HANDS ON LAB WITH SHAREPOINT SYNTEX

Creating and Applying a Document Understanding Model with SharePoint Syntex

Description
SharePoint Syntex provides the ability to leverage machine learning and machine teaching technologies to automate the classification of files and the extraction of metadata from files that are stored SharePoint Online libraries. For unstructured documents where the text entities reside in sentences or specific regions of the document (i.e., letters or contracts), SharePoint Syntex allows you to build a model that can designate both the type of file (its *classification*) and extract entities and data from the document (its *extractors*). This set of models is a special content type in SharePoint and can be subsequently published to other libraries in a Microsoft 365 tenant. Like the AI Builder form processing model, the service will write out the extracted file properties to metadata columns in the library at runtime.

Objective: The purpose of this guided lab is to understand:

- How to build and apply a document understanding model, including the classifier and one extractor
- How to create three types of explanations
- How to test your model on similar unlabeled documents to validate your *model*
- How to apply the model to one or more libraries so it can process files at runtime

BACKGROUND: CONTENT CENTER, SHAREPOINT SITE AND TRAINING DOCUMENTS

You are going to need a content center and a document library to publish the completed model. The content center is a new site template in SharePoint Syntex, serving as a location for the creation and management of models, but it will eventually include additional features and capabilities. This site template is not available in self-service and must be created by a SharePoint admin in the admin center. You can find the template in the *Other options* template menu.

For this lab, a ***Content Center*** has already been created and can be accessed at the following URL:

For users 1 to 24:	https://potctraining1.sharepoint.com/sites/ContentCenter/
For users 25 to 48:	https://potctraining2.sharepoint.com/sites/ContentCenter/
For users 49 to 72:	https://potctraining3.sharepoint.com/sites/ContentCenter/
For users 73 to 96:	https://potctraining4.sharepoint.com/sites/ContentCenter/

Later in the lab, we will use the ***SharePoint Syntex Lab*** site and a document library that you will created in it. This site is located at the following URL:

For users 1 to 24:	https://potctraining1.sharepoint.com/sites/SharePointSyntexLab
For users 25 to 48:	https://potctraining2.sharepoint.com/sites/SharePointSyntexLab
For users 49 to 72:	https://potctraining3.sharepoint.com/sites/SharePointSyntexLab
For users 73 to 96:	https://potctraining4.sharepoint.com/sites/SharePointSyntexLab

Creating a model requires a set of example files to train (and test) on. You will need about 5-25 DOCX or PDF files of similar type and layout, but with different values for the information you wish to extract. This set also needs to include 1-3 *negative* examples (files of what the trained file is not; the model uses these negative examples to improve its accuracy). You can apply these exercises to your own files, but for illustrating the concepts and steps in this lab, we will use the Contoso Electronics benefits contracts example set.

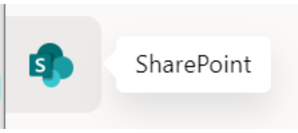
These 14 files (twelve examples and two negative examples) are available in the ***Lab Documents***, which is found in the site referenced above.

STEP 1: CREATE YOUR SHAREPOINT LIBRARY WHERE YOU WILL PUBLISH YOUR MODEL

□ Login to Microsoft 365 by navigating to https://portal.office.com and using the username and password that you claimed. If you are asked "Stay signed in?" select Yes.

- Your username will be something like *username@potctraining1.onmicrosoft.com (use the username you claimed)*
- *If you see a windows that says "Help us protect your account" please click "Skip for now (14 days until required"*

□ Click on the SharePoint icon in the left hand menu.



□ Click the SharePoint Syntex Lab site in the Featured Links section (in the left navigation menu).

