



First Look: Text Analytics with InterSystems Products

Version 2019.4
2020-01-28

First Look: Text Analytics with InterSystems Products
InterSystems IRIS Data Platform Version 2019.4 2020-01-28
Copyright © 2020 InterSystems Corporation
All rights reserved.

InterSystems, InterSystems IRIS, InterSystems Caché, InterSystems Ensemble, and InterSystems HealthShare are registered trademarks of InterSystems Corporation.

All other brand or product names used herein are trademarks or registered trademarks of their respective companies or organizations.

This document contains trade secret and confidential information which is the property of InterSystems Corporation, One Memorial Drive, Cambridge, MA 02142, or its affiliates, and is furnished for the sole purpose of the operation and maintenance of the products of InterSystems Corporation. No part of this publication is to be used for any other purpose, and this publication is not to be reproduced, copied, disclosed, transmitted, stored in a retrieval system or translated into any human or computer language, in any form, by any means, in whole or in part, without the express prior written consent of InterSystems Corporation.

The copying, use and disposition of this document and the software programs described herein is prohibited except to the limited extent set forth in the standard software license agreement(s) of InterSystems Corporation covering such programs and related documentation. InterSystems Corporation makes no representations and warranties concerning such software programs other than those set forth in such standard software license agreement(s). In addition, the liability of InterSystems Corporation for any losses or damages relating to or arising out of the use of such software programs is limited in the manner set forth in such standard software license agreement(s).

THE FOREGOING IS A GENERAL SUMMARY OF THE RESTRICTIONS AND LIMITATIONS IMPOSED BY INTERSYSTEMS CORPORATION ON THE USE OF, AND LIABILITY ARISING FROM, ITS COMPUTER SOFTWARE. FOR COMPLETE INFORMATION REFERENCE SHOULD BE MADE TO THE STANDARD SOFTWARE LICENSE AGREEMENT(S) OF INTERSYSTEMS CORPORATION, COPIES OF WHICH WILL BE MADE AVAILABLE UPON REQUEST.

InterSystems Corporation disclaims responsibility for errors which may appear in this document, and it reserves the right, in its sole discretion and without notice, to make substitutions and modifications in the products and practices described in this document.

For Support questions about any InterSystems products, contact:

InterSystems Worldwide Response Center (WRC)
Tel: +1-617-621-0700
Tel: +44 (0) 844 854 2917
Email: support@InterSystems.com

Table of Contents

First Look: Text Analytics with InterSystems Products.....	1
1 Why NLP Text Analytics Is Important	1
2 How InterSystems IRIS Implements NLP Text Analytics	1
3 Trying NLP Text Analytics for Yourself	2
3.1 Before You Begin	2
3.2 Create a Domain and Add Data Locations	3
3.3 Explore the Data	3
4 Learn More About NLP Text Analytics	5

First Look: Text Analytics with InterSystems Products

This First Look guide introduces you to InterSystems IRIS® data platform support for Natural Language Processing (NLP) text analytics, which provides semantic analysis of unstructured text data in a variety of natural languages. This enables you to discover useful information about the contents of a large number of text documents without any prior knowledge of the contents of the texts.

This First Look guide presents an introduction to InterSystems IRIS Natural Language Processing, and walks through some initial tasks associated with indexing text data for semantic text analysis. Once you've completed this exploration, you will have indexed a group of texts and performed analysis determining the most common entities in those texts, metrics about those entities, various kinds of associations found between entities, and viewing the appearances of an entity in the source texts. These activities are designed to use only the default settings and features, so that you can acquaint yourself with the fundamentals of NLP text analysis. For the full documentation on Text Analytics, see the [InterSystems IRIS Natural Language Processing \(NLP\) Guide](#).

