LexNLP Documentation

Release 0.1.6

ContraxSuite, LLC

Mar 22, 2018

Contents

LEXNLP

CHAPTER 1

Table of Contents

1.1 About LexNLP

1.1.1 Purpose

LexNLP is a library for working with real, unstructured legal text, including contracts, plans, policies, procedures, and other material. LexNLP provides functionality such as:

Segmentation and tokenization, such as

- A sentence parser that is aware of common legal abbreviations like LLC. or F.3d.
  - Pre-trained segmentation models for legal concepts such as pages or sections.
  - Pre-trained word embedding and topic models, broadly and for specific practice areas
- Pre-trained classifiers for document type and clause type
- Broad range of fact extraction, such as:
  - Monetary amounts, non-monetary amounts, percentages, ratios
  - Conditional statements and constraints, like "less than" or "later than"
  - Dates, recurring dates, and durations
  - Courts, regulations, and citations
- Tools for building new clustering and classification methods
- Hundreds of unit tests from real legal documents

1.1.2 ContraxSuite Projects

LexNLP is often used as part of ContraxSuite, an open source contract analytics and document exploration platform built by LexPredict. LexNLP and ContraxSuite are related through the following project structure:

- ContraxSuite web application: https://github.com/LexPredict/lexpredict-contraxsuite

- LexNLP library for extraction: https://github.com/LexPredict/lexpredict-lexnlp
- ContraxSuite pre-trained models and "knowledge sets": https://github.com/LexPredict/lexpredict-legal-dictionary
- ContraxSuite agreement samples: https://github.com/LexPredict/lexpredict-contraxsuite-samples
- ContraxSuite deployment automation: https://github.com/LexPredict/lexpredict-contraxsuite-deploy

1.1.3 Citing for academic use

We are currently drafting and submitting a technical whitepaper describing the LexNLP library and documenting its performance on a large corpus of gold-standard contracts. Please contact us at support@contraxsuite.com for more information on citing in academic use.

1.2 LexNLP package

1.2.1 `lexnlp.extract`: Extracting structured data from unstructured text

The `lexnlp.extract` module contains methods that allow for the extraction of structured data from unstructured textual sources. Supported data types include a wide range of facts relevant to contract or document analysis, including dates, amounts, proper noun types, and conditional statements.

This module is structured along ISO 2-character language codes. Currently, the following languages are stable:

- English: `lexnlp.extract.en`

Extraction methods follow a simple *get_X* pattern as demonstrated below:

```
>>> import lexnlp.extract.en.amounts
>>> text = "There are ten cows in the 2 acre pasture."
>>> print(list(lexnlp.extract.en.amounts.get_amounts(text)))
[10, 2.0]
```

1.2.1.1 Pattern-based extraction methods

The full list of supported pattern-based structured data types is below:

- amounts, e.g., "ten pounds" or "5.8 megawatts"
- citations, e.g., "10 U.S. 100" or "1998 S. Ct. 1"
- conditions, e.g., "subject to . . . " or "unless and until . . . "
- constraints, e.g., "no more than" or "
- copyright, e.g., "(C) Copyright 2000 Acme"

- courts, e.g., “Supreme Court of New York”
- dates, e.g., “June 1, 2017” or “2018-01-01”
- definitions, e.g., “Term shall mean . . . ”
- distances, e.g., “fifteen miles”
- durations, e.g., “ten years” or “thirty days”
- geographic and geopolitical entities, e.g., “New York” or “Norway”
- money and currency usages, e.g., “$5” or “10 Euro”
- percents and rates, e.g., “10%” or “50 bps”
- PII, e.g., “212-212-2121” or “999-999-9999”
- ratios, e.g.,” 3:1” or “four to three”
- regulations, e.g., “32 CFR 170”
- trademarks, e.g., “MyApp (TM)”
- URLs, e.g., “http://acme.com/”

1.2.1.2 NLP-based extraction methods

In addition to pattern-based structured data types, the *lexnlp.extract* module also supports NLP methods based on tagged part-of-speech classifiers. These classifiers are based on NLTK and, optionally, Stanford NLP libraries. The list of these modules is below:

- named entity extraction with NLTK maximum entropy classifier
- named entity extraction with NLTK and regular expressions
- named entity extraction with Stanford Named Entity Recognition (NER) models

1.2.2 `lexnlp.nlp`: Natural language processing

The `lexnlp.nlp` module contains methods that assist in natural language processing (NLP) tasks, especially in the context of developing unsupervised, semi-supervised, or supervised machine learning. Methods range from tokenizing, stemming, and lemmatizing to the creation of custom sentence segmentation or word embedding models.