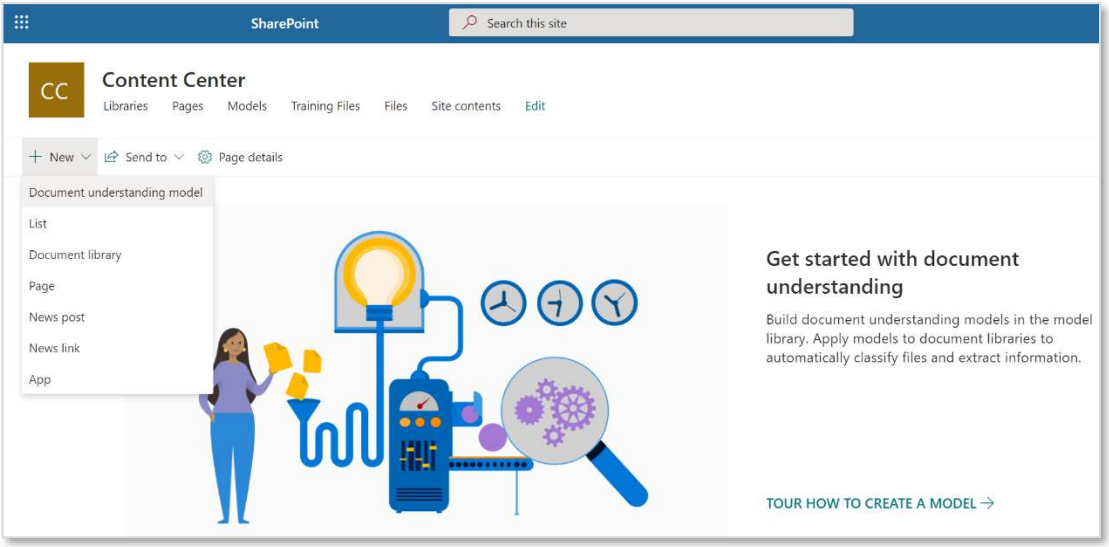
□ Create a new Document Library by clicking the +New menu and selecting *Document library*. Enter your username as your document library name – for example, if your username is user1@potctraining1.onmicrosoft.com then use *user1* as your library name.

STEP 2: BUILD YOUR CLASSIFIER MODEL

□ Navigate to the Content Center site by doing the following:

- Click the waffle icon in the top left of your browser (icon that looks like a waffle) ⋮⋮⋮
- Click the SharePoint icon SharePoint
- Click the Content Center tile

□ From the content center home page, click the +New button in the ribbon, and then click Document understanding model



Get started with document understanding

Build document understanding models in the model library. Apply models to document libraries to automatically classify files and extract information.

□ This will open the "New model" panel where you can associate a model to an existing content type or create a new one. For this exercise we will create a new model named *Benefits Change*.

- Enter your "*username + Benefits Change*" in the Name field. For example, if your username is user1@potctraining1.onmicrosoft.com, then enter "user1 Benefits Change" as your model name.

  - Leave the Advanced Settings with their default settings.
  - Click Create.

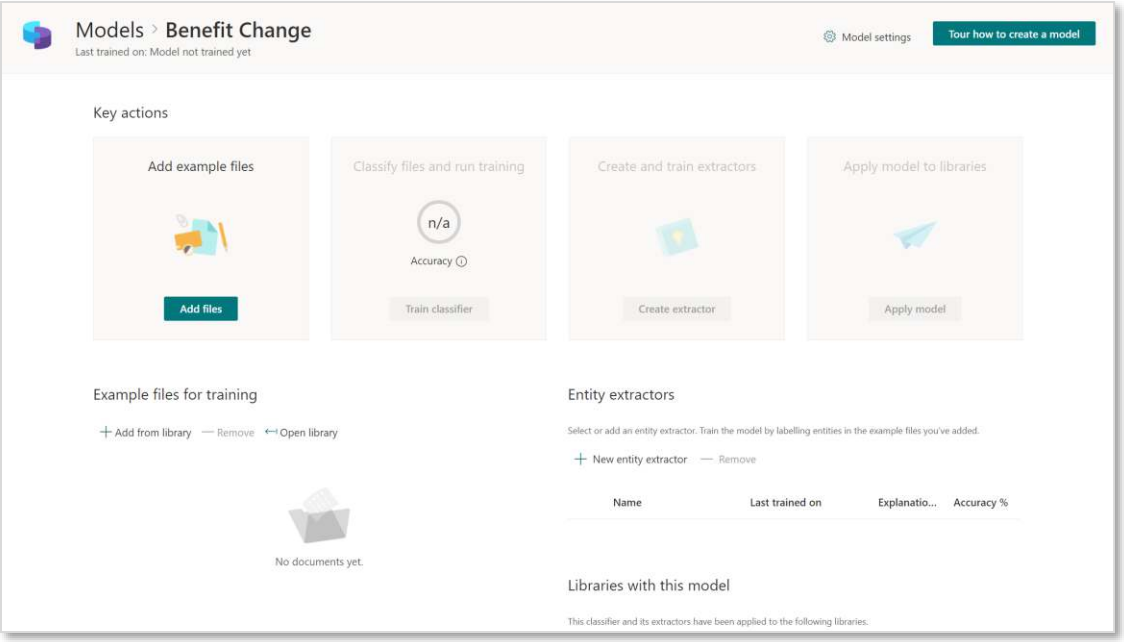
□ The model file will be created in the model library, the content type will be created (currently as a site content type) and you will be redirected to the “model home page”.

