Named Entity Recognition

01/23/2018 5 minutes to read Contributors

In this article

Recognizes named entities in a text column

Category: Text Analytics

Module overview

This article describes how to use the **Named Entity Recognition** module in Azure Machine Learning Studio, to identify the names of things, such as people, companies, or locations in a column of text.

Named entity recognition is an important area of research in machine learning and natural language processing (NLP), because it can be used to answer many real-world questions, such as:

- Does a tweet contain the name of a person? Does the tweet also provide his current location?
- Which companies were mentioned in a news article?
- Were specified products mentioned in complaints or reviews?

To get a list of named entities, you provide a dataset as input that contains a text column. The **Named Entity Recognition** module will then identify three types of entities: people (PER), locations (LOC), and organizations (ORG).

The module also labels the sequences by where these words were found, so that you can use the terms in further analysis.

For example, the following table shows a simple input sentence, and the terms and values generated by the module:

Input text	Module output

Input text	Module output
“Boston is a great place to live.”	0,Boston,0,6,LOC

The output can be interpreted as follows:

- The first ‘0’ means that this string is the first article input to the module.

  Because a single article can have multiple entities, including the article row number in the output is important for mapping features to articles.

- `Boston` is the recognized entity.
- The `0` that follows `Boston` means the entity `Boston` starts from the first letter of the input string. Indices are zero-based.
- `6` means the length of the entity `Boston` is 6.
- `LOC` means the entity `Boston` is a place, or location. Other supported named entity types are person (`PER`) and organization (`ORG`).

How to configure Named Entity Recognition

1. Add the **Named Entity Recognition** module to your experiment in Studio. You can find the module in the **Text Analytics** category.

2. On the input named **Story**, connect a dataset containing the text to analyze.

   The "story" should contain the text from which to extract named entities.

   The column used as **Story** should contain multiple rows, where each row consists of a string. the string can be short, like a sentence, or long, like a news article.

   You can connect any dataset that contains a text column. However, if the input dataset contains multiple columns, use Select Columns in Dataset to choose only the column that contains the text you want to analyze

   > ⓘ Note
   >
   > The second input, **Custom Resources (Zip)**, is not supported at this time.

   In future, you can add custom resource files here, for identifying different entity types.

3. Run the experiment.

Results

The module outputs a dataset containing a row for each entity that was recognized, together with the offsets.

Because each row of input text might contain multiple named entities, an article ID number is automatically generated and included in the output, to identify the input row that contained the named entity. The article ID is based on the natural order of the rows in the input dataset.

You can convert this output dataset to CSV for download or save it as a dataset for re-use.

Use named entity recognition in a web service

If you publish a web service from Azure Machine Learning Studio and want to consume the web service by using C#, Python, or another language such as R, you must first implement the service code provided on the help page of the web service.

If your web service provides multiple rows of output, the URL of the web service that you add to your C#, Python, or R code should have the suffix `scoremultirow` instead of `score`.

For example, assume you use the following URL for your web service:

```
https://ussouthcentral.services.azureml.net/workspaces/<workspace id>/services/<service id>/score
```

To enable multi-row output, change the URL to

```
https://ussouthcentral.services.azureml.net/workspaces/<workspace id>/services/<service id>/scoremultirow
```

To publish this web service, you should add an additional Execute R Script module after the Named Entity Recognition module, to transform the multi-row output into a single delimited with semi-colons (;). The reason for consolidating the multiple rows of output into a single row is to return multiple entities per input row.

For example, let’s assume you have an input sentence with two named entities. Rather than returning two rows for each row of input, you can return a single rows with multiple entities, separated by semi-colons as shown here:

Input Text	Output of Web Service
Microsoft has two office locations in Boston.	0,Microsoft,0,9,ORG,;,0,Boston,38,6,LOC,;

The following code sample demonstrates how to do this:

text Copy

```
# Map 1-based optional input ports to variables
d <- maml.mapInputPort(1) # class: data.frame
y=length(d) ##size of cols
x=dim(d)[1] ##size of rows
longd=matrix("NA",nrow=1,ncol=x*(y+1))
for (i in 1:x)
  {
     for (j in 1:y)
     {
       longd[1,j+(i-1)*(y+1)]=toString(d[i,j])
     }
     longd[1,j+(i-1)*(y+1)+1]=c(";")
  }

final_output=as.data.frame(longd)
# Select data.frame to be sent to the output Dataset port
maml.mapOutputPort("final_output");
```

Examples

This blog provides an extended explanation of how named entity recognition works, its background, and possible applications:

- Machine learning and text analytics

Also, see the following sample experiments in the Azure AI Gallery for demonstrations of how to use text classification methods commonly used in machine learning:

- News Categorization sample: Uses feature hashing to classify articles into a predefined list of categories.
- Similar Companies sample: Uses the text of Wikipedia articles to categorize companies.
- Text-Classification Step 1 of 5: Data preparation: In this five-part walkthrough of text classification, text from Twitter messages is used to perform sentiment analysis. A variety of text pre-processing techniques are also demonstrated.

Technical notes

Language support

Currently, the **Named Entity Recognition** module supports only English text. It can detect organization names, personal names, and locations in English sentences. If you use the module on other languages, you might not get an error, but the results are not as good as for English text.

In future, support for additional languages can be enabled by integrating the multilingual components provided in the Office Natural Language Toolkit.

Expected inputs

Name	Type	Description
Story	Data Table	An input dataset (DataTable) that contains the text column you want to analyze.
CustomResources	Zip	(Optional) A file in ZIP format that contains additional custom resources. This option is not available currently and is provided for forward compatibility only.

Outputs

Name	Type	Description
Entities	Data Table	A list of character offsets and entities

See also

Text Analytics
Feature Hashing
Score Vowpal Wabbit 7-4 Model
Train Vowpal Wabbit 7-4 Model

ⓘ **Note**

The feedback system for this content will be changing soon. Old comments will not be carried over. If content within a comment thread is important to you, please save a copy. For more information on the upcoming change, we invite you to read our blog post.