM Watson Experience Manager
Configuration
System Administration
ccu_administrator
IBM
New Project
Rename this project
Corpus Created

Projects

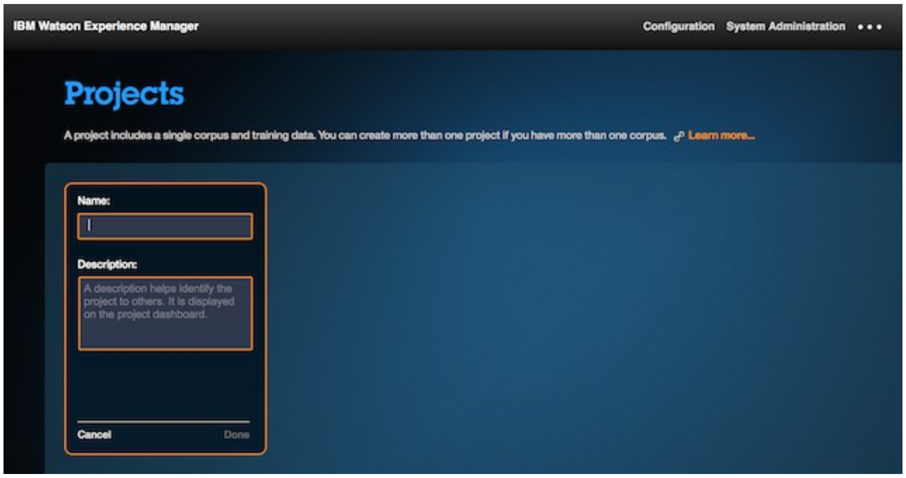
A project includes a single corpus and training data. You can create more than one project if you have more than one corpus. Learn more...

Created when your information was migrated from the previous version. Your administrato...

Create a new project

To create a new project complete the following steps:

1. Click the plus (+) sign to create anew project



Projects

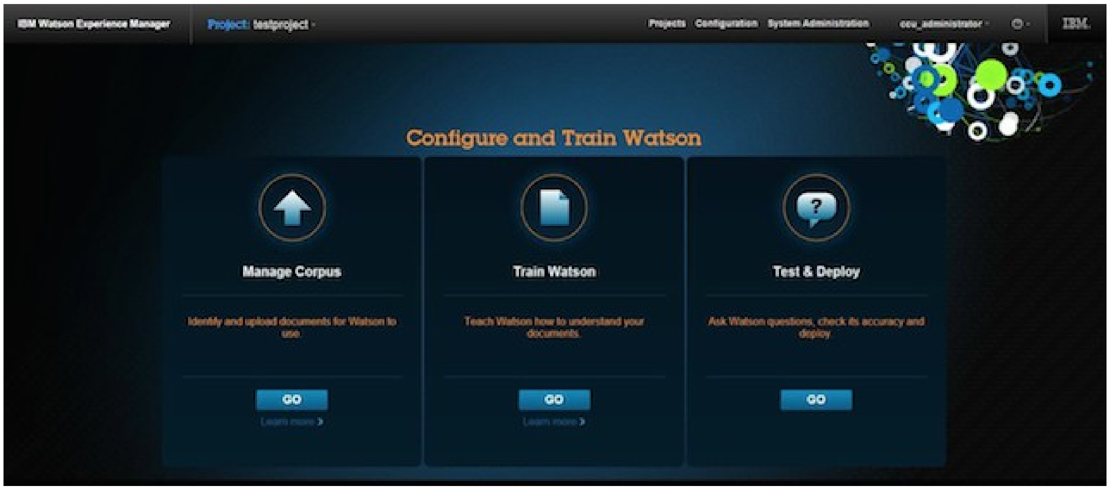
A project includes a single corpus and training data. You can create more than one project if you have more than one corpus. Learn more...

Name:

Description:

A description helps identify the project to others. It is displayed on the project dashboard.

2. Specify a unique name and description for your project. You can create up to 5 projects.
3. Click **Done**.
4. Click the center of your newly created project.



Identify and upload documents for Watson to use.

Teach Watson how to understand your documents.

Ask Watson questions, check its accuracy and deploy

You are now presented with the starting UI, also referred to as the Watson Experience Manager (WEM). You use the **Manage Corpus** to upload your documents, **Train Watson** to build the ground truth, (question-answer pairs) and **Test and Deploy** (or just Test) to check the quality of answers returned by Watson.

The test and deploy module infers from the training question-answer pairs that you entered to reveal answers that Watson deemed with highest confidence to be the proper response. This event takes advantage of the Watson pipeline and uses the out of the box machine learning models.

A well trained system, 50 to 100 uploaded documents (under 1 GB of space per project), plus 400 or more question answer pairs (again, per project), is best poised to answer domain related questions that was not in the ground truth. There are different levels of training that your system can undergo.

Your IBM Tech Mentor can take this a step further by running a **test set** to generate accuracy, recall and precision metrics. With the help of your Tech Mentor you can then perform some accuracy analysis tasks in identifying areas where you can improve on your ground truth.

Just as an additional background, a deeper level of training, known as a **training set**, is done on clusters by IBM development and is outside of the scope of the academic partners due to time and cost considerations. These tests are run on practically super computers, they are very human and machine resource intensive and are not for free. These training scripts create specific SPSS machine learning models based on your domain (not out of the box models that all partners get).

For large enterprise engagements and business partners, IBM also runs a **blind set** test to better gauge Watson performance.

Curate the documents prior to upload

Your objective is to curate or sanitize the raw documents before uploading them to the system. You likely have three document types: PDF, HTML and MS Word.

Watson can process many Microsoft Word documents. However, the following formats are not supported:

- Word 95 and earlier
- Word 2003 XML formats

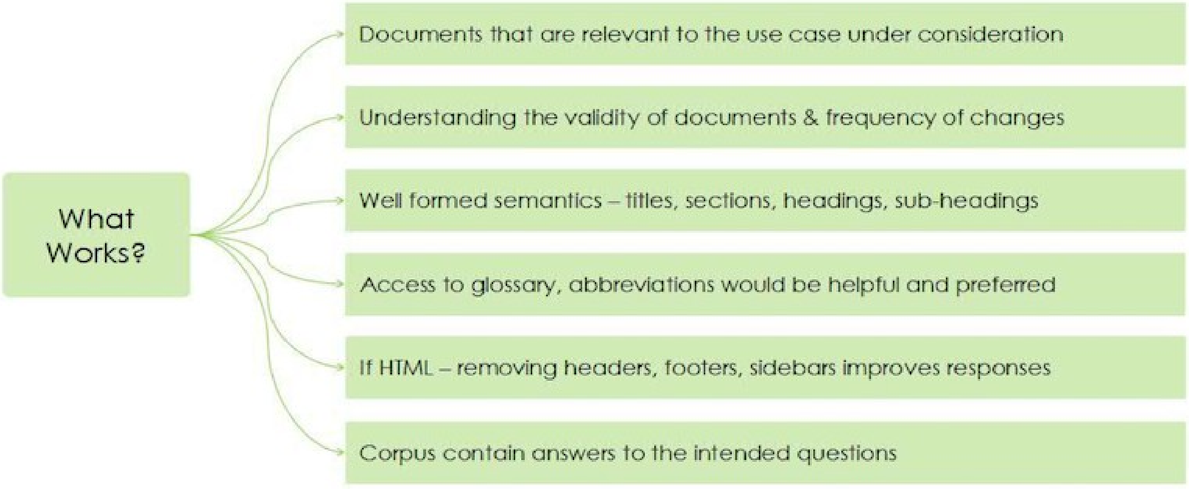
Both Word 95 and Word 2003 XML files use .doc extensions, so it can be difficult to tell if Watson can process them. Here is one method to make sure that you do not use these file formats. Open the document in version 2007 or later and save it with a .docx extension.

Enriching Watson's dictionaries with glossary and abbreviations relevant to your application domain help improve Watson's responses.

Please note that at this time, Watson does not understand graphics (images, videos, audio, etc.), non-English documents, or data with complex relationships such as nested tabular data, call data records, sensor logs, meter data, etc. Figure 5 illustrates examples of content that is "not good" such as scanned documents (OCR), charts, password-protected documents, noisy chat logs, nested tables and image-rich documents with no description. To reiterate, documents whose text content is not relevant to the application domain or don't include answers to end users' questions would not be very useful for Watson.

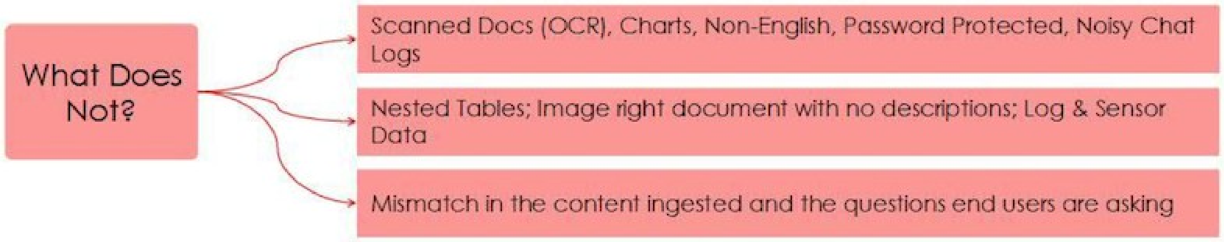
Guideline for preferred content

Documents that have a well defined and relevant title and are rich with headings and sub headings, are deemed as preferred content plus other consideration outlined in the figure below:


What Works?
Documents that are relevant to the use case under consideration
Understanding the validity of documents & frequency of changes
Well formed semantics – titles, sections, headings, sub-headings
Access to glossary, abbreviations would be helpful and preferred
If HTML – removing headers, footers, sidebars improves responses
Corpus contain answers to the intended questions

Characteristics of troublesome documents

Documents that contain tables (especially nested tables, or happen to be OCR scans masquerading as text as summarized below, may cause ingestion errors.


What Does Not?
Scanned Docs (OCR), Charts, Non-English, Password Protected, Noisy Chat Logs
Nested Tables; Image right document with no descriptions; Log & Sensor Data
Mismatch in the content ingested and the questions end users are asking

Admittedly, curating your raw content before uploading it to the system is a time consuming effort. However, these best practices will result in Watson returning succinct answers with highest confidence.

Curating PDF documents

PDF documents are not a favorite of Watson's ingestion engine. The PDF preprocessor, converts the PDF document to HTML before presenting it as answer units. It relies on font size and congruity of text to identify what is title and what is text.

Curating HTML documents

HTML pages, although a favorite of the Watson preprocessor, can be quite busy with graphics, advertisements, table sections, all of which are distractions from the pure text, the data that you are interested in.

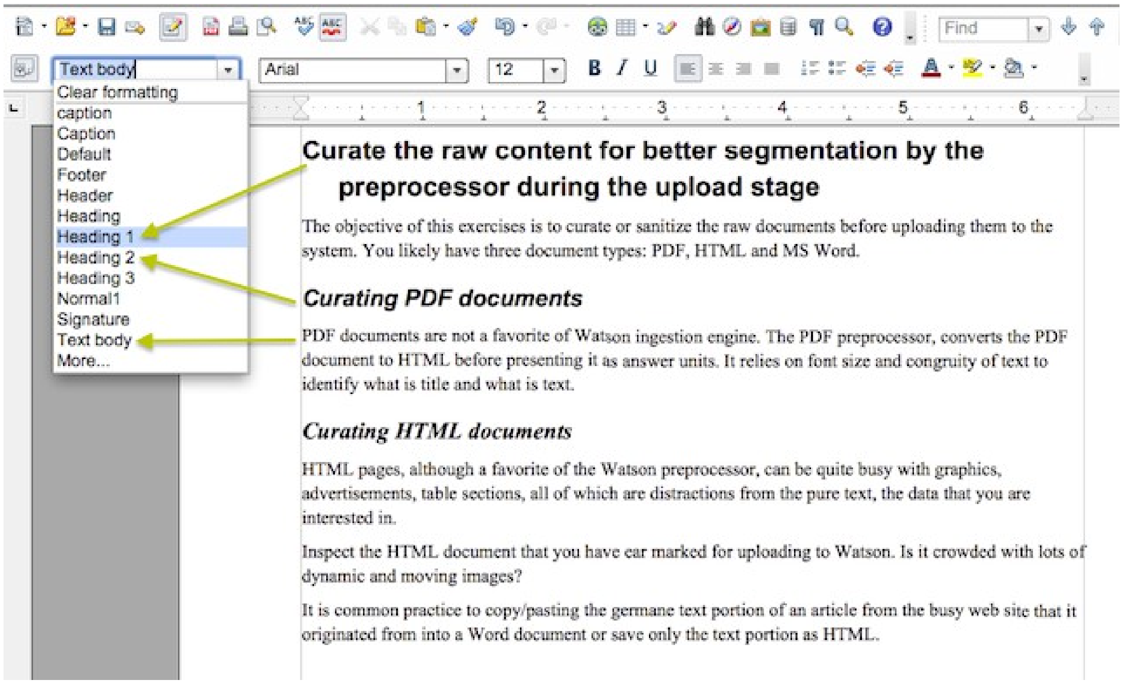
Inspect the HTML document that you have ear marked for uploading to Watson. Is it crowded with lots of dynamic and moving images?

It is common practice to copy/paste the germane text portion of an article from the busy web site that it originated from into a Word document or save only the text portion as HTML.

Curating MS Word documents

The Watson preprocessor works best in extracting titles and text from MS Word and HTML documents, simply because, unlike PDF, there happens to be tags for various segments. However, even an MS Word document may be crafted poorly, where the author, instead of selecting Head1 from the drop down loot box, has simply bolded a regular default text.

Open your Word documents and spend some time ensuring that the headings and texts are actual MS Word section headings or numbers as depicted in the image below:



Curate the raw content for better segmentation by the preprocessor during the upload stage

The objective of this exercises is to curate or sanitize the raw documents before uploading them to the system. You likely have three document types: PDF, HTML and MS Word.

Curating PDF documents

PDF documents are not a favorite of Watson ingestion engine. The PDF preprocessor, converts the PDF document to HTML before presenting it as answer units. It relies on font size and congruity of text to identify what is title and what is text.

Curating HTML documents

HTML pages, although a favorite of the Watson preprocessor, can be quite busy with graphics, advertisements, table sections, all of which are distractions from the pure text, the data that you are interested in.

Inspect the HTML document that you have ear marked for uploading to Watson. Is it crowded with lots of dynamic and moving images?

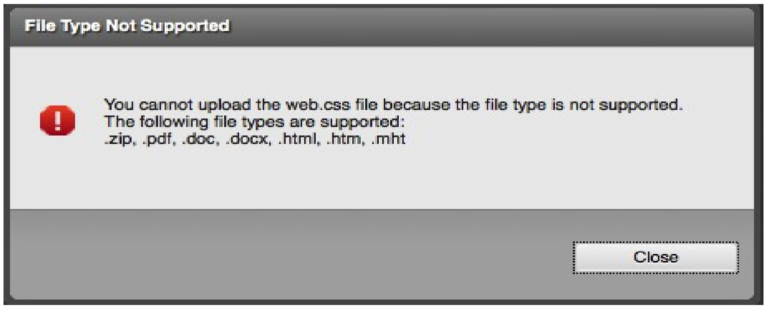
It is common practice to copy/pasting the germane text portion of an article from the busy web site that it originated from into a Word document or save only the text portion as HTML.

Upload the curated domain-specific content and inspect the TREC files

You upload documents via the Content Management Tool (CMT). Uploading the documents also runs the preprocessor code. The preprocessor creates the TREC (Text Retrieval Conference) files and if you have done a thorough job of tagging the original document with proper headings in Word or tags in the HTML, it will reveal smaller and more precise or succinct answer units which you can then link to your primary question.

TREC is an abbreviation for NIST's [National Institute of Standards and Technology] Text REtrieval Conference. In order for the indexer to know where the document boundaries are within files, each document must have begin document and end document tags. These tags are similar to HTML or XML tags and are actually the format for TREC documents.

If you happen to upload a file type that is not supported you will get the following message indicating the exact file types that Watson can ingest:



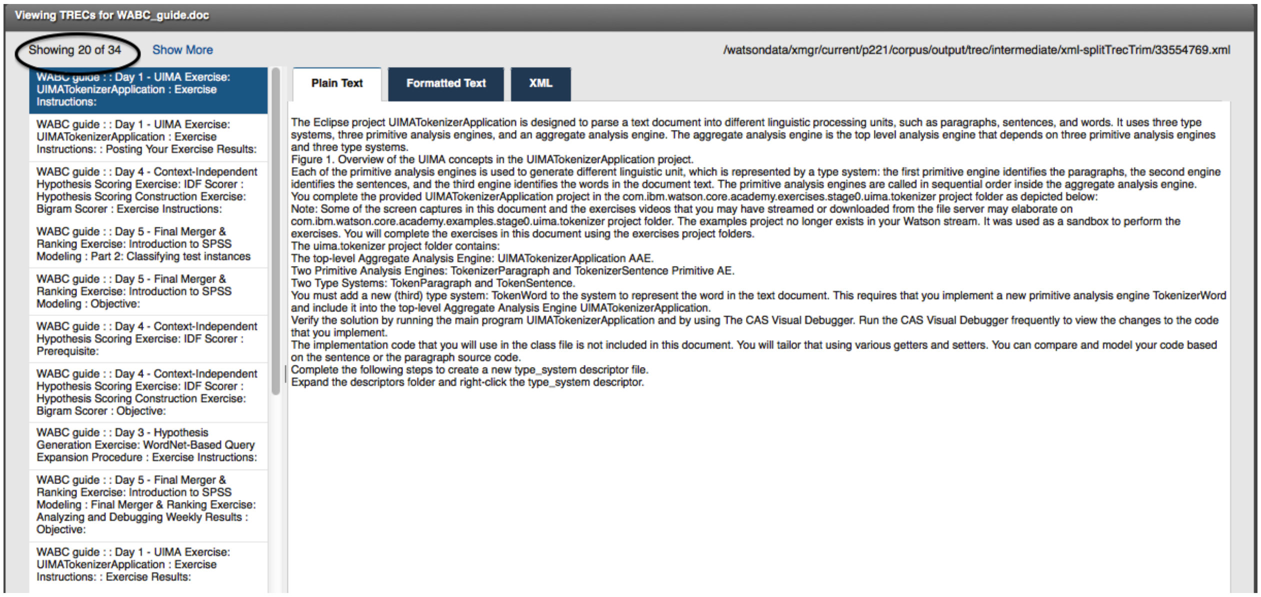
You cannot upload the web.css file because the file type is not supported.
The following file types are supported:
.zip, .pdf, .doc, .docx, .html, .htm, .mht

To begin uploading documents to WEM, complete the following steps:

1. From the Home drop down menu (upper left corner of the WEM tool) select **Corpus Management** or click the **Manage Corpus** box from the landing page.
2. Click **Upload** and browse to the location of the file (or zip file, if you want to do it in a batch) on your computer. For this exercise, start with the MS Word document.

    > Clicking the Content Store will download literally, thousands of mini html documents that you can begin to work with without having to start from scratch. However, this content currently is demo data bound to travel or healthcare domain and the sheer number of downloaded documents may cause performance issues. It is recommended that you use the upload option and grab your own domain specific documents.

    Allow for the preprocessor to complete it's task.
3. Click **View TRECs** and inspect the segmented answer units. In this example the default configuration resulted in 34 answer units, one or more of which can be later mapped to the primary question.



/watsondata/xmgr/current/p221/corpus/output/trec/intermediate/xml-splitTrecTrim/33554769.xml

WABC guide : : Day 1 - UIMA Exercise: UIMATokenizerApplication : Exercise Instructions:

WABC guide : : Day 1 - UIMA Exercise: UIMATokenizerApplication : Exercise Instructions: : Posting Your Exercise Results:

WABC guide : : Day 4 - Context-Independent Hypothesis Scoring Exercise: IDF Scorer : Hypothesis Scoring Construction Exercise: Bigram Scorer : Exercise Instructions:

WABC guide : : Day 5 - Final Merger & Ranking Exercise: Introduction to SPSS Modeling : Part 2: Classifying test instances

WABC guide : : Day 5 - Final Merger & Ranking Exercise: Introduction to SPSS Modeling : Objective:

WABC guide : : Day 4 - Context-Independent Hypothesis Scoring Exercise: IDF Scorer : Prerequisite:

WABC guide : : Day 4 - Context-Independent Hypothesis Scoring Exercise: IDF Scorer : Hypothesis Scoring Construction Exercise: Bigram Scorer : Objective:

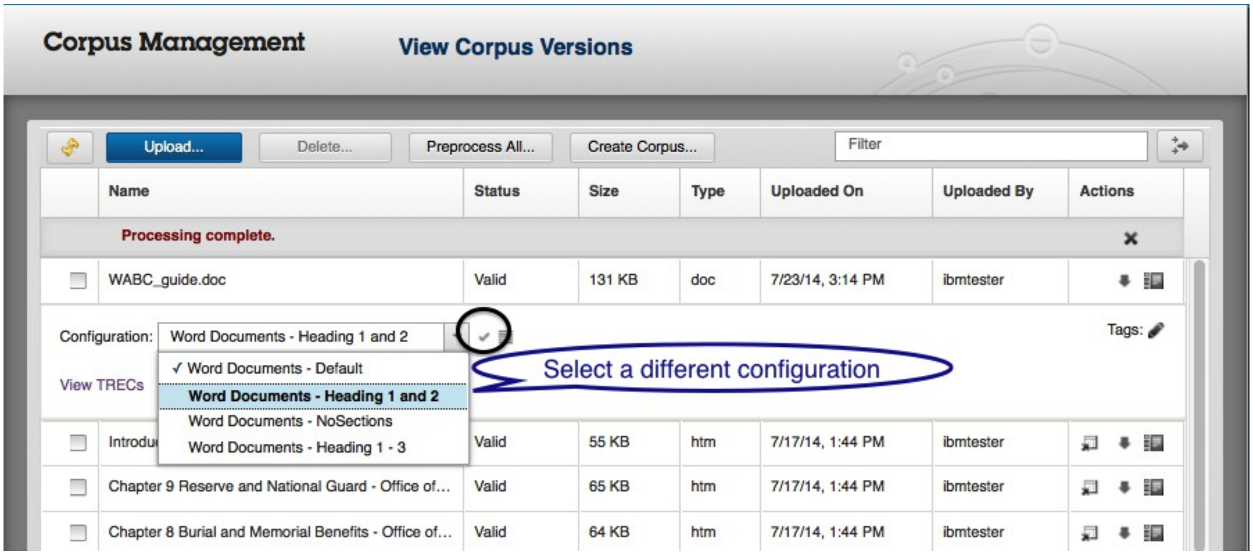
WABC guide : : Day 3 - Hypothesis Generation Exercise: WordNet-Based Query Expansion Procedure : Exercise Instructions:

WABC guide : : Day 5 - Final Merger & Ranking Exercise: Introduction to SPSS Modeling : Final Merger & Ranking Exercise: Analyzing and Debugging Weekly Results : Objective:

WABC guide : : Day 1 - UIMA Exercise: UIMATokenizerApplication : Exercise Instructions: : Exercise Results:

The Eclipse project UIMATokenizerApplication is designed to parse a text document into different linguistic processing units, such as paragraphs, sentences, and words. It uses three type systems, three primitive analysis engines, and an aggregate analysis engine. The aggregate analysis engine is the top level analysis engine that depends on three primitive analysis engines and three type systems.
Figure 1. Overview of the UIMA concepts in the UIMATokenizerApplication project.
Each of the primitive analysis engines is used to generate different linguistic unit, which is represented by a type system: the first primitive engine identifies the paragraphs, the second engine identifies the sentences, and the third engine identifies the words in the document text. The primitive analysis engines are called in sequential order inside the aggregate analysis engine.
You complete the provided UIMATokenizerApplication project in the com.ibm.watson.core.academy.exercises.stage0.uima.tokenizer project folder as depicted below:
Note: Some of the screen captures in this document and the exercises videos that you may have streamed or downloaded from the file server may elaborate on com.ibm.watson.core.academy.examples.stage0.uima.tokenizer project folder. The examples project no longer exists in your Watson stream. It was used as a sandbox to perform the exercises. You will complete the exercises in this document using the exercises project folders.
The uima.tokenizer project folder contains:
The top-level Aggregate Analysis Engine: UIMATokenizerApplication AAE.
Two Primitive Analysis Engines: TokenizerParagraph and TokenizerSentence Primitive AE.
Two Type Systems: TokenParagraph and TokenSentence.
You must add a new (third) type system: TokenWord to the system to represent the word in the text document. This requires that you implement a new primitive analysis engine TokenizerWord and include it into the top-level Aggregate Analysis Engine UIMATokenizerApplication.
Verify the solution by running the main program UIMATokenizerApplication and by using The CAS Visual Debugger. Run the CAS Visual Debugger frequently to view the changes to the code that you implement.
The implementation code that you will use in the class file is not included in this document. You will tailor that using various getters and setters. You can compare and model your code based on the sentence or the paragraph source code.
Complete the following steps to create a new type_system descriptor file.
Expand the descriptors folder and right-click the type_system descriptor.

4. Click **Close**.
5. Open the arrow next to the MS Word document you uploaded and from the **Configuration** drop down list, select a different configuration and click the *check mark* to the right of the configuration. This action runs the preprocessor again.



Processing complete.

Name	Status	Size	Type	Uploaded On	Uploaded By	Actions
WABC_guide.doc	Valid	131 KB	doc	7/23/14, 3:14 PM	ibmtester	
Introdu	Valid	55 KB	htm	7/17/14, 1:44 PM	ibmtester	
Chapter 9 Reserve and National Guard - Office of...	Valid	65 KB	htm	7/17/14, 1:44 PM	ibmtester	
Chapter 8 Burial and Memorial Benefits - Office of...	Valid	64 KB	htm	7/17/14, 1:44 PM	ibmtester	

Configuration: Word Documents - Heading 1 and 2

- ✓ Word Documents - Default
- Word Documents - Heading 1 and 2
- Word Documents - NoSections
- Word Documents - Heading 1 - 3

Select a different configuration

6. Open the newly generated TREC file. Does it look any different? Compare the answer units that were generated against the Default configuration.
7. Upload all of the remaining domain specific documents that you prepared for this workshop.

Enter domain relevant questions

The objective is to begin the process of collecting questions, leading to what constitutes “representative questions.” Typically, at a client engagement, representative questions number around 2000. For the purposes of academic partners, it is recommended that you enter upwards of 400 or more questions.

Typically, the role of GTT_AUTHOR is responsible for entering questions.

Note that the answer(s) for the representative questions for a project must exist in the uploaded content and associated corpus for that same project.

You are now ready to enter questions from the ground truth tool, also know as the Train Watson module or **Expert Training** from the home pull down menu.

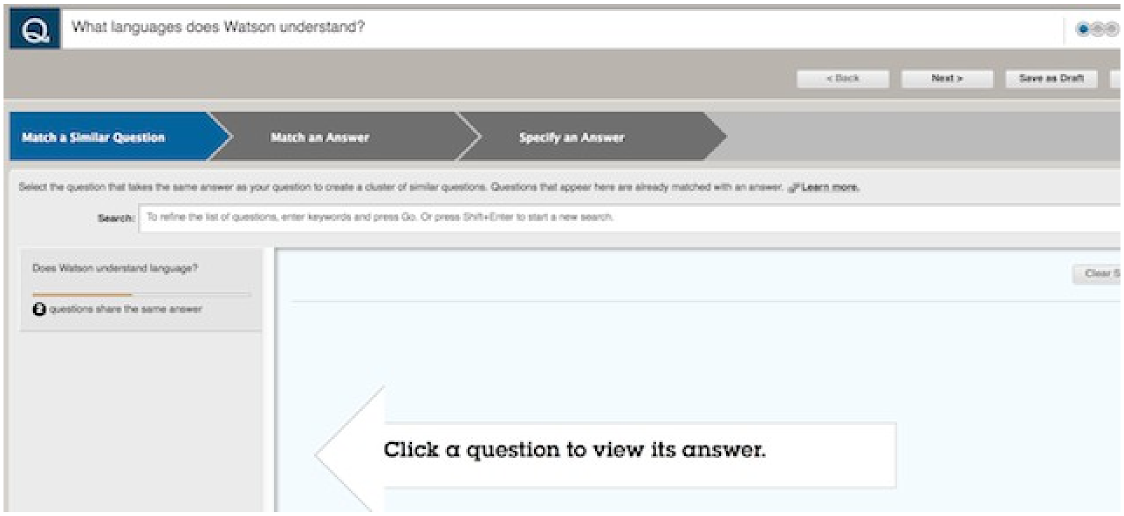
1. Click the **Train Watson** module (middle box).
2. Enter the questions that you prepared as part of the pre-req into the system and click **Add Question.**
3. Click **Save as Draft** after entering each question.

You can either choose to answer the questions right away and cluster them later, or upon entering questions, you can save it as draft and return back to answer them. In this section, you focus on just collecting questions (and delaying matching them with an answer until later).

Cluster question paraphrases

Clustering paraphrases of the original question directly affects the accuracy metrics of Watson. Typically, you may want to paraphrase the same question up to 4 or 5 times and cluster those handful of paraphrases. By matching a question to another question you help Watson learn to associate questions and answers that may contain different terms, idioms, phrases and other natural language nuances that are unique to your domain. This helps improve the quality of answers that Watson provides at run-time. Complete the following steps:

1. Enter a paraphrase of one of the questions that you entered earlier.
2. The system may detect a similar question and prompt you to cluster the new question with the parent question. For a new question, it will prompt you to Match an Answer. Again, you can continue with Matching an Answer from the generated list, or merely saving the question as draft and coming back to answering them later.



What languages does Watson understand?

Select the question that takes the same answer as your question to create a cluster of similar questions. Questions that appear here are already matched with an answer. Learn more.

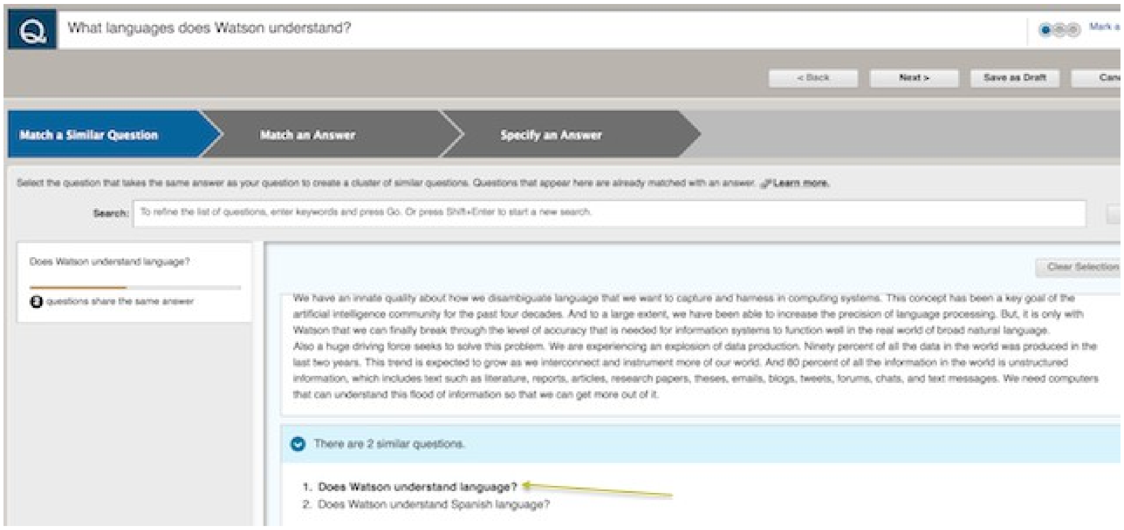
Search: To refine the list of questions, enter keywords and press Go. Or press Shift+Enter to start a new search.

Does Watson understand language?

2 questions share the same answer

Click a question to view its answer.

3. Select the system discovered similar question; scroll to the bottom of the passage and open the collapse section. The question that is bolded is the parent question. **Click Select** to cluster this paraphrase with the parent question.



What languages does Watson understand?

Select the question that takes the same answer as your question to create a cluster of similar questions. Questions that appear here are already matched with an answer.

To refine the list of questions, enter keywords and press Go. Or press Shift+Enter to start a new search.

Does Watson understand language?

3 questions share the same answer

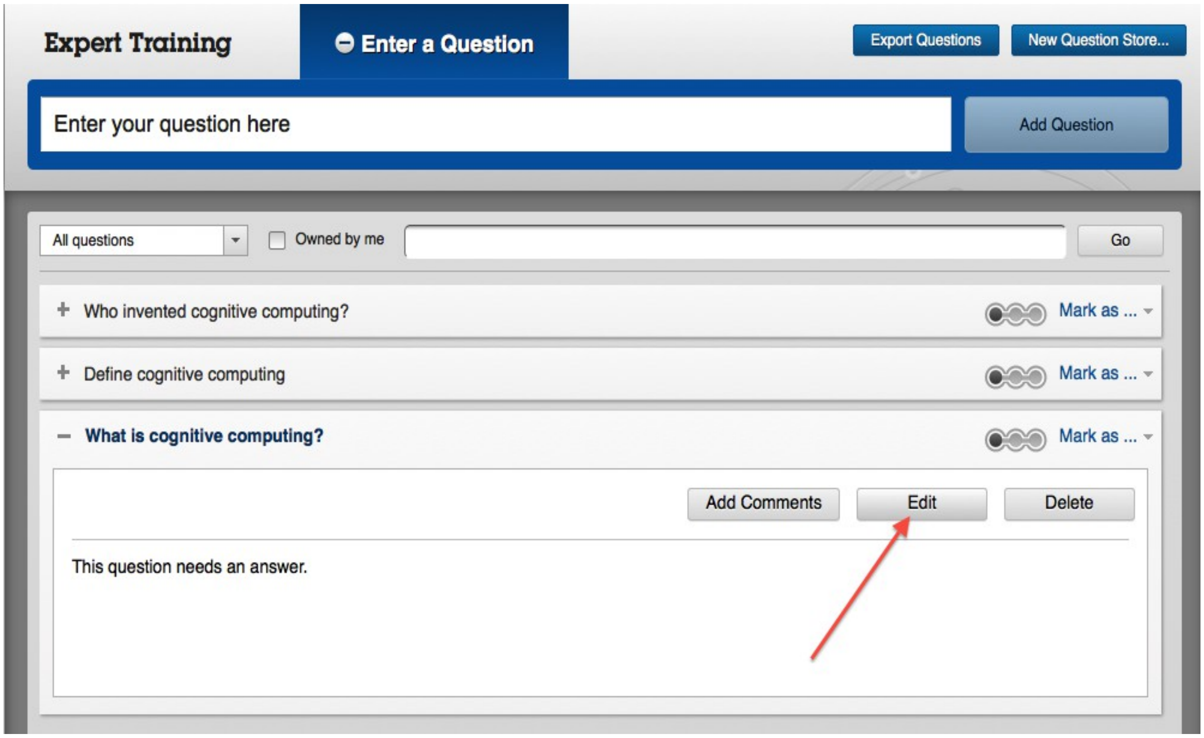
We have an innate quality about how we disambiguate language that we want to capture and harness in computing systems. This concept has been a key goal of the artificial intelligence community for the past four decades. And to a large extent, we have been able to increase the precision of language processing. But, it is only with Watson that we can finally break through the level of accuracy that is needed for information systems to function well in the real world of broad natural language.
Also a huge driving force seeks to solve this problem. We are experiencing an explosion of data production. Ninety percent of all the data in the world was produced in the last two years. This trend is expected to grow as we interconnect and instrument more of our world. And 80 percent of all the information in the world is unstructured information, which includes text such as literature, reports, articles, research papers, theses, emails, blogs, tweets, forums, chats, and text messages. We need computers that can understand this flood of information so that we can get more out of it.

There are 2 similar questions.

1. Does Watson understand language?
2. Does Watson understand Spanish language?

At this time you are not interested in matching answers to your questions, but merely creating a robust set of question paraphrases. You don't have to do this for all questions, but statistically you will get better inferencing with clusterings that are 4 or 5 levels deep. Each time as you click **Save as Draft**, it will allow you to move on to the next question.

4. You can also add clustered questions by opening the details of a question and clicking **Edit**.



Enter your question here

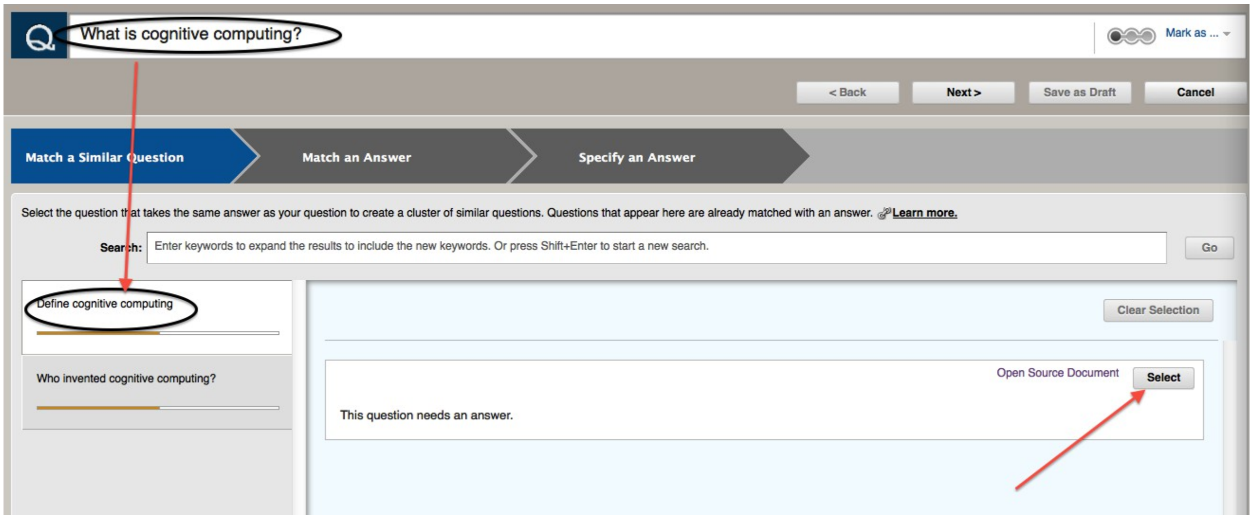
+ Who invented cognitive computing?

+ Define cognitive computing

− What is cognitive computing?

This question needs an answer.

5. Select one or more questions that the system has deemed similar, and that you agree, and cluster or link them to your primary question by clicking the **Select** button.



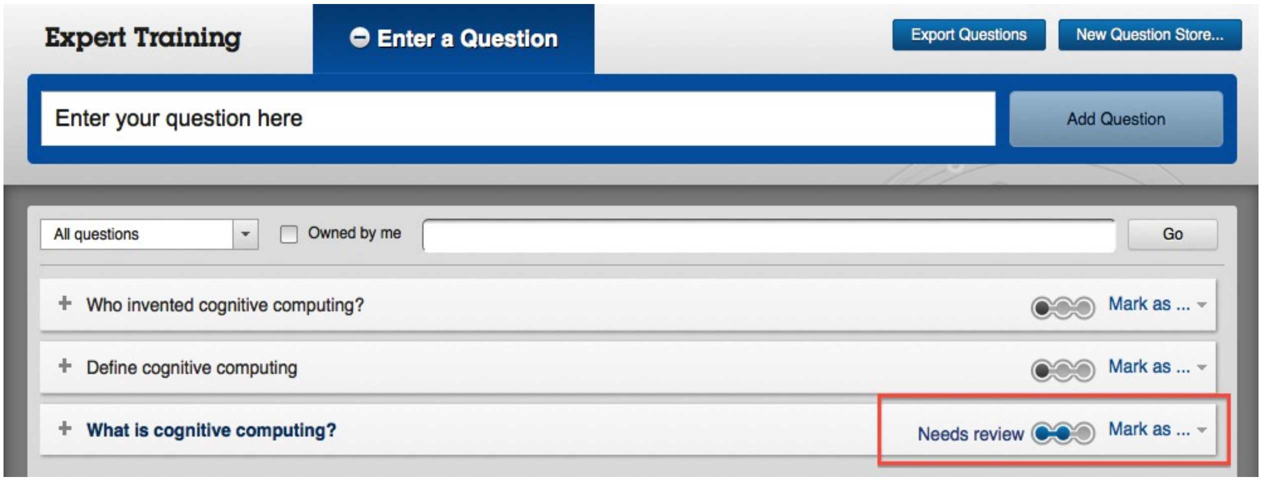
Select the question that takes the same answer as your question to create a cluster of similar questions. Questions that appear here are already matched with an answer. Learn more.

Enter keywords to expand the results to include the new keywords. Or press Shift+Enter to start a new search.

This question needs an answer.

6. Spend a few minutes and link as many question to other similar question that the system has picked as paraphrases of the original or parent question. Typically, 4 to 5 levels results in higher quality of answers by Watson.

7. Click **Save**. Notice that the workflow progress bar indicates that the question needs review.



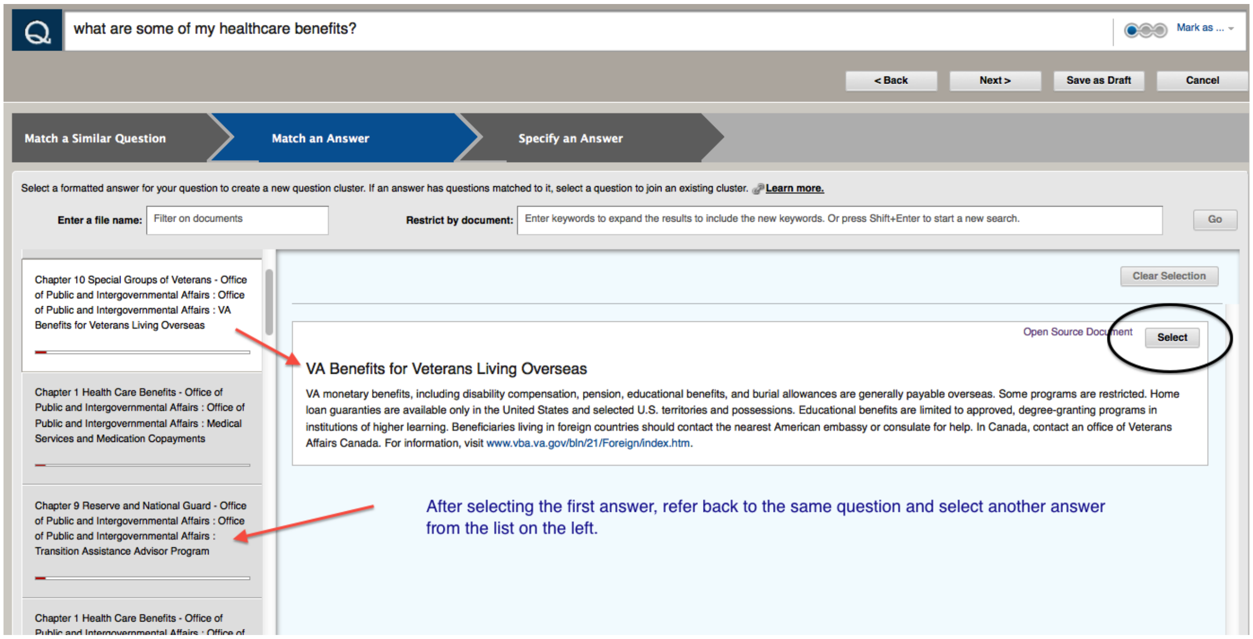
Mapping answers to questions

You are ready to link one or more answer units to the questions that appear in the Ground Truth.

Complete the following steps:

1. From the **Expert Training** module (also known as the Ground Truth Tool), select a question that you had entered and expand the plus sign to the left of it.
2. Click **Next**. Notice that the Match an Answer section is highlighted.

   You will also notice that a number of relevant answer units are displayed. At this point, the system has merely used a term lookup search capability to retrieve these relevant documents.
3. Click some of the passages that appear to the left and study the retrieved segment.
4. When you find a relevant answer, click **Select** to pick the answer unit that you feel best answers the question. Preferably, repeat this procedure and select more than one answer per question.



what are some of my healthcare benefits?

Select a formatted answer for your question to create a new question cluster. If an answer has questions matched to it, select a question to join an existing cluster.

Enter a file name: Filter on documents

Restrict by document: Enter keywords to expand the results to include the new keywords. Or press Shift+Enter to start a new search.

- Chapter 10 Special Groups of Veterans - Office of Public and Intergovernmental Affairs : Office of Public and Intergovernmental Affairs : VA Benefits for Veterans Living Overseas
- Chapter 1 Health Care Benefits - Office of Public and Intergovernmental Affairs : Office of Public and Intergovernmental Affairs : Medical Services and Medication Copayments
- Chapter 9 Reserve and National Guard - Office of Public and Intergovernmental Affairs : Office of Public and Intergovernmental Affairs : Transition Assistance Advisor Program
- Chapter 1 Health Care Benefits - Office of Public and Intergovernmental Affairs : Office of

VA Benefits for Veterans Living Overseas

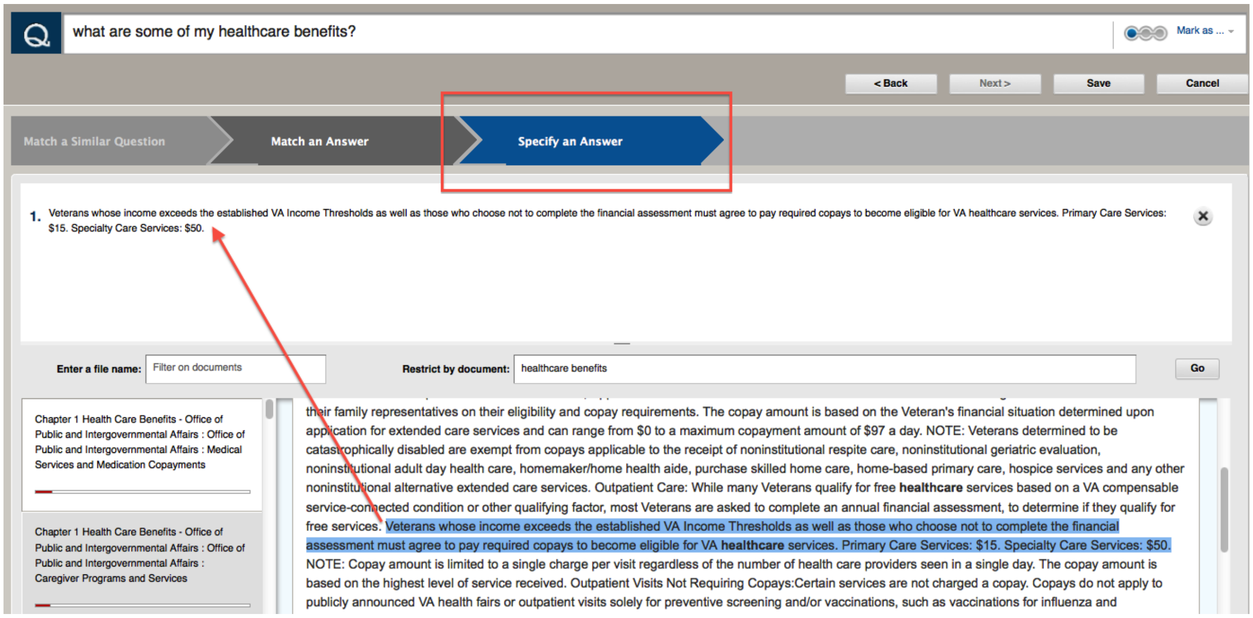
VA monetary benefits, including disability compensation, pension, educational benefits, and burial allowances are generally payable overseas. Some programs are restricted. Home loan guaranties are available only in the United States and selected U.S. territories and possessions. Educational benefits are limited to approved, degree-granting programs in institutions of higher learning. Beneficiaries living in foreign countries should contact the nearest American embassy or consulate for help. In Canada, contact an office of Veterans Affairs Canada. For information, visit www.vba.va.gov/bln/21/Foreign/index.htm.

After selecting the first answer, refer back to the same question and select another answer from the list on the left.

There are times, that you may have selected the “sibling” or child question. Should you decide to match an answer to the child question, the system invokes the **Specify and Answer** tab, where you must highlight a particular sentence or paragraph from the provided answer unit.

It is important to note, that at this point, it is good practice to *stop* and find the parent question, which appears in bold in the detail section, and answer the parent question instead even if you are adding additional answer units. That way, you get the chance to click **Select** from **Match an Answer** tab. That is the preferred method. The code behind the Select button, does not take your answer to a Derby database (as is the case with highlighting) but stores it in a special directory to be indexed and available to Lucene's search engine.

5. Although you may want to answer a few questions using the Specify and Answer method: click **Next**. Notice that the **Specify an Answer** section is now highlighted.
6. In the **Restricted by document** text box, type in one or more words pertinent to your question and press **Shift+Enter**.
7. Highlight a sentence or two that best answers the question.



1. Veterans whose income exceeds the established VA Income Thresholds as well as those who choose not to complete the financial assessment must agree to pay required copays to become eligible for VA healthcare services. Primary Care Services: $15. Specialty Care Services: $50.

Enter a file name: Filter on documents

Restrict by document: healthcare benefits

Chapter 1 Health Care Benefits - Office of Public and Intergovernmental Affairs : Office of Public and Intergovernmental Affairs : Medical Services and Medication Copayments

Chapter 1 Health Care Benefits - Office of Public and Intergovernmental Affairs : Office of Public and Intergovernmental Affairs : Caregiver Programs and Services

their family representatives on their eligibility and copay requirements. The copay amount is based on the Veteran's financial situation determined upon application for extended care services and can range from $0 to a maximum copayment amount of $97 a day. NOTE: Veterans determined to be catastrophically disabled are exempt from copays applicable to the receipt of noninstitutional respite care, noninstitutional geriatric evaluation, noninstitutional adult day health care, homemaker/home health aide, purchase skilled home care, home-based primary care, hospice services and any other noninstitutional alternative extended care services. Outpatient Care: While many Veterans qualify for free **healthcare** services based on a VA compensable service-connected condition or other qualifying factor, most Veterans are asked to complete an annual financial assessment, to determine if they qualify for free services. Veterans whose income exceeds the established VA Income Thresholds as well as those who choose not to complete the financial assessment must agree to pay required copays to become eligible for VA **healthcare** services. Primary Care Services: $15. Specialty Care Services: $50. NOTE: Copay amount is limited to a single charge per visit regardless of the number of health care providers seen in a single day. The copay amount is based on the highest level of service received. Outpatient Visits Not Requiring Copays:Certain services are not charged a copay. Copays do not apply to publicly announced VA health fairs or outpatient visits solely for preventive screening and/or vaccinations, such as vaccinations for influenza and

8. Click **Save**.

   After you save the mapped question-answer pair, expand the plus sign next to the question and you will be presented with a workflow of other actions that you must take such as **Add Comments, Approve, Reject, Edit,** and **Delete**. Depending on your role and user privileges set in the Administrative tool, you may have some or all of these buttons visible.

9. Once satisfied with your selection, click **Approve**.
10. Match all of the questions that pertain to you with an answer or two and approve them.

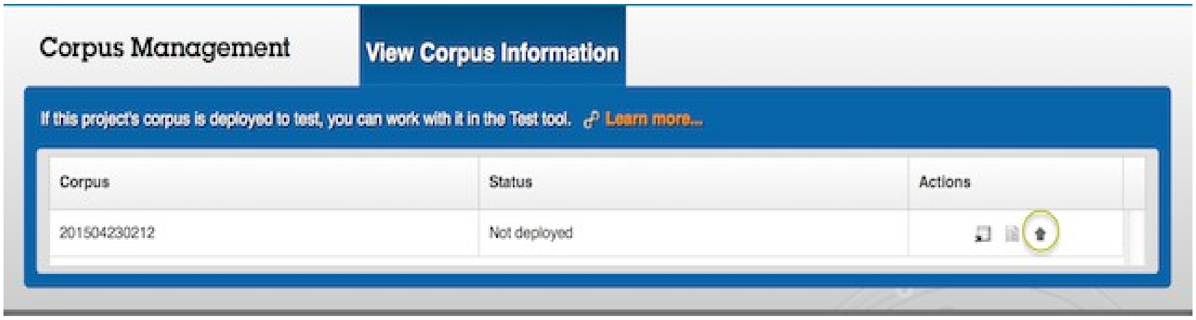
Ingest and deploy the corpus

Given the appropriate role, you are now ready to ingest to the corpus plus the question and answer pairs. This is a very simple step for mankind, but a real complex undertaking for machine. Lucene indexes are created at this point and the pipeline services are started.

1. Open the **Corpus Management** tool.
2. Click Create Corpus.

This may take upwards of 10 minutes for a corpus of 60-100 documents of about 60MB in total corpus size.

3. Once the deployment is complete as indicated by the displayed message, click the **View Corpus Versions** link to view more detail.
4. Click the up arrow to deploy the ground truth and bind it to the pipeline.



If this project's corpus is deployed to test, you can work with it in the Test tool. Learn more...

Corpus	Status	Actions
201504230212	Not deployed	

Test the quality of answers generated by Watson

You are now ready to ask domain relevant questions.

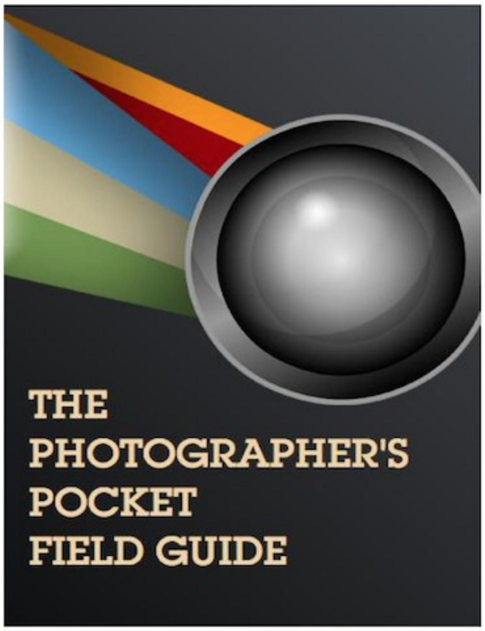
1. Open the Test module from the WEM front page.
2. Type some of the questions that you have seen in the Ground Truth and observe the returned answers.
3. Enter some questions that were not in the ground truth, but you have seen the answer in your ingested documents.

Bear in mind that the system is using default pipeline algorithms and default SPSS models. You would need to run a test set to establish base accuracy, recall and precision metrics. This activity can be done by your Tech Mentor who has Lab_Services role. Test sets often reveal how one can improve on the ground truth for better matched answers. Typically in a commercial BP engagement, after a few test sets and GT corrections, the accuracy may be close to 50% at which time, IBM development would run a training set. Now, the SPSS models have changed to better depict the domain. Please remember that this is just background information, and that running a training set is not done for the academic partners.

Working with the Watson QAAPI

In this section, we describe the three sample reference mobile applications that were designed and built by IBM to help illustrate the process of building a powered-by- Watson application. The reference mobile application, “Watson Photography,” is created for:

≥ Native Android ≥ Native iOS ≥ Hybrid (HTML, JavaScript, CSS, Apache Cordova) using Apache Cordova that can be deployed to both Android and iOS platforms. The reference applications connect over https to Watson using the Question and Answer API (QAAPI).



Watson Photography App

The Photographer’s Pocket Field Guide is a mobile reference application that answers users’ questions related to photography.

Sample code and platform specific documentation can be found on the following public Github repositories:

iOS: https://github.com/IBMMobileCoC/watson-photography-ios

Android: https://github.com/IBMMobileCoC/watson-photography-android

Hybrid: https://github.com/IBMMobileCoC/watson-photography-hybrid

Watson Question Answer API

Once your Watson instance is trained to your application domain, it is ready to be integrated into your application. In this guide, we will focus on using the Question Answer API (QAAPI) to interact with Watson. The Watson QAAPI exposes a REST service interface that allows applications to access Watson's cognitive computing capabilities. You can access your API endpoint at the following url:

https://watson- wdc01.ihost.com/instance/<*nnn*>/deepqa/v1/question

In what follows, we describe how to use the QAAPI to submit questions and retrieve answers. Although the Watson Question and Answer REST service accepts many parameters, only one parameter, questionText, is required to post a question. The following shows a sample JSON defining the question to post:

```
{
“question” : { “questionText” : “What is White Balance?” }
}
```

When you post a question, include an HTTP header. The header must include the content-type. The header must also include a value for X_SyncTimeout since we are using synchronous mode. The X_SyncTimeout value represents how long the server waits after the question is submitted until it times out. The value does not represent how long the client waits for a response from the server. The header should also include the authorization (base64 encoded) required to access your Watson instance.

```
Accept: application/json Content-Type: application/json X-SyncTimeout: 30
Authoritzation: Basic username:password
```

```
// create the question portion

@SuppressWarnings("serial") HashMap<String, Object> question_info = new
HashMap<String, Object>() {{

put("questionText", "What is aperture?"); }};

// given the collection we created above, serialize it into a JSON object

JSONObject question = new JSONObject().put("question",question_info); // construct a
resource using the REST Client. The URL is the URL of the QA API

Resource resource = restClient.resource(QUESTION_POST_URL); ClientResponse response =
null;

// HTTP header that designates the REST service to be synchronous // The service will
wait 30 seconds after receiving the request and giving it // to the Watson pipeline.
After 30 seconds, the service will return with // a timeout in the status field. A -1
indicates to wait indefinitely. resource.header("X-SyncTimeout", "30");

String auth = new String(Base64.encode("username:password".getBytes()));
resource.header("Authorization","Basic " + auth);

// post the question to the Q/A API

//  Accept : application/json

//  Content-Type : application/json

//  HTTP POST

      try{ response = resource.accept(MediaType.APPLICATION_JSON)
```

```
        .contentType(MediaType.APPLICATION_JSON) .post(question.toString());

        // check HTTP response code

        if (response.getStatusType() != javax.ws.rs.core.Response.Status.OK)
        { System.out.println("HTTP Response code returned an error:

        "+response.getStatusCode()); System.exit(1);

        } }

        catch(Exception e){ System.out.println("Unable to connect to server on " +
        QUESTION_POST_URL); System.exit(1);

        }
```

After you submit the question, the response includes an HTTP status code. For the synchronous mode, the successful HTTP status code is 200 Created. The successful response includes the original question and the reported answers sorted in decreasing order of confidence. The answers can be obtained from the evidencelist array where each entry in that array includes the answer text and the confidence of that answer

The following sample Java code illustrates how to obtain the response from WATSON REST service:

```
// marshal the entity response from the POST into a JSON Object that we can process

JSONObject payloadResponse = response.getEntity(JSONObject.class);

// pull the question from the JSON Object

JSONObject questionResponse = payloadResponse.getJSONObject("question");

// We can also see that the question was Accepted if the 'status' is queued. Other
possible values include ERROR

String status = questionResponse.getString("status");

if (status.equals(STATE_COMPLETE)) { // Here we have gotten an answer back and we are
going to display some results System.out.println("------------ Portions of the Answer
Payload shown as an example -------");

//display various Example answer values

System.out.println("status :"+questionResponse.getString("status"));
System.out.println("questionText:"+questionResponse.getString("questionText"));

//loop through the answers displaying some information

System.out.println("answer id :"+answers.getJSONObject(i).getString("id"));
System.out.println(" answer text :"+answers.getJSONObject(i).getString("text"));
System.out.println(" answer confidence

:"+answers.getJSONObject(i).getString("value")); }

} else { // Question failed or timed out System.out.println("Question response
returned status: "+status); System.exit(1);

}
```

Please consult more detailed documentation on the QAAPI and full Java examples in Watson Experience Manager by going to Home Developer Portal and then following the API Reference link: https://watson-wdc01.ihost.com/instance/<nnn>/predeploy/index.jsp?scope=gtthelp&topic=/com.ibm.watson.xmgr.gt.trident.do c/topics/sdk_qaapi_overview.html

Glossary:

Acronym Definition

Corpus - Collection of documents representing the content uploaded to Watson.

QAAPI - Question Answer Application Programming Interface

JSON - JavaScript Object Notation REST: Representational State Transfer

WDC - Watson Developer Cloud

WEM - Watson Experience Manager

GIT - Source code control management system