Note: you may see a step by step tour window appear before the model home page. Click the X to close the tour.

STEP 3: ADD FILES

Notice that all the action tiles are grayed out except one, *Add files*. Your first step is to select a set of example files. Note: these same files should be used for both the classifier and extractor training. You always have the option to add more, but typically you can add a full set of example files, label them to train your model, and test the remaining unlabeled files to validate your model.

□ Click the *Add files* button and Navigate to the *Example Training Files* folder



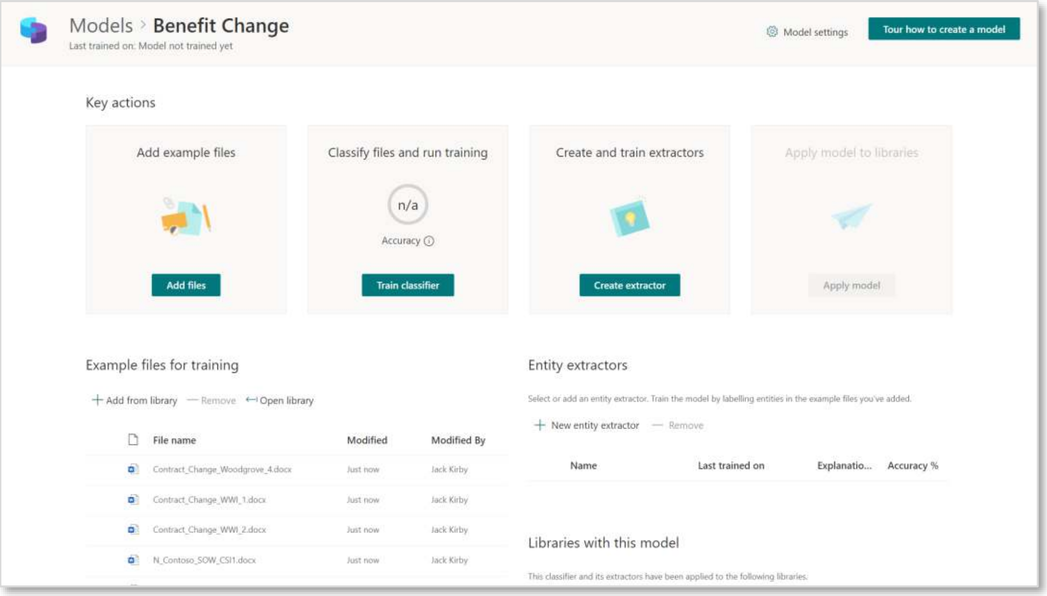
Select or add an entity extractor. Train the model by labelling entities in the example files you've added.

This classifier and its extractors have been applied to the following libraries.

□ Select all 6 files and click Add to add them to your model. They will appear in the *Example files for training* section.

STEP 4: CREATE INSURANCE PROVIDER ENTITY EXTRACTOR & LABEL EXAMPLES

Notice that once the example files are added the *Create new entity extractor* and *Label examples* tiles are enabled. You can start by either training your classifier (*Label examples*) or creating and training an extractor (*Create a new entity extractor*). Extractors are used by Syntex as explanations to understand the document classification, and since one explanation is needed for each classifier or extractor you train, it is a helpful strategy to start with an extractor.



