Text Data Processing

SAP Data Services Text Data Processing enables you to perform natural language processing and extraction processing on unstructured text. This capability was introduced in Data Services 4.0 and has been enhanced further in subsequent releases. Text Data Processing now supports extracting information from binary documents, such as Word and PDF, richer entity extraction in 31 different languages, and can be pushed down to execute directly in Hadoop. In the latest Data Services 4.2 release, the Entity Extraction transform has added language identification, pre-defined entity type support for Dutch and Portuguese, and sentiment analysis extraction in Simplified Chinese.

Entity Extraction Transform

The Entity Extraction transform found under the Text Data Processing feature on Data Services Designer performs linguistic processing on content by using semantic and syntactic knowledge of words. You can configure the transform to identify paragraphs, sentences, and clauses and it can extract entities and facts from text. Typically, you use the Entity Extraction transform when you have text with specific information you want to extract and then use in downstream analytics and applications.

- How to Customize Rules
- Mapping Input and Output Fields
- Configuring Extraction Options
- Recommendations on Best Practices
- Best Practices for Entity Extraction Using Text Analysis XI 3.x
- CGUL Tips and Tricks for Entity Extraction

Language Modules

Each language module (formally known as Lingware) consists of a set of files that include system dictionaries containing words to support the language processing operations for the given natural language. It is the language modules that enable linguistic analysis and extraction of unstructured text in a given language.

- Language Modules Reference
  - Out-of-the-box entities, relations, and events pre-configured to extract key information – the who, what, where, when, and how in text
  - Available in English, French, German, Simplified Chinese, Spanish, Italian, Russian, Arabic, Farsi, Portuguese, Dutch, Bokmal, Catalan, Croatian, Czech, Danish, Greek, Hebrew, Hungarian, Nynorsk, Polish, Romanian, Serbian, Slovenian, Swedish, Thai, Traditional Chinese, Turkish

  Note: The languages in bold support pre-defined entity types out-of-the-box while the remainder may only support the NOUN_GROUP entity type. All of the languages support dictionary- and rule-based extraction.

Certain language modules include specialized content that provides entity types and sets of rules that address specific needs:

- Voice of the Customer
  - Out-of-the-box entities, relations, and requests pre-configured to extract key information for sentiments: strong positive, weak positive, neutral, weak negative, strong negative, problems
  - Available in English, French, German, Spanish, and Simplified Chinese

- Public Sector
  - Out-of-the-box entities and relations pre-configured to extract key information such as person-organization, person-alias, travel events and security (weapons, vehicles, facilities, etc.)
  - Available in English
  - Available in Arabic and Simplified Chinese for security only

- Enterprise
  - Out-of-the-box entities, relations, and events pre-configured to extract key information for mergers and acquisitions, as well as executive job changes
  - Available in English
Blueprints

End-to-end jobs posted on the SAP Community Network that are configured with best practices to solve a specific scenario which include sample data and may be run in a customer's environment with only a few modifications.

- **Text Data Processing Language 4.2**
  - Contains the necessary project, jobs, data flows, file formats, sample data, template tables, and custom functions configured to illustrate the use of the Entity Extraction transform in common scenarios
  - One English blueprint job highlights retrieving tweets from Twitter performing sentiment analysis on them to quantify sentiment around one or more search terms
  - Available in English, German, Japanese, and **Simplified Chinese**

- **Text Data Processing Data Quality 4.2**
  - Contains the necessary project, jobs, data flows, file formats, sample data, template tables, and custom functions configured to illustrate the use of the Entity Extraction transform in conjunction with Data Quality transformations
  - One blueprint job showcases sentiment analysis and utilizes the Match transform to group similar topics for visualizing via a SAP BusinessObjects Universe and two Web Intelligence reports

- **Text Data Processing Miscellaneous 4.2**
  - Contains sample jobs configured to illustrate the use of Text Data Processing for language identification and how input text is parsed to support custom rule creation.

Relevant Data Services Documentation

- **Designer Guide** -- contains an overview of Text Data Processing and its Entity Extraction transform
- **Reference Guide** -- provides details on the options and input/output fields supported by the Entity Extraction transform
- **Text Data Processing Extraction Customization Guide** – contains information about building dictionaries and extraction rules to create your own extraction patterns to use with the Entity Extraction transform
- **Text Data Processing Language Reference Guide** – contains information about the linguistic analysis and extraction features of the Entity Extraction transform provides, as well as a reference section for each language module

Other Data Services Content

- **Data Services Tutorials**
- **Data Quality**
- **Data Services Installation and Configuration**
- **Data Services Usage Questions**
- **HANA**
- **Performance of Data Services**
- **SAP Data Services Blueprints**
- **SAP Topics**
- **Text Data Processing**
- **Troubleshooting Guides**

Search the EIM Wiki: