A Data Pro

Semantic technologies research and development group

Annotation services documentation

Contents:

1. Overview

The purpose of this document is to describe annotation service as part of “Identrics Trinity” text mining framerwork. Document describes behavior and usage of annotation service. It includes detailed overview of request and response parameters with working examples.

2. Annotation service description

Trinity Annotation service is entry point for all kinds of machine learning analysis provided by Identrics technology stack. It takes the content and initial metadata of particular text document and organizes other background services to process that document. The output of background services is then organized in single array of typed objects and finally returned as annotation service output. Analytical tasks such as “sentiment analysis”, “document similarity”, “NER” etc. are also part of the input, so the request object contains information which machine learning tasks to be executed against one particular text document.

This documentation covers the following tasks:

- sentiment analysis
- multi-label taxonomy classification
- named entity recognition and classification
- document similarity clustering

The output of annotation service is array of annotation objects with specific schemata related to the task performed by background services.

3. Web API documentation

Annotation service is available through Web-based API with REST-full support. The client-server communication is performed using JSON serialization of request parameters and response objects.

Queries are organized in such manner:

http://<host>:<port>/services/annotation/annotate?json=[<JSON list of parameters>]

Responses are organized in such manner:

[{response object 1},{response object 2},{response object 3}]

3.1 Request parameters

Annotation server has one method signature with four parameters described in usage order as follows:

contextName: Json string representation of particular use case scenario that applies machine learning analysis. By default it is “annotation”.

documentId: Json string representation of unique persistent identifier for the document.

documentText: Json string representation of document text body. This parameter contains entire text block to be analyzed.

taskList: Json array containing string representations of machine learning tasks to be performed by background services. Task list may contains all or part of the following items:

“classifier” - multi-label taxonomy classification task
”sentiment” - sentiment classification task
”ner” - named entity recognition task
”docsim” - document similarity task

Annotation query should be specified as:

http://<host>:<port>/services/annotation/annotate?json=[“annotation”,”docID_123”,”Document text body”,[“classifier”,”sentiment”,”ner”,”docsim”]]

Note that Json request parameters must contains escaped forms of all Json reserved symbols, otherwise web server will return Json parse error. At least that escapes must be performed in Java and Javascrip like notation:

"\\", "\\\\"
"\"", "\\\""
"\n", "\\n"
"\r", "\\r"
"\b", "\\b"
"\f", "\\f"
"\t", "\\t"

3.2 Response objects

Annotation server returns three different types of Json response objects. Taxonomy classifier and sentiment classifier are represented with one and the same annotation object scheme but they are explicitly marked as “taxonomy” or “sentiment” by Json property “**annotationType**”. Another Json types of response objects are for named entity recognition and document similarity. All they are (if available in the result) combined in single Json array and that array can be empty if no results are returned from the server, at least one or many objects of any of the mentioned types. Not that the order

of the response object types matches the order of annotation tasks from the request.

3.2.1 Sentiment response object

Is specified as:

{"anotationType":"sentiment","uri":null,"annotationClass":"<sentiment response value>"}

anotationType: Type of the class annotation. For sentiment its always “sentiment”
uri: Reserved to represent response in URI standard. Not implemented yet – always “null”
annotationClass: Actual value of the sentiment class – can be one of the following values:

-1	negative class
0	neutral class
1	positive class

Annotation service response array is restricted to contains zero or one sentiment object.

3.2.2 Taxonomy classifier response object

Is specified as:

{"anotationType":"taxonomy","uri":null,"annotationClass":"<some taxonomy class>"}

anotationType: Type of the class annotation. For sentiment its always “taxonomy”
uri: Reserved to represent response in URI standard. Not implemented yet – always “null”
annotationClass: Actual value of the taxonomy class.

Annotation service response array is restricted to contains zero, one or more taxonomy objects.

