
CoreNLP XML Library Documentation

Release 0.0.1

Robert Elwell

June 12, 2014

1	Installing the Library	3
2	What You Can Do With This Library	5
2.1	The Document Model	5
2.2	Interacting with Dependencies	8
2.3	Coreference Resolution	10
3	Indices and tables	13
	Python Module Index	15

This library is designed to add a data model over Stanford CoreNLP's basic XML output.

The Document class is designed to provide lazy-loaded access to information from syntax, coreference, and dependency parse structures within the XML.

Installing the Library

It's as easy as

```
pip install corenlp_xml
```

What You Can Do With This Library

Some code examples:

```

from corenlp_xml import Document

doc = Document(xml_string)

# The first sentence
s1 = doc.sentences[0]

# Noun phrases for the first sentence
s1_nps = s1.phrase_strings("np")

# Text of semantic head of first sentence
s1_head = s1.semantic_head.text

# Find all representative coreferences matching noun phrases in sentence 1
s1_corefs = [coref for coref in doc.coreferences
              if coref.representative and coref.sentence == s1]

```

Contents:

2.1 The Document Model

Sub-module for handling document-level stuff

class `corenlp_xml.document.Document` (*xml_string*)

This class abstracts a Stanford CoreNLP Document

coreferences

Returns a list of Coreference classes

Getter Returns a list of coreferences

Type list of `corenlp_xml.coreference.Coreference`

get_sentence_by_id (*id*)

Gets sentence by ID

Parameters *id* (*int*) – the ID of the sentence, as defined in the XML

Returns a sentence

Return type `corenlp_xml.document.Sentence`

sentences

Returns the ordered dict of sentences as a list.

Getter returns list of sentences, in order

Type list of `corenlp_xml.document.Sentence`

sentiment

Returns average sentiment of document. Must have sentiment enabled in XML output.

Getter returns average sentiment of the document

Type float

class `corenlp_xml.document.Sentence` (*element*)

This abstracts a sentence

basic_dependencies

Accesses basic dependencies from the XML output

Getter Returns the dependency graph for basic dependencies

Type `corenlp_xml.dependencies.DependencyGraph`

collapsed_ccprocessed_dependencies

Accesses collapsed, CC-processed dependencies

Getter Returns the dependency graph for collapsed and cc processed dependencies

Type `corenlp_xml.dependencies.DependencyGraph`

collapsed_dependencies

Accesses collapsed dependencies for this sentence

Getter Returns the dependency graph for collapsed dependencies

Type `corenlp_xml.dependencies.DependencyGraph`

get_token_by_id (*id*)

Accesses token by the XML ID

Parameters *id* (*int*) – The XML ID of the token

Returns The token

Return type `corenlp_xml.document.Token`

id

Returns the ID attribute of the sentence

Return type `int`

parse

Accesses the parse tree based on the S-expression parse string in the XML

Getter Returns the NLTK parse tree

Type `nltk.Tree`

parse_string

Accesses the S-Expression parse string stored on the XML document

Getter Returns the parse string

Type `str`

phrase_strings (*phrase_type*)

Returns strings corresponding all phrases matching a given phrase type

Parameters **phrase_type** (*str*) – POS such as “NP”, “VP”, “det”, etc.

Returns a list of strings representing those phrases

semantic_head

Returns the semantic head of the sentence – AKA the dependent of the root node of the dependency parse

Returns the mention related to the semantic head

Return type `corenlp_xml.coreference.Mention`

sentiment

The sentiment of this sentence

Getter Returns the sentiment value of this sentence

Type `int`

subtrees_for_phrase (*phrase_type*)

Returns subtrees corresponding all phrases matching a given phrase type

Parameters **phrase_type** (*str*) – POS such as “NP”, “VP”, “det”, etc.

Returns a list of `NLTK.Tree.Subtree` instances

Return type list of `NLTK.Tree.Subtree`

tokens

The tokens related to this sentence

Getter Returns a a list of `Token` instances

Type `corenlp_xml.document.TokenList`

class `corenlp_xml.document.Token` (*element*)

Wraps the token XML element

character_offset_begin

Lazy-loads character offset begin node

Getter Returns the integer value of the beginning offset

Type `int`

character_offset_end

Lazy-loads character offset end node

Getter Returns the integer value of the ending offset

Type `int`

id

Lazy-loads ID

Getter Returns the ID of the token element

Type `int`

lemma

Lazy-loads the lemma for this word

Getter Returns the plain string value of the word lemma

Type `str`

ner

Lazy-loads the NER for this word

Getter Returns the plain string value of the NER tag for the word

Type str

pos

Lazy-loads the part of speech tag for this word

Getter Returns the plain string value of the POS tag for the word

Type str

speaker

Lazy-loads the speaker for this word

Getter Returns the plain string value of the speaker tag for the word

Type str

word

Lazy-loads word value

Getter Returns the plain string value of the word

Type str

2.2 Interacting with Dependencies

This component is responsible for managing dependency parses

class `corenlp_xml.dependencies.DependencyGraph` (*element*)

Dependency graph, models a dependency parse

get_node_by_idx (*idx*)

Stores each distinct node in a dict

Parameters *idx* (*int*) – the “idx” value of the node

Returns the node instance for that index

Type `corenlp_xml.dependencies.DependencyNode`

links

Accesses links within the graph

Returns a list of `corenlp_xml.dependencies.DependencyLink` instances

Type list of `corenlp_xml.dependencies.DependencyLink`

links_by_type (*dep_type*)

Accesses links within the graph

Parameters *dep_type* (*str*) – the dependency type

Returns a list of `corenlp_xml.dependencies.DependencyLink` instances

Type list of `corenlp_xml.dependencies.DependencyLink`

class `corenlp_xml.dependencies.DependencyLink` (*graph*, *element*)

Represents a relationship between two nodes in a dependency graph

dependent

Accesses the dependent node

Getter returns the Dependent node

Type `corenlp_xml.dependencies.DependencyNode`

governor

Accesses the governor node

Getter Returns the Governor node

Type `corenlp_xml.dependencies.DependencyNode`

class `corenlp_xml.dependencies.DependencyNode` (*graph, element*)

Represents a node in a dependency graph

dependent (*dep_type, node*)

Registers a node as dependent on this node

Parameters

- **dep_type** (*str*) – The dependency type
- **node** (`corenlp_xml.dependencies.DependencyNode`) – The node to be registered as a dependent

Returns self, provides fluent interface

Return type `corenlp_xml.dependencies.DependencyNode`

dependents

Gets dependent nodes

Getter returns a flat list of all governing nodes

Type list of `corenlp_xml.dependencies.DependencyNode`

dependents_by_type (*dep_type*)

Gets the dependents of this node by a given dependency type

Parameters **dep_type** (*str*) – The dependency type

Returns dependents matching the provided type

governor (*dep_type, node*)

Registers a node as governing this node

Parameters

- **dep_type** (*str*) – The dependency type
- **node** –

Returns self, provides fluent interface

Return type `corenlp_xml.dependencies.DependencyNode`

governors

Gets governing nodes

Getter returns a flat list of all governing nodes

Type list of `corenlp_xml.dependencies.DependencyNode`

governors_by_type (*dep_type*)

Gets the governors of this node filtered by a dependency type

Parameters **dep_type** (*str*) – The dependency type

Returns governors matching the provided type

classmethod `load` (*graph*, *element*)

Instantiates the node in the graph if it's not already stored in the graph

Parameters

- **graph** (*corenlp_xml.dependencies.DependencyGraph*) – The dependency graph this node is a member of
- **element** (*lxml.ElementBase*) – The lxml element wrapping the node

text = None

These properties are dicts of link type to node

2.3 Coreference Resolution

This library is responsible for handling coreference resolution parsing from the XML output

class `corenlp_xml.coreference.Coreference` (*document*, *element*)

Reflects a grouping of mentions

mentions

Returns mentions

Returns list of mentions

Return type list

representative

Representative mention

Returns representative Mention

Return type `corenlp_xml.coreference.Mention`

class `corenlp_xml.coreference.Mention` (*coref*, *element*)

Reflects a given mention

head

The token serving as the “head” of the mention

Getter the token corresponding to the head

Type `corenlp_xml.document.Token`

representative

Interprets and normalizes the “representative” attribute”

Getter determines whether the mention is representative

Type bool

sentence

The sentence related to this mention

Getter returns the sentence this mention relates to

Type `corenlp_xml.document.Sentence`

siblings

Accesses other mentions in this coref group

Getter the other mentions for this coref group

Type list of `corenlp_xml.coreference.Mention`

tokens

A list of tokens related to this mention

Getter returns a list of tokens relating to this mention

Type list of corenlp_xml.document.Token

Indices and tables

- *genindex*
- *modindex*
- *search*

C

`corenlp_xml.coreference`, 10
`corenlp_xml.dependencies`, 8
`corenlp_xml.document`, 5