文章
Louis Lu · 十一月 2, 2021 阅读大约需 11 分钟

IRIS 2021 技术文档 First Look 30 -- 使用 InterSystems 产品进行文本分析

本技术概览（ First Look ）介绍了 InterSystems IRIS® 数据平台 支持使用 Natural Language Processing（NLP，自然语言处理 ）文本分析的能力，NLP
文本分析以各种自然语言对非结构化文本数据进行语义分析。能让您发现有关大量文本文档内容的有用信息，而无需事先了解文本内容。

本技术概览（ First Look ）介绍了 InterSystems IRIS Natural Language Processing
（自然语言处理），并介绍了一些与索引文本数据相关的初始任务，以进行语义文本分析。完成这些任务后，您将对一组文本建立索引并执行分析，以确定这些文本中
最常见的实体（entity
）、关于这些实体的度量指标、实体之间的各种关联，以及查看实体在源文本的表现形式。这些活动仅设置使用默认设置和功能，以便您熟悉 NLP 文本分析的基础知识。有关Text
Analytics（文本分析）的完整文档，请参阅InterSystems IRIS Natural Language Processing (NLP) Guide （《InterSystems IRIS 自然语言处理 (NLP) 指南》）。

处理非结构化文本的一个相关但独立的工具是InterSystems IRIS SQL Search。SQL Search允许您*search*（搜索）这些相同的实体，以及多个文本中的单个单词、正则表达式（regular expression）和其他结构。本质上，搜索解决方案的前提是您知道自己要寻找的目标。NLP
文本分析旨在帮助您发现内容和内容实体之间的联系，而无需明确要寻找的目标。

要浏览技术概要（First Look）的所有内容，包括可以在InterSystems IRIS的免费评估实例上执行的许多内容，请参阅InterSystems First Looks （《InterSystems 技术概要》）。

1. 为什么 NLP 文本分析很重要

企业逐渐累积的越来越多的非结构化文本数据，远远超出了他们阅读或记载这些文本的能力。通常，企业可能对这些文本文档的内容知之甚少。基于纯*search*
（搜索）技术传统的“ 自上而下 ”文本分析，对这些文本的内容做出假设，可能会遗漏重要内容。

InterSystems IRIS Natural Language Processing （NLP
）可以在预先不了解这些文本的主题的情况下，对这些文本进行文本分析。它通过应用识别语义实体的特定语言规则来实现这项功能。由于这些规则是专门针对语言而非内容，因此 NLP
可以在不使用字典或本体的情况下提供对文本内容的深入分析。

2. InterSystems IRIS 如何实施 NLP 文本分析

要为 NLP 分析准备文本，您必须将这些文本加载到domain
（域）中，然后构建域。基于对这些文本的分析，NLP为该域构建索引，
NLP可以使用域快速分析大量文本。文本可以从各种数据位置输入，包括 SQL 表、文本文件、strings（字符串）、globals 和 RSS 数据。

NLP 支持以下功能：

- Language models（语言模型）：识别单词之间的语义关系是专门针对语言的。NLP 包含十种自然语言的语义规则（语言模型），可以分析用该语言编写的任何主题的文本。如果指定的语言不止一种，NLP 通过确定每个文本中的每个句子与指定语言之间的最佳匹配来执行自动语言识别。NLP 分析不需要预先创建或关联字典或本体，尽管您可以通过添加它们来扩展其功能。
- Entity analysis（实体分析）：一个或多个单词的语义群，被称为实体，NLP对其进行操作。实体被定义为Concepts（包括名词和名词短语）或Relations（包括动词和介词）。通常，要考虑的最相关实体是Concepts，但也可以分析Relations。句子和单词始终泾渭分明。忽略字母大小写。
- Path analysis（路径分析）：NLP 将Concepts和Relations的连贯序列分组为Paths。一个句子通常由一个单独的 Path 组成。Path 反映了实体之间的联系。
- Attributes（属性）：NLP 标记语义属性，例如否定，以便您可以区分肯定的文本序列（“ 结构损坏的证据 ”）和否定的文本序列（“ 没有结构损坏的证据 ”）。
- Frequency、Spread 和 Dominance：这些是为实体计算的度量指标。Frequency 是实体在一组文本中出现的次数。Spread是包含该实体的文本数。Dominance 是一个更细微的指标，通过将实体重复出现次数相对于每个文本的长度、具有共同单词的其他实体的重复出现次数以及其他因素考虑在内，形成该指标。实体通常按这些指标按降序排列。通过这些指标可以了解文本内容，您能更加深入地分析特定实体。
- Similar Entities、 Related Concepts 和 Proximity Profile。给定一个实体，这些功能可以发现其他相关实体。例如，给定一个简短的实体，相似的实体将包括域中包含相同词的其他更长的实体，从而是比种子实体更具体的实体。给定一个实体，相关实体是同一句子中通过单个Relation与指定实体相关联的其他实体。给定一个实体，Proximity 指标度量计算指定实体与其他实体之间在路径内的距离。
- Dictionaries（字典）：您可以添加可选的字典来识别实体的同义词。
- Summarization（摘要）：您可以使用 NLP 生成文本摘要，要求摘要按整体文本的一定百分比生成。例如，50% 的摘要将包含原始文本中一半的句子，NLP 会选择那些被计算为与整个源文本最相关的句子。