A related, but separate, tool for handling unstructured texts is [InterSystems IRIS SQL Search](#). SQL Search allows you to *search* for these same entities, as well as single words, regular expressions and other constructs in multiple texts. Inherently, a search solution presupposes that you know what you are looking for. NLP text analytics is designed to help you discover content and connections between content entities without necessarily starting from an idea to look for.

To browse all of the First Looks, including those that can be performed on a [free evaluation instance of InterSystems IRIS](#), see [InterSystems First Looks](#).

1 Why NLP Text Analytics Is Important

Increasingly, organizations are amassing larger and larger quantities of unstructured text data, far in excess of their ability to read or catalog these texts. Frequently, an organization may have little or no idea what the contents of these text documents are. Conventional “top-down” text analysis based on pure *search* technologies makes assumptions about the contents of these texts, which may miss important content.

InterSystems IRIS Natural Language Processing (NLP) allows you to perform text analysis on these texts without any upfront knowledge of the subject matter. It does this by applying language-specific rules that identify semantic entities. Because these rules are specific to the language, not the content, NLP can provide insight into the contents of texts without the use of a dictionary or ontology.

2 How InterSystems IRIS Implements NLP Text Analytics

To prepare texts for NLP analytics you must load those texts into a domain, and then build the domain. Based on its analysis of the texts, NLP builds indices for the domain that NLP can use to rapidly analyze large quantities of text. Texts can be input from a variety of data locations, including SQL tables, text files, strings, globals, and RSS data.

NLP supports the following functionality:

- Language models: identifying semantic relationships between words is language-specific. NLP contains semantic rules (language models) for ten natural languages that enable analysis of a text on any subject written in that language. If you specify more than one language, NLP performs automatic language identification by determining the best match between each sentence in each text and the specified languages. NLP analysis does not require the upfront creation or association of dictionaries or ontologies, although you can expand its functionality by adding them.
- Entity analysis: NLP operates on semantic groups of one or more words known as entities. Entities are identified as either Concepts (which include nouns and noun phrases) or Relations (which include verbs and prepositions). Commonly, the most relevant entities to consider are Concepts, though it is also possible to analyze Relations. Sentence and word boundaries are always observed. Letter case is ignored.
- Path analysis: NLP groups coherent sequences of Concepts and Relations into Paths. A sentence usually consists of a single Path. A Path reveals the connections between entities.
- Attributes: NLP flags semantic attributes such as negation, so that you can differentiate text sequences that are positive (“evidence of structural damage”) from those that are negative (“no evidence of structural damage”).
- Frequency, Spread, and Dominance: These are metrics calculated for an entity. Frequency is the number of times an entity appears in a group of texts. Spread is the number of texts that contain that entity. Dominance is a more nuanced metric generated by factoring in the entity frequency relative to the length of each text, the frequency of other entities having words in common, and other factors. Entities are commonly returned sorted in descending order by these metrics. These metrics provide insight into the content of texts, enabling you to perform deeper analysis of specific entities.
- Similar Entities, Related Concepts, and Proximity Profile. Given an entity, these discover other relevant entities. For example, given a short entity, similar entities would include other, longer entities in the domain that contain the same words, thereby being more specific entities than the seed one. Given an entity, related entities are other entities in the same sentence that are associated to the specified entity by a single Relation. Given an entity, the Proximity metric calculates the proximity within paths between the specified entity and other entities.
- Dictionaries: you can add optional dictionaries to identify synonyms for an entity.
- Summarization: you can use NLP to generate a summary of a text, requesting the summary as a percentage of the whole. For example, a 50% summary would consist of half of the sentences in the original text, with NLP selecting those sentences that are calculated as most relevant to the overall source text.

3 Trying NLP Text Analytics for Yourself

It is easy to use InterSystems IRIS Text Analytics. This simple procedure walks you through the basic steps of generating NLP metrics.

This example is provided to give you some initial experience with InterSystems IRIS Natural Language Processing. You should not use this example as the basis for developing a real application. To use NLP in a real situation you should fully research the available choices provided by the software, then develop your application to create robust and efficient code.