3.2.3 Named entity recognition (NER) response object

Is specified as:

{"startOffset":<integer value>,"endOffset":<integer value>,"annotationClass":"<string value>","word":"<string value>","annotationURI":null,"index":<integer value>,"annotationType":"ner"}

startOffset: Start position of the named entity mention in the text
endOffset: End position of the named entity mention in the text
annotationClass: Actual class of the entity. Can be one of the following:
PERSON, LOCATION, ORGANIZATION, DATE, PERCENT, MONEY, MISCELANEOUS

word: String representation of named entity as it is found in the text.
annotationURI: Reserved to represent response in URI standard. Not implemented yet – always “null”
index: Word index of the entity in tokenized text.

annotationType: Type of the annotation response object. For that object it's always “ner”

Annotation service response array is restricted to contains zero, one or more NER objects.

3.2.4 Document similarity response object

Is specified as:

{"annotationClass":null,"docId":<string value>,"annotationType":"docsim"}

annotationClass: Reserved for additional annotation metadata. Not implemented yet – always “null”
docId: String representation of unique persistent identifier of document which is contextually similar to the document from the request.
annotationType: Type of the annotation response object. For that object it's always “docsim”

Annotation service response array is restricted to contains zero, one or more Docsim objects.

3.3 Usage

The following is a real usage example of annotation request and response:

Query:

http://10.50.30.23:8080/services/annotation/annotate*?json=[**"annotation","111",**"Миналата година се е оказала поредната успешна за българските публични компании. Първите публикувани отчети до фондовата борса показват ръст на приходи и печалби, като последните при някои се удвояват. Повишеното вътрешно потребление и външно търсене са в основата на добрите резултати. Отчетите на борсово търгуваните дружества са първият индикатор за случващото се в икономиката на корпоративно ниво. Добрата новина е, че растеж има в разнообразни сектори като услуги, машиностроене и фармация. От публикуваните дотук данни за по-големи борсови дружества се вижда, че преобладаващо растат печалбите, а при повечето - и приходите. Например при куриерската Спиди продажбите се вдигат със 7%, а печалбата с 19%. Производителят на радиатори Корадо - България почти удвоява печалбата, а оборотът се вдига с 30%. Европейските пазари увеличиха и приходите, и финансовия резултат на Софарма, нагоре с двуцифрени проценти са и показателите на козметичната Лавена. Годината е била силна и за машиностроителите: приходите на М+С хидравлик - Казанлък, и ХЕС - Ямбол, растат с 1 от 5, показват данните на Стара планина холд. И кабелният производител Емка влиза в групата на удвоили печалбата си и отчели силен ръст в продажбите.",**["sentiment","classifier","ner","docsim"]]***

Response:


```
[
{"anotationType":"sentiment","uri":null,"annotationClass":"1"},
{"uri":null,"annotationClass":"PROSAL","annotationType":"taxonomy"},
{"annotationClass":"LOCATION","word":"България","annotationURI":null,"startOffset":748,"endOffset":756,"index":18,"annotationType":"ner"},
{"annotationClass":"ORGANIZATION","word":"Софарма","annotationURI":null,"startOffset":877,"endOffset":884,"index":39,"annotationType":"ner"},
{"annotationClass":"PRODUCT","word":"ХЕС","annotationURI":null,"startOffset":1044,"endOffset":1047,"index":17,"annotationType":"ner"},
{"annotationClass":"PRODUCT","word":"Ямбол","annotationURI":null,"startOffset":1050,"endOffset":1055,"index":19,"annotationType":"ner"},
{"annotationClass":"ORGANIZATION","word":"Стара","annotationURI":null,"startOffset":1094,"endOffset":1099,"index":30,"annotationType":"ner"},
{"annotationClass":null,"docId":"31","annotationType":"docsim"},
{"annotationClass":null,"docId":"502","annotationType":"docsim"},
{"annotationClass":null,"docId":"634","annotationType":"docsim"},
{"annotationClass":null,"docId":"213","annotationType":"docsim"},
{"annotationClass":null,"docId":"592","annotationType":"docsim"},
{"annotationClass":null,"docId":"514","annotationType":"docsim"},
{"annotationClass":null,"docId":"284","annotationType":"docsim"},
{"annotationClass":null,"docId":"330","annotationType":"docsim"},
]
```