txtDataNLP Framework – Documentation

Table of Contents

[1 Introduction 1](#_Toc31549385)

[2 Central Classes and Concepts 2](#_Toc31549386)

[2.1 TextUnit 2](#_Toc31549387)

[2.2 Word 3](#_Toc31549388)

[2.3 Annotation 3](#_Toc31549389)

[2.4 Span 3](#_Toc31549390)

[2.5 IAnnotationObject 3](#_Toc31549391)

[2.6 AbstractAnnotator 3](#_Toc31549392)

[2.7 AbstractCreator 3](#_Toc31549393)

[2.8 Language 3](#_Toc31549394)

[3 Annotators 3](#_Toc31549395)

[3.1 What is an Annotator? 3](#_Toc31549396)

[3.2 WhitelistAnnotator 3](#_Toc31549397)

[3.3 PosPatternAnnotator 3](#_Toc31549398)

[3.4 SubsumedAnnotationsRemover 3](#_Toc31549399)

[3.5 RecursiveDictionaryAnnotator 3](#_Toc31549400)

[4 Common use cases 3](#_Toc31549401)

[4.1 Creating a simple dictionary-based annotator 4](#_Toc31549402)

[4.2 Creating a more complex annotator based on “POS-Rules” 4](#_Toc31549403)

[4.3 Working with annotations from a third-party Machine Learning model 4](#_Toc31549404)

[4.4 Implementing white- and blacklists on top of a third-party Machine Learning model 4](#_Toc31549405)

[4.5 Building a complex NLP-Pipeline that combines various Machine Learning and/or rule-based approaches 4](#_Toc31549406)

# Introduction

txtDataNLP is a light Java framework for Natural Language Processing (NLP) projects.

Typical use cases include:

* Combining various third-party NLP and Machine Learning (ML) tools into one setup.
* Implementing algorithms that combine output from various NLP/ML modules and return one coherent text analysis—even if their output is contradictory.
* Extraction of features from low-level NLP tools (tokenizers, taggers) for downstream ML models.
* Implementing dictionary-based, rule-based or linguistically motivated approaches to NLP.
* Implementing white- or backlists that often sit on top of ML approaches in order to deal with specific mistakes the ML models make.

txtDataNLP contains no ML algorithms or models itself but is was built to support the use of ML models in NLP and to provide infrastructure to combine various NLP/ML modules, use them together reliably and to fill the gaps that they commonly leave.

# Central Concepts and corresponding classes

## TextUnit

In many aspects the TextUnit class is the central class of txtDataNLP. It represents a piece of text that is supposed to be analyzed and annotated. In many cases, the text represented in one instance of TextUnit is a sentence, but this does not have to be the case. In some scenarios, a TextUnit might also represent different things, e.g. a search query, a headline, a sub-heading or even a complete paragraph or article.

Each TextUnit instance keeps track of three kinds of information:

1. The surface of the respective text, as a Java String. It can be retrieved with the getSurfaceText() method.
2. A list of words that represent the words the text contains, using the Word class, see below. Words can be retrieved with getWords().
3. A list of annotations, that represent interesting bits in the text, using the Annotation class, see below. All annotations of this TextUnit can be retrieved with getAnnotations().

The TextUnit class contains a toString method which is useful to get a quick idea of what is stored in an instance:

|  |
| --- |
| He lives in London and she in New York City.  He PRP (0-2)  lives VBZ (3-8)  in IN (9-11)  London NNP (12-18)  and CC (19-22)  she PRP (23-26)  in IN (27-29)  New NNP (30-33)  York NNP (34-38)  City NNP (39-43)  . . (43-44)  [12-18] 'London' Location  [30-43] 'New York City' Location |

The first line contains the surface text of the analyzed sentence. The following lines list each of the words that are present in the sentence, together with their part-of-speech (POS) tags and the start and end positions of each word in the surface text. Finally, the last to lines show annotations: Here the NLP has detected two entities that are present in the input text: “London” and “New York City”. These are given together with the position where they occur in the sentence and their semantic class: “Location”.

Note that TextUnit merely provides a container to store such information. The actual analysis necessary to determine word boundaries, POS tags and entities is done somewhere else and can be configured.

## Word

The Word class represents a word that occurs in a text. It contains various pieces of information about that word, for example its surface appearance, the position in characters where the word starts in the text in which it was found, the position where it ends etc.

Since words appear in text and text in txtDataNLP is typically represented with the TextUnit class, the words you will be working with often originate from TextUnit’s getWords() method.

Important methods of the Word class are:

* getSurface() Returns the surface structure of the word.
* getStarts() Returns the position where this word starts in the text in which it occurs.
* getEnds() Returns the position where this word ends in the text in which it occurs.

In a typical scenario the three just mentioned method should always return a value and never return null.

Other methods that might be set, or return null, depending on the scenario are:

* getRoot(), returns the root form of the word, e.g. “house” for “houses”.
* getPOS(), return the part-of-speech for the word, e.g. “nn” for “house”.
* getMorph(), returns morphological information for the word.
* getIdf(), returns an IDF score for the word.

For these four methods, the values returned is highly dependent on the actual project setup. The project needs to be configured so that these fields are filled with relevant information. In many project scenarios, some or all of these fields might just return null.

## Annotation

An annotation represents a piece of text that is for some reason interesting or relevant in a certain scenario. If you are for example looking to detect locations in text and one sentence of that text is “He lives in London and she in New York City” the most likely parts of this text that will be annotated are “London” and “New York City”.

As the Word class, the Annotation class has the following three methods:

* getSurface() Returns the surface structure of the annotation.
* getStarts() Returns the position where this annotation starts in the text in which it occurs.
* getEnds() Returns the position where this annotation ends in the text in which it occurs.

Each Annotation class also has a dedicated slot where semantic information belonging to the annotation can be stored: IAnnotationObject, see below.

## Span

## IAnnotationObject

## AbstractAnnotator

## AbstractCreator

## Language

# Annotators

## What is an Annotator?

## WhitelistAnnotator

## PosPatternAnnotator

## SubsumedAnnotationsRemover

## RecursiveDictionaryAnnotator

# Common use cases

## Creating a simple dictionary-based annotator

## Creating a more complex annotator based on “POS-Rules”

## Working with annotations from a third-party Machine Learning model

## Implementing white- and blacklists on top of a third-party Machine Learning model

## Building a complex NLP-Pipeline that combines various Machine Learning and/or rule-based approaches