Models > Benefit Change

Last trained on: Model not trained yet

Model settings
Tour how to create a model

Key actions

Add example files	Classify files and run training	Create and train extractors	Apply model to libraries
Add files	n/a Accuracy Train classifier	Create extractor	Apply model

Example files for training

+ Add from library — Remove ⟻ Open library

File name	Modified	Modified By
Contract_Change_Woodgrove_4.docx	Just now	Jack Kirby
Contract_Change_WWI_1.docx	Just now	Jack Kirby
Contract_Change_WWI_2.docx	Just now	Jack Kirby
N_Contoso_SOW_CSI1.docx	Just now	Jack Kirby

Entity extractors

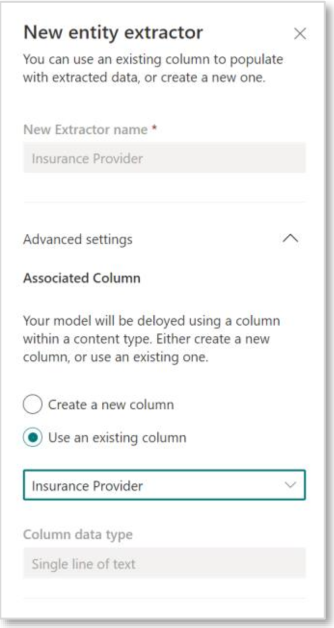
Select or add an entity extractor. Train the model by labelling entities in the example files you've added.

Name	Last trained on	Explanatio...	Accuracy %

Libraries with this model

This classifier and its extractors have been applied to the following libraries.

- ☐ Click the *Create extractor* button in the *Create and Train Extractors* tile.
    - ○ Click the Advanced Settings dropdown
    - ○ Select the radio button *Use an existing column*
    - ○ Select the column *Insurance Provider*
    - ○ The Advanced Settings will remain with their default settings.
    - ○ Click Create



New entity extractor

You can use an existing column to populate with extracted data, or create a new one.

New Extractor name *

Insurance Provider

Advanced settings

Associated Column

Your model will be deloyed using a column within a content type. Either create a new column, or use an existing one.

○ Create a new column

● Use an existing column

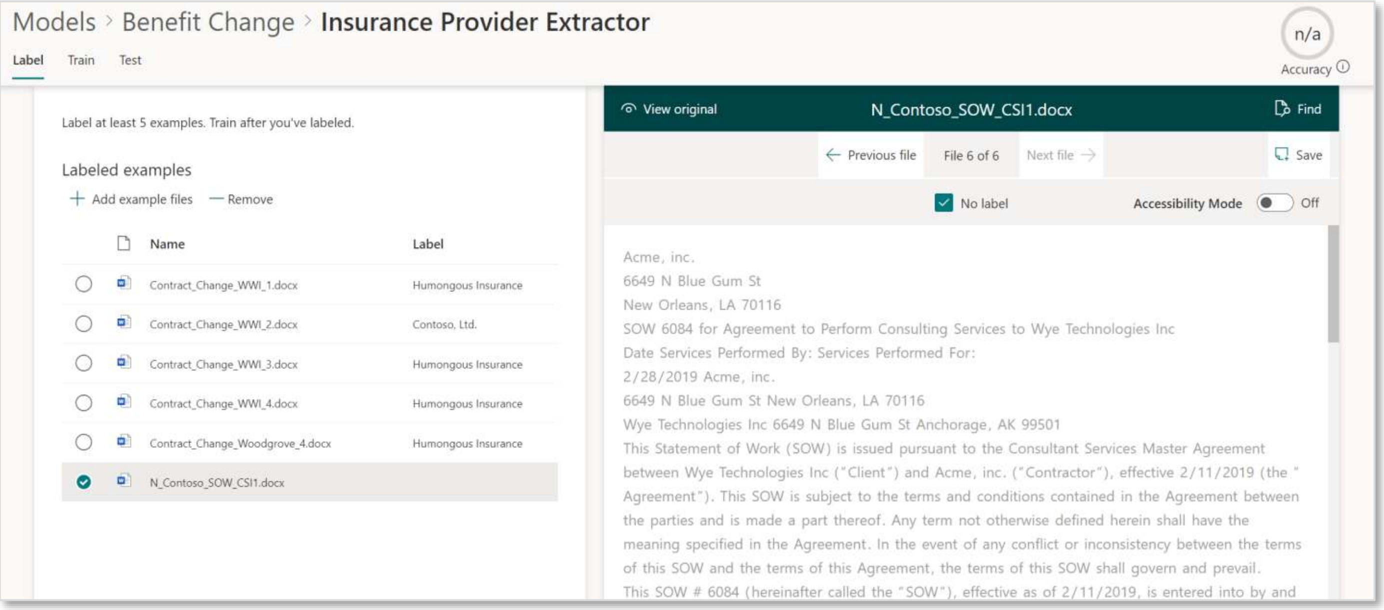
Insurance Provider

Column data type

Single line of text

- ☐ The first document in your sample set will be displayed in the viewer in plain text view; the full list of added example files is displayed in the labeled examples section. NOTE: you can render the formatted view of a document by selecting *View original* link. Label the insurance provider name Humongous Insurance by highlighting it. If you select too little or too much text, just select the highlighted selection to clear it and try again. Alternatively, the text may say Contoso Ltd.
- ☐ Click *Next file* when the string is correctly highlighted to submit the labeled file and advance to the next one.
- ☐ Once you advance or click to the next file, it will be displayed in the viewer and the label you applied to the previous file will be displayed in the "Labeled examples" section.
- ☐ Label the next 4 files for a total of 5.

□ The 6th file is a negative sample and does not contain an insurance provider. Click the "No label found" option in the title and click *Save*.



Label at least 5 examples. Train after you've labeled.

Labeled examples

	Name	Label
○	Contract_Change_WWI_1.docx	Humongous Insurance
○	Contract_Change_WWI_2.docx	Contoso, Ltd.
○	Contract_Change_WWI_3.docx	Humongous Insurance
○	Contract_Change_WWI_4.docx	Humongous Insurance
○	Contract_Change_Woodgrove_4.docx	Humongous Insurance
✓	N_Contoso_SOW_CSI1.docx	

N_Contoso_SOW_CSI1.docx

☑ No label

Acme, inc.
6649 N Blue Gum St
New Orleans, LA 70116
SOW 6084 for Agreement to Perform Consulting Services to Wye Technologies Inc
Date Services Performed By: Services Performed For:
2/28/2019 Acme, inc.
6649 N Blue Gum St New Orleans, LA 70116
Wye Technologies Inc 6649 N Blue Gum St Anchorage, AK 99501
This Statement of Work (SOW) is issued pursuant to the Consultant Services Master Agreement between Wye Technologies Inc ("Client") and Acme, inc. ("Contractor"), effective 2/11/2019 (the " Agreement"). This SOW is subject to the terms and conditions contained in the Agreement between the parties and is made a part thereof. Any term not otherwise defined herein shall have the meaning specified in the Agreement. In the event of any conflict or inconsistency between the terms of this SOW and the terms of this Agreement, the terms of this SOW shall govern and prevail. This SOW # 6084 (hereinafter called the "SOW"), effective as of 2/11/2019, is entered into by and

□ Once you have labeled five files you should see a notification banner displayed near the top of the screen, informing you to move to training. Even if the banner does not appear at the top, please move onto the next step.

STEP 5: TRAIN INSURANCE PROVIDER ENTITY EXTRACTOR

□ Click the Train tab in the top left of the Model page. The training phase is where you add explanations to help the model understand how to classify your document or how to identify the information you want to extract.

□ The first labeled file will be displayed and you will be prompted to add an explanation. At least one is required before the model can be trained. NOTE: only the example files that you label in the previous phase are displayed, with a status of *not trained*. For an entity extractor, explanations can be hints about the label format itself or strings around the label to help identify it. To identify the *Insurance Provider* we will create two phrase list explanations, one before the label and one after.

□ Click *+New* from the Explanations pane and select *Blank* from the New menu.

- From the *Type* dropdown select *Phrase list*.
- Enter "String before" for the Explanation name. In the Phrase list text box enter the preceding part of the sentence "network participation with your insurance provider". The case sensitivity should be set to off.
- Click Save.

□ Once this first explanation is saved, we will automatically train the model. The training results will be displayed in the "Trained files" section, showing both the predicted label value and the evaluation. A match means the model is correctly predicting your label, a mismatch means it's not correctly predicting the label for that file. You can look in the "Prediction" column or load the file into the viewer to see the mismatch.

