PaperParser

1. Search for and download papers in a relevant (clean energy) field
   - Test case: perovskite solar cells
2. Process and clean papers into a usable format
3. Search and parse papers for relevant information
   - Synthesis parameters
   - Performance metrics
4. Collect and clean relevant information into a usable format
5. Visualize relationships between synthetic parameters and device performance

PaperParser

1. Search for and download papers in a relevant (clean energy) field
   - Test case: perovskite solar cells
2. Process and clean papers into a usable format
3. Search and parse papers for relevant information
   - Synthesis parameters
   - Performance metrics
4. Collect and clean relevant information into a usable format
5. Visualize relationships between synthetic parameters and device performance

Packages

- spaCy
  - What: Organize text into useful categories
  - Why: Cleaning data (Step 4)
- ChemicalTagger
  - What: Find and organize chemistry-relevant parameters
  - Why: Search and parse (Step 3) (and 4?)
- ChemDataExtractor
  - What: Find chemistry-relevant parameters
  - Why: Search and parse (Step 3)

spaCy-NLP

- https://spacy.io/
- Kim, E., Huang, K., Saunders, A., McCallum, A., Ceder, G., & Olivetti, E. (2017). *Materials Synthesis Insights from Scientific Literature via Text Extraction and Machine Learning. Chemistry of Materials, 29(21), 9436–9444.*

- Free, open-source lib for adv NLP in python
- Fastest syntactic parser in the world
- Its accuracy is within 1% of the best available
- Supports 34+ languages
- Easy deep learning integration
- Easy model packaging and deployment
- Pre-trained statistical models and word vectors
- ~9000 commits and recent commits within a month
- Cited in research journals for Text extraction and machine learning in Materials Science domain

	SPACY	SYNTAXNET	NLTK	CORENLP
Programming language	Python	C++	Python	Java
Neural network models	✓	✓	✗	✓
Integrated word vectors	✓	✗	✗	✗
Multi-language support	✓	✓	✓	✓
Tokenization	✓	✓	✓	✓
Part-of-speech tagging	✓	✓	✓	✓
Sentence segmentation	✓	✓	✓	✓
Dependency parsing	✓	✓	✗	✓
Entity recognition	✓	✗	✓	✓
Coreference resolution	✗	✗	✗	✓

Input

v2.0.18 · Python 3 · via Binder

```
# pip install spacy
# python -m spacy download en_core_web_sm

import spacy

# Load English tokenizer, tagger, parser, NER and word vectors
nlp = spacy.load('en_core_web_sm')

# Process whole documents
text = (u"When Sebastian Thrun started working on self-driving cars at "
        u"Google in 2007, few people outside of the company took him "
        u"seriously. “I can tell you very senior CEOs of major American "
        u"car companies would shake my hand and turn away because I wasn’t "
        u"worth talking to,” said Thrun, now the co-founder and CEO of "
        u"online higher education startup Udacity, in an interview with "
        u"Recode earlier this week.")
doc = nlp(text)

# Find named entities, phrases and concepts
for entity in doc.ents:
    print(entity.text, entity.label_)

# Determine semantic similarities
doc1 = nlp(u"my fries were super gross")
doc2 = nlp(u"such disgusting fries")
similarity = doc1.similarity(doc2)
print(doc1.text, doc2.text, similarity)
```

Output

```
Sebastian Thrun PERSON
Google ORG
2007 DATE
American NORP
Thrun PERSON
Recode ORG
earlier this week DATE
my fries were super gross such disgusting fries 0.7139701576579747
```

Hawizy et al. Journal of Cheminformatics 2011, 3:17
http://www.jcheminf.com/content/3/1/17

Chemical Tagger

- How the package works?
  - Text normalization – preprocessing
  - Tokenization – split into elements
  - Tagging – chemical entities, chemistry related, parts-of-speech
  - Phrase parsing – assign structure to text
  - Action phrase identification – e.g. add, dissolve, stir
- Appeal
  - Separates out action steps, synthesis parameters/conditions etc. in a structured database output
- Drawbacks
  - Java

JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

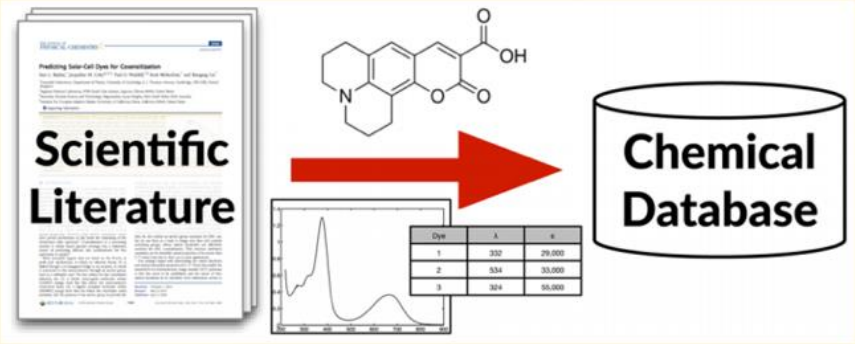
pubs.acs.org/jcim

ChemDataExtractor: A Toolkit for Automated Extraction of Chemical Information from the Scientific Literature

Matthew C. Swain and Jacqueline M. Cole*

Cavendish Laboratory, University of Cambridge, J. J. Thomson Avenue, Cambridge, CB3 0HE, U.K.

"(A) toolkit for the **automated extraction of chemical entities and their associated properties**, **measurements, and relationships from scientific documents** that can be used to populate structured chemical databases."



Characteristics of note;

- chemistry-aware natural language processing pipeline
  - trained by unsupervised word clustering in a "massive corpus of chemistry articles"

JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

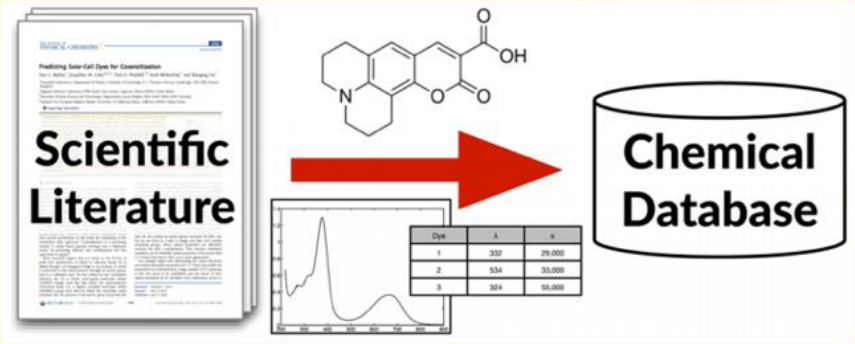
pubs.acs.org/jcim

ChemDataExtractor: A Toolkit for Automated Extraction of Chemical Information from the Scientific Literature

Matthew C. Swain and Jacqueline M. Cole*

Cavendish Laboratory, University of Cambridge, J. J. Thomson Avenue, Cambridge, CB3 0HE, U.K.

"(A) toolkit for the **automated extraction of chemical entities and their associated properties**, **measurements, and relationships from scientific documents** that can be used to populate structured chemical databases."



Characteristics of note;

- chemistry-aware natural language processing pipeline
  - trained by unsupervised word clustering in a "massive corpus of chemistry articles"
- document-level processing to resolve data interdependencies

JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

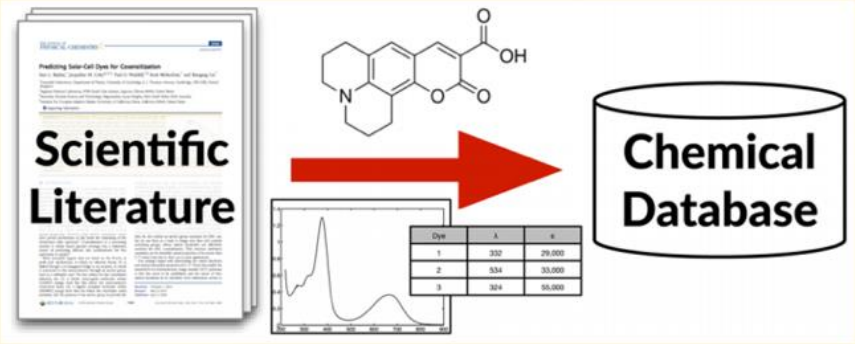
pubs.acs.org/jcim

ChemDataExtractor: A Toolkit for Automated Extraction of Chemical Information from the Scientific Literature