This module is structured along ISO 2-character language codes. Currently, the following languages are stable:

- English: *lexnlp.nlp.en*

Extraction methods follow a simple *get_X* pattern as demonstrated below:

```
>>> import lexnlp.nlp.en.tokens
>>> text = "There are ten cows in the 2 acre pasture."
>>> print(list(lexnlp.nlp.en.tokens.get_nouns(text)))
['cows', 'pasture']
```

The methods in this package are primarily built on the Natural Language Toolkit (NLTK), but some functionality from the Stanford NLP, gensim, and spaCy packages is available to users depending on their use case.

Attention: The sections below are a work in progress. Thank you for your patience while we continue to expand and improve our documentation coverage.

If you have any questions in the meantime, please feel free to log issues on GitHub at the URL below or contact us at the email below:

- GitHub issues: https://github.com/LexPredict/lexpredict-lexnlp
- Email: support@contraxsuite.com

1.2.2.1 Tokenization and related methods

- Extracting tokens, stems, lemmas, and parts of speech

1.2.2.2 Segmentation and related methods for real-world text

- Sentences
- Paragraphs
- Sections
- Pages
- Titles
- Utilities

1.2.2.3 Transforming text into features

- Character Transforms
- Token Transforms, including n-grams and skip-grams

1.3 Changelog

1.3.1 0.1.6 - March 1, 2018

- Improved LexNLP unit test coverage.

1.3.2 0.1.5 - Feb 1, 2018

- Improved LexNLP unit test coverage.

1.3.3 0.1.4 - Jan 1, 2018

- Improved LexNLP unit test coverage.
- Implemented method to get sentence ranges in addition to sentence texts.

1.3.4 0.1.3 - Dec 1, 2017

- Improved LexNLP unit test coverage.

1.3.5 0.1.2 - Nov 1, 2017

- Implemented LexNLP title locator.
- Implemented additional LexNLP transforms for skipgrams and n-grams.
- Improved LexNLP handling for parties with abbreviations and other cases.
- Improved LexNLP handling for amounts with mixed alpha and numeric characters.
- Improved LexNLP unit test coverage.

1.3.6 0.1.1 - Oct 1, 2017

- Improve unit test framework handling for language and locales.
- Implemented method and input-level CPU and memory benchmarking for unit tests.
- Migrated all unit tests to 60 separate CSV files.
- Added over 1,000 new unit tests for most LexNLP methods.
- Reduced memory usage for paragraph and section segmenters.
- Improved handling of brackets and parentheses within noun phrases.
- Added URL locator to LexNLP.
- Added trademark locator to LexNLP.
- Added copyright locator to LexNLP.
- Improved default Punkt sentence boundary detection.
- Added custom sentence boundary training methods.
- Improved handling of multilingual text, especially around geopolitical entities.
- Improved default handling of party names with non-standard characters.
- Enhanced metadata related to party type in LexNLP.
- Improved continuous integration for public repositories.

1.3.7 0.1.0 - Sep 1, 2017

- Refactored and integrate core extraction into separate LexNLP package.
- Released nearly 200 unit tests with over 500 real-world test cases in LexNLP.
- Improved definition, date, and financial amount locators for corner cases.
- Integrated PII locator for phone numbers, SSNs, and names from LexNLP.
- Integrated ratio locator from LexNLP.
- Integrated percent locator from LexNLP.
- Integrated regulatory locator from LexNLP.

- Integrated distance locator from LexNLP.
- Integrated case citation locator from LexNLP.
- Improved geopolitical locator to allow non-master-data entity location.
- Improved party locator to allow configuration and better handle corner cases

1.4 License

1.4.1 AGPL License

LexNLP is available by default under the terms of the GNU Affero General Public License v3.0. https://github.com/LexPredict/lexpredict-lexnlp/blob/master/LICENSE

1.4.2 License Release

If you would like to request a release from the terms of the default AGPLv3 license, please contact us at: ContraxSuite Licensing <license@contraxsuite.com>.

CHAPTER 2

Indices and tables

- genindex
- modindex
- search