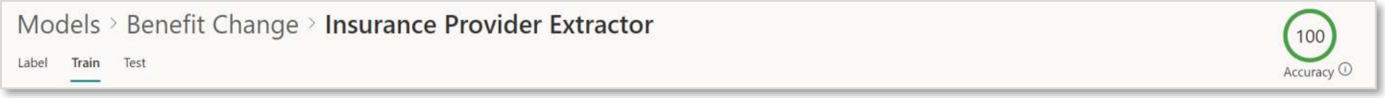
□ In our case, notice that one of our files is predicting a mismatch.

□ To correct, let's add another explanation.

- Click *+New* from the Explanations pane and select *Blank* from the New menu.
- From the *Type* dropdown select *Phrase list*.
- Enter "String after" for the Explanation name. In the Phrase list text box enter the preceding part of the sentence "(http://www.". The case sensitivity should be set to off.
- Click Save.

- ☐ We will initiate retraining once the new explanation is saved. Notice that now the mismatch is fixed for that file. The additional explanation helped the model to understand the variety of insurance provider names and correctly predicted this longer name.

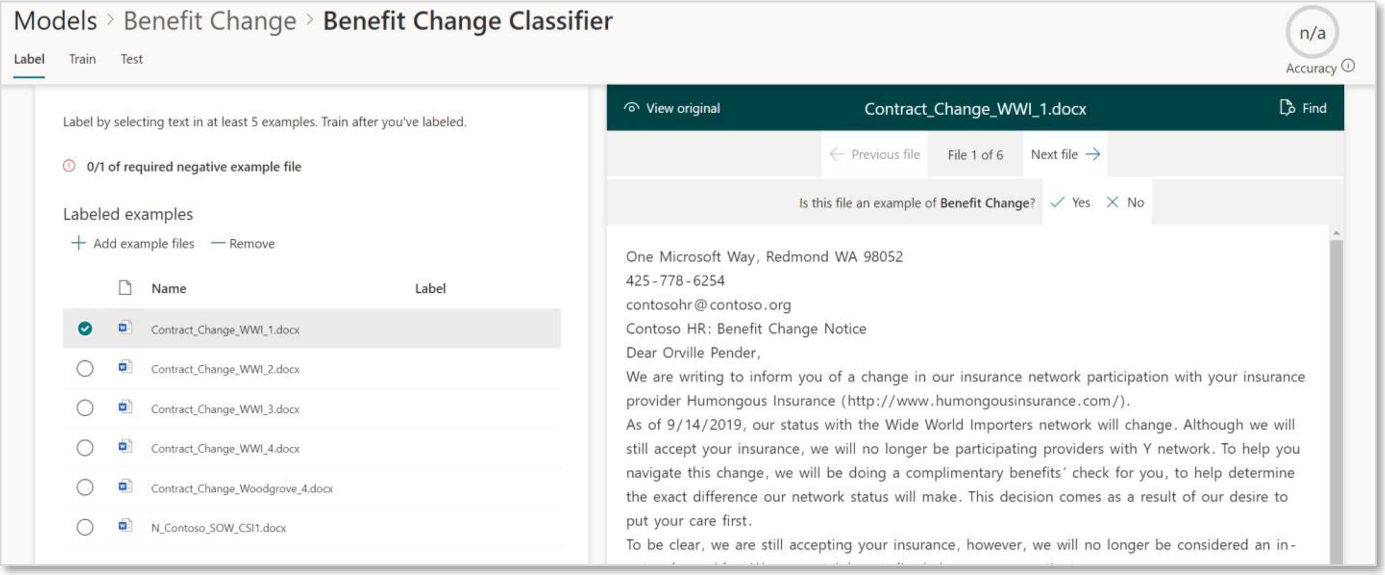
- ☐ Now that your *Insurance Provider* entity extractor is trained and accurately predicting labels you can exit this training and return to the model home page.
  - o At the top of the page, where it says **Model > Benefits Change > <username> Insurance Provider Name**, click *Benefits Change* to return to the main model page.



STEP 6: TRAIN BENEFITS CHANGE CLASSIFIER

In the "Classify files and run training" tile of the model page, notice that your classifier has no accuracy score (in the middle of the second tile). This is because it has not been trained yet.

- ☐ Click the *Train Classifie*r button.
- ☐ Like the entity extractor experience, the first labeled document in your example set will be displayed in the viewer; the list of files and their labels are displayed in the "Labeled examples" section on the left.