3.1 Before You Begin

To use the procedure, you will need a running InterSystems IRIS instance. Your choices for InterSystems IRIS include several types of licensed and free evaluation instances; the instance need not be hosted by the system you are working on (although they must have network access to each other). For information on how to deploy each type of instance if you do not already have one to work with, see [Deploying InterSystems IRIS](#) in *InterSystems IRIS Basics: Connecting an IDE*.

You also need to obtain the Aviation.Event SQL table, which is available on GitHub at <https://github.com/intersystems/Samples-Aviation>. Follow the instructions provided in [Downloading and Setting up the Sample Files](#) in *First Look: SQL Search with InterSystems Products* to download and set up the files.

3.2 Create a Domain and Add Data Locations

All NLP analysis occurs within a domain. You associate multiple texts with a domain. You then build the domain, creating indices that are used by NLP queries.

A domain is created within a namespace, such as the **SAMPLES** namespace you created by following the procedure in *First Look: SQL Search with InterSystems Products* in the previous section. You can create multiple domains within a namespace. You can associate a text with multiple domains.

There are several ways to create, populate, and build a domain. The following example uses the **Domain Architect** interface.

1. Open the Management Portal for your instance in your browser, using the [URL described for your instance](#) in *InterSystems IRIS Basics: Connecting an IDE*.
2. Navigate to the Domain Architect page (**Analytics > Text Analytics > Domain Architect**). Before using the **Analytics** options, you may need to switch to the analytics-enabled **SAMPLES** namespace.
3. Click the **New** button to define a domain. Specify the following domain values (in the order given):
 - **Domain name:** The name you assign to a domain must be unique for the current namespace (not just unique within its package class); domain names are not case-sensitive. For this example, specify **MyTest**.
 - **Definition class name:** the domain definition package name and class name, separated by a period. From the **Domain name** field press the Tab key to generate a default **Definition class name:** **Samples .MyTest**.
 - **Allow Custom Updates:** this check box enables adding data or dictionaries to this domain manually. For this example, do not check the box.

Click the **Finish** button to create the domain. This displays the **Model Elements** selection screen.

4. Within a domain you can define data locations and other model elements for the domain. To add or modify model elements, click on the expansion triangle next to one of the headings. Initially, no expansion occurs. Once you have defined some model elements, clicking the expansion triangle shows the model elements you have defined.

Click the **Data Locations** triangle to display the Details tab on the right side of the screen. The Details tab shows five Add Data options. Select **Add data from table**.

This option allows you to specify data stored in an SQL table. In this example we will specify the following fields:

- **Name:** A name for the set of extracted data files. Use the default: **Table_1**.
- **Schema:** From the drop-down list select **Aviation**.
- **Table Name:** From the drop-down list select **Event**.
- **ID Field:** From the drop-down list select **ID**.
- **Data Field:** From the drop-down list select **NarrativeFull**.

The **Domain Architect** page heading is followed by an asterisk (*) if there are unsaved changes to the current domain definition. Click **Save** to save your changes.

5. Compile the Domain by pressing the **Compile** button.
6. Build the NLP indices for the data sources by pressing the **Build** button.

3.3 Explore the Data

Explore the data using the procedure that follows:

1. On the Domain Architect page, select the **Tools** tab on the right side of the screen, then click the **Domain Explorer** button.
2. The Domain Explorer initially displays a list of the most significant concepts in the source texts:
 - The **frequency** tab displays the **Top Concepts** in descending order by frequency. Each listed item is shown with its frequency count (number of occurrences) and spread count (number of sources containing that concept).
For example, the concept `pilot` has a frequency of 6206 and a spread of 1085; the concept `student pilot` has a frequency of 319 and a spread of 141.
 - The **dominance** tab displays the **Dominant Concepts** in descending order by dominance calculation.
For example, the concept `pilot` has a dominance of 351.6008; the concept `student pilot` has a dominance of 49.3625.
3. When you select one of these concepts the other **Domain Explorer** listings are displayed:
 - **Similar Entities** lists the selected concept, and any other concepts that contain that concept, each with its frequency and spread.
For example, selecting `student pilot` lists the Similar Entities including `student pilot`, `student pilot certificate`, `student pilot's logbook`, `solo student pilot`.
 - **Related Concepts** lists other concepts that are related to the selected concept, with the frequency and spread of the instances of those concepts in this related context.
For example, selecting `student pilot` lists Related Concepts including `flight instructor` and `airplane`.
 - **Proximity Profile** lists other concepts found in proximity to the selected concept, with calculated proximity score of the instances of those concepts when found in the same sentence as the selected concept.
For example, selecting `student pilot` lists a Proximity Profile including `airplane` with a proximity of 2702, and `flight instructor` with a proximity of 1662.

By selecting a concept in any of these lists, these listings are refreshed based on that concept. Alternatively, you can also type an entity (Concept or Relation) into the **Domain Explorer Explore** area and click the **Explore!** button.

By using these listings, you can determine what concepts appear in the source documents, how significant they are, and what other concepts are associated with them.

The lower portion of the **Domain Explorer** allows you to view how a selected concept appears in the source texts:

- The **Sources** tab lists by source all of the sentences that contain the selected concept. The concept is highlighted and red text is used to indicate negation that involves the concept.
- The **Paths** tab lists all of the paths that contain the selected concept. The path text is highlighted to show NLP indexing: the selected concept is highlighted in orange, other concepts in the path are highlighted in blue, path-relevant concepts (commonly pronouns) are highlighted in light blue. Relations are shown white. Red text is used to indicate negation that involves the concept.

By clicking the **eye** icon, you can display the complete text of the source, with the selected concept highlighted, and red text used to indicate negation.

- The **indexing** toggle button displays the complete text of the source with highlighting showing NLP indexing. Because this is the source text, capitalization, punctuation, and non-relevant words are shown; these aspects of the text are not shown in the **Paths** listing.
- The **%** option allows you to display a summary of the text. Specify a percentage. The total number of sentences in the text is reduced to that percentage. The sentences that NLP includes in the summary are determined by metrics calculating their significance to the full text.

4. You can add a blacklist to exclude undesired concepts. Often the list of top concepts begins with those that are too common or have little value in discovering useful information. These may be words or phrases that appear in all of the sources (such as "accident report" or "conclusions"), general concepts (such as "airplane" or "pilot"), or concepts not relevant to your use of the data (such as a list of cities). You can use a blacklist to prevent the display of these concepts. A blacklist affects only the display of concepts in certain query results; it has no effect on NLP indexing of concepts.
 - a. In the Domain Architect click the **Open** button and select Samples >> then MyTest to open the existing domain Samples.MyTest.
 - b. Click the **Blacklists** expansion triangle. This displays the **Add blacklist** button in the Details tab on the right side of the screen. Click **Add blacklist** to display the Name and Entries fields. Accept the default name for the blacklist (Blacklist_1). In the **Entries** box list entries (concepts) one concept per line; entries are not case-sensitive. In this example list the concepts: pilot, student pilot, co-pilot, passenger, instructor, flight instructor, certified flight instructor.
 - c. **Save** and **Compile** the domain. (You do not need to **Build** the domain to add, modify, or remove blacklists).
 - d. In the Domain Explorer click the **sunglasses** icon in the upper right corner. This displays a list of the blacklists defined for this domain that you can apply. Select Blacklist_1. Note that the **Top Concepts** listing no longer lists the blacklist concepts.

4 Learn More About NLP Text Analytics

InterSystems has other resources to help you learn more about NLP Text Analytics, including:

- [InterSystems IRIS Natural Language Processing \(NLP\) Guide](#)