3. 亲自尝试 NLP 文本分析

使用 InterSystems IRIS 文本分析很容易。这个简单的程序将引导您按基本步骤生成NLP 度量指标。

提供此示例是为了让您初步体验InterSystems IRIS Natural Language Processing。不应将此示例用作开发实际应用程序的基础。如要在真实情况下使用 NLP，您应该充分研究该软件提供的可用选项，然后开发您的应用程序以生成稳健且高效的代码。

用前须知

要使用该程序，您需要一个正在运行的 InterSystems IRIS 实例。您的选择包括多种类型的已授权的和免费的评估实例；该实例不需要在您工作的系统中（尽管它们必须相互具有网络访问权限）。如果您还没有一个可以使用的实例，如何部署每种类型实例的有关信息，请参阅*InterSystems IRIS Basics: Connecting an IDE*（*《INTERSYSTEMS IRIS 基础：连接一个IDE》*）中的Deploying InterSystems IRIS（部署 InterSystems IRIS）。

您还需要获取 Aviation.Event SQL 表，该表可在 GitHub 上找到，网址为https://github.com/intersystems/Sam-ples-

Aviation。按照*First Look: SQL Search with InterSystems Products*（*《技术概览：使用 InterSystems 产品进行 SQL 搜索》*）中的Downloading and Setting up the Sample Files（下载和设置示例文件）提供的说明下载和设置文件

创建域并添加数据位置

所有 NLP 分析都发生在一个域内。将多个文本关联到一个域。然后构建域，创建 NLP 查询使用的索引。

域是在namespace（命名空间）中创建的，例如在上一节中按照*First Look: SQL Search with InterSystems Products*（*《技术概览：使用 InterSystems 产品进行 SQL 搜索》*）的过程，创建的**SAMPLES**命名空间。可以在一个命名空间内创建多个域。可以将一个文本与多个域进行关联。

有多种方法可以创建、填充和构建域。以下示例使用**Domain Architect**界面。

1. 在浏览器（browser）中为实例打开 Management Portal （管理门户），使用 *InterSystems IRIS Basics: Connecting an IDE*（*《INTERSYSTEMS IRIS 基础：连接一个IDE 》*）的URL described for your instance（实例适用的URL）。
2. 导航到 Domain Architect 页面（**Analytics > Text Analytics > Domain Architect**）。使用**Analytics**选项之前，可能需要切换到启用分析功能的**SAMPLES**命名空间。
3. 点击 **New** **(新建)** 按钮定义域。指定以下域值（按给定的顺序）：
   - **Domain name**：分配给域的名称在当前命名空间必须是唯一的（不仅仅是在其package class（包类）中唯一）；域名不区分大小写。本示例指定名称为 MyTest。
   - **Definition class name**：域定义 package name （包名）和 class name（类名），用句号分隔。从 **Domain name**字段按 Tab 键生成默认**Definition class name**：Samples.MyTest 。
   - **Allow Custom Updates**：此复选框允许手动将数据或字典添加到此域。本示例请勿选中此框。

点击 **Finish** **(结束)** 按钮来创建域。在屏幕中将显示 **Model Elements** 选项。

4. 在域内，可以为域定义数据位置和其他模型元素。要添加或修改模型元素，请单击标题之一旁边的扩展三角形。一开始，没有扩展三角形出现。定义一些模型元素后，单击扩展三角形会显示您定义的模型元素。

点击 **Data Locations** 三角形以在屏幕右侧显示 Details 选项卡。Details选项卡显示五个 Add Data 选项。选择 **Add data from table**。

此选项允许您指定存储在 SQL 表中的数据。在本例中，我们将指定以下字段：

Name：提取的数据文件集的名称。使用默认值：Table1 。

Schema：从下拉列表中选择 Aviation。

Table Name：从下拉列表中选择 Event。

ID Field：从下拉列表中选择 ID。

Data Field：从下拉列表中选择 NarrativeFull。

如果当前域定义有未保存的更改，**Domain Architect** 页面标题后跟一个星号 (*)。点击 **Save** **(保存)** 保存更改。

5. 按 **Compile** **(编译)** 按钮，编译 Domain。

6. 按 **Build** 按钮，为源数据构建NLP索引。

分析数据

使用以下过程分析数据：

1. 在 Domain Architect 页面上，选择屏幕右侧的 **Tools** 选项卡，然后单击 **Domain Explorer** 按钮。
2. Domain Explorer 最初显示源文本中最重要概念的列表：
   - **frequency**选项卡按频率降序显示**Top Concepts**。每个列出的项目都列示其频度计数（出现次数）和分布计数（包含该概念的源数量）。

例如， pilot概念的重复出现次数为6206，分布计数为1085；student pilot概念的重复出现次数为 319，分布计数为141。

   - **dominance**选项卡按 dominance 计算结果，降序显示**Dominant Concepts**。

例如，pilot概念拥有 351.6008 的dominance；student pilot概念拥有 49.3625 的dominance。

3. 当您选择上述其中一项概念，将显示该概念的另一个**Domain Explorer**列表：
   - **Similar Entities** 列出所选概念以及包含该概念的所有其他概念，每个概念都有其重复出现次数和分布计数。

例如，选择student pilot，Similar Entities 将显示包括student pilot, student pilot certificate, student pilot's logbook, solo student pilot在内的列表。

   - **Related Concepts**
     列出与所选概念相关的其他概念，以及在上下文语境中该实例的这些概念的重复出现次数和分布技术。

例如，选择student pilot，Related Concepts 将显示包括flight instructor和airplane在内的列表。

   - **Proximity Profile**
     列出所选概念相邻的其他概念，以及当发现在相同的句子中有与所选概念相同的概念，将计算这些概念的实例的相邻程度。

例如，选择student pilot，Proximity Profile 将列示 airplane的相邻程度为 2702，以及flight instructor的相邻程度为 1662 。

在上述任一列表中的
选择一个概念，这些列表将根据该概念进行刷新。或者，您也可以将实体（Concept或Relation）键入到 **Domain Explorer Explore** 区域并单击 **Explore!** 按钮。

通过使用这些列表，可以确定源文档中出现哪些概念、它们的重要性以及与它们相关联的其他概念。

Domain Explorer 的下部分允许您查看所选概念在源文本中的显示方式：

Sources
选项卡按来源列出包含所选概念的所有句子。该概念被高亮显示，红色文本用于表示涉及该概念的否定。

Paths
选项卡列出了包含所选概念的
所有路径。路径文本被突出显示，以显示 NLP
索引：所选概念以橙色高亮显示，路径中的其他概念以蓝色高亮显示，与路径相关的概念（通常是代词）以浅蓝色高亮显示。关系显示为白色。红色文本用于表示涉及该概念的否定。

通过点击 **eye** 图标，可以显示源的完整文本，所选概念将被高亮显示，并使用红色文本表示否定。

Indexing 切换按钮显示源的完整文本，高亮显示 NLP

Paths列表中。

%
选项允许您显示文
本摘要。指定一个百分比。文本中的
句子总数减少到该百分比。NLP 包含在摘要中的句子由它们对全文的重要性计算确定

4. 您可以添加一个 skiplist 来排除不需要的概念。通常，top
concepts
（上层概念）列表始于那些非常常见或在几乎没有什么有用信息的概念。这些可能是出现在所有来源中的词或短语（例如 "accident
report（事故报告）"或"conclusions（结论）"）、一般概念（例如"airplane（飞机）"或"pilot
（飞行员）"），或与您所使用数据无关的概念（例如城市列表）。您可以使用
skiplist 避免显示这些概念。Skiplist 仅影响某些查询结果中概念的显示；它对概念的 NLP 索引没有影响。

a. 在 Domain Architect 中单击 **Open** 按钮并选择 Samples >>，然后是 MyTest 打开现有域 Samples.MyTest。

b. 点击 **Skiplists** 扩展三角形。将在屏幕右侧的 Details 选项卡显示 **Add skiplist** 按钮。点击 **Add skiplist** 以显示
Name 和 Entries 字段。接受 skiplist 的默认名称 （Skiplist1 ）。在
Entries 框列示条目（概念），一行一个概念；条目不区分大小写。本例中列出如下概念：pilot（飞行员）
student pilot（学生飞行员）、 co-pilot（副驾驶）、 passenger（乘客）、instructor（教官）、flight
instructor（飞行教官）、certified flight instructor（认证飞行教官）。

c. **Save** **（保存）** 和**Compile** **（编译）** 域。（不需要使用 **Build** 功能去添加、修改或删除 skiplists 的域）。

d. 在Domain Explorer 中单击
sunglasses 右上角的图标。将显示您可以应用的为该域定义的skiplists的列表。选择Skiplist1 。请注意，**Top
Concepts** 不再列出 skiplists 的概念。

4. 了解有关 NLP 文本分析的更多信息

InterSystems 有其他相关资料可帮助您了解有关 NLP 文本分析的更多信息，包括：

- InterSystems IRIS Natural Language Processing (NLP) Guide（《InterSystems IRIS自然语言处理 (NLP) 指南》）

#InterSystems IRIS #InterSystems IRIS for Health

源URL:
https://cn.community.intersystems.com/post/iris-2021-%E6%8A%80%E6%9C%AF%E6%96%87%E6%A1%A3-first-look-30-%E4%BD%BF%E7%94%A8-intersystems-%E4%BA%A7%E5%93%81%E8%BF%9B%E8%A1%8C%E6%96%87%E6%9C%AC%E5%88%86%E6%9E%90