Label by selecting text in at least 5 examples. Train after you've labeled.

ⓘ 0/1 of required negative example file

Labeled examples

	Name	Label
●	Contract_Change_WWI_1.docx	
○	Contract_Change_WWI_2.docx	
○	Contract_Change_WWI_3.docx	
○	Contract_Change_WWI_4.docx	
○	Contract_Change_Woodgrove_4.docx	
○	N_Contoso_SOW_CSI1.docx	

Is this file an example of **Benefit Change**?

One Microsoft Way, Redmond WA 98052
425-778-6254
contosohr@contoso.org
Contoso HR: Benefit Change Notice
Dear Orville Pender,
We are writing to inform you of a change in our insurance network participation with your insurance provider Humongous Insurance (http://www.humongousinsurance.com/).
As of 9/14/2019, our status with the Wide World Importers network will change. Although we will still accept your insurance, we will no longer be participating providers with Y network. To help you navigate this change, we will be doing a complimentary benefits' check for you, to help determine the exact difference our network status will make. This decision comes as a result of our desire to put your care first.
To be clear, we are still accepting your insurance, however, we will no longer be considered an in-

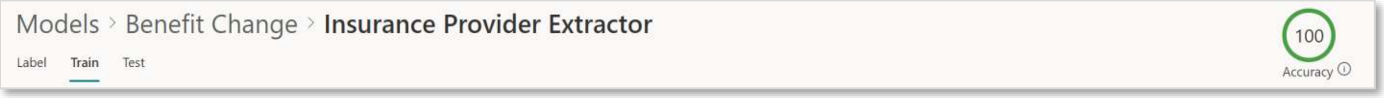
In this training you are effectively labeling the *entire* document as either a positive or negative example of the content type.

- ☐ Answer the question in the viewer ribbon "Yes" or "No" for that displayed document. Making a selection will submit the answer, label the document as "Positive" or "Negative" and advance the viewer to the next example file.
  NOTE: you must submit at least one negative example before you can proceed to the next step. Once you have labeled five files a notification banner will display informing you to move to training. You can label more documents or advance to training.

- ☐ Click the Train tab in the top left of the model page to advance to training.
- ☐ The service will train on the labeled set immediately. You do not need to provide an additional explanation because you already provided explanations when you trained the *Insurance Provider* entity extractor.

- □ You will now see that all of your trained and labeled files show a status of *Match*. You can select any one to see the viewed file highlighted in green. If there was a mismatch, the file would show in red and you could remediate by either labeling more files or adding more explanations.

- □ Now that you have trained your model to classify files you can exit this training and return to the model home page.
  - ○ At the top of the page, where it says **Model > Benefits Change > <username> Insurance Provider Name**, click *Benefits Change* to return to the main model page.



STEP 7: APPLY MODEL TO A LIBRARY

You are now ready to apply your model to a document library.

- □ On the model home page click "Apply Model" from the tile.
- □ This will display a panel where you can navigate to or search for a site you have access to. Select the *SharePoint Syntex Lab* site.
- □ Select the library that you created in Step 1 of this lab and click "Add". Remember, if your username was user1@potctraining1.onmicrosoft.com then you created a library called *user 1*. The library location of the applied model will now appear in the "Libraries with this model" list.

  Now that you have published your model, we will validate it by classifying and extracting metadata from documents in your library.

- □ Open a new browser tab and navigate to SharePoint Syntex Lab site.
- □ Navigate to the *Lab Documents* library, and then to a folder called *Document Understanding Sample Files*.
- □ Select all the documents in this folder and click *Copy To* in the ribbon bar.
- □ As a destination, select the SharePoint Syntex Lab site and then select the library that you created.
- □ Click *Copy Here*.

- □ Navigate to your library. Select the new view that was created, which should be named after your model "*username + Benefits Change*".

  After a few minutes you should start to see the content type for your documents get set to "*username + Benefits Change*" and you should see the Insurance Provider metadata column populated with the insurance provider from inside the documents.

  Note: the negative examples which you copied into this library will not have the content type set, nor the Insurance Provider metadata column set.

Congratulations! You have used SharePoint Syntex to create a document understanding model, published it to a library and used it to classify files and extract metadata.

FEEDBACK

How did it go? Please record any issues, questions or feedback that arise as you work through the lab.