Matthew C. Swain and Jacqueline M. Cole*

Cavendish Laboratory, University of Cambridge, J. J. Thomson Avenue, Cambridge, CB3 0HE, U.K.

"(A) toolkit for the **automated extraction of chemical entities and their associated properties**, **measurements, and relationships from scientific documents** that can be used to populate structured chemical databases."



Characteristics of note;

- chemistry-aware natural language processing pipeline
  - trained by unsupervised word clustering in a "massive corpus of chemistry articles"
- document-level processing to resolve data interdependencies
- The performance of the toolkit to correctly extract various types of data was evaluated

JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

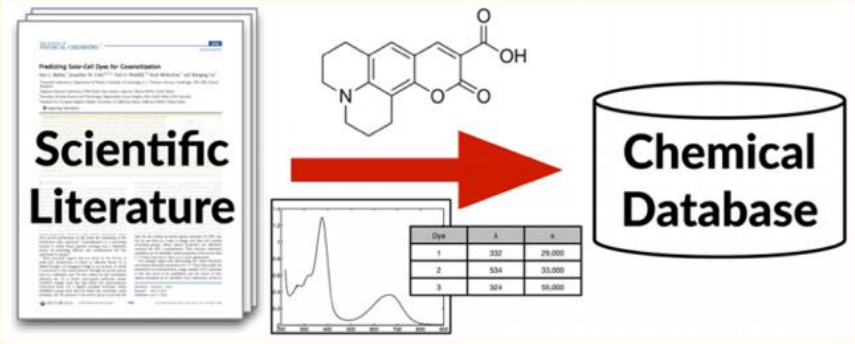
pubs.acs.org/jcim

ChemDataExtractor: A Toolkit for Automated Extraction of Chemical Information from the Scientific Literature

Matthew C. Swain and Jacqueline M. Cole*

Cavendish Laboratory, University of Cambridge, J. J. Thomson Avenue, Cambridge, CB3 0HE, U.K.

"(A) toolkit for the **automated extraction of chemical entities and their associated properties, measurements, and relationships from scientific documents** that can be used to populate structured chemical databases."



Characteristics of note;

- chemistry-aware natural language processing pipeline
  - trained by unsupervised word clustering in a "massive corpus of chemistry articles"
- document-level processing to resolve data interdependencies
- The performance of the toolkit to correctly extract various types of data was evaluated
- released under the MIT license and are available to download

JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

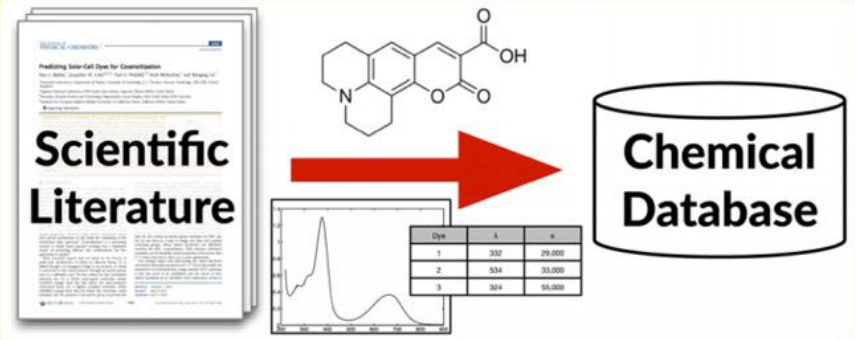
pubs.acs.org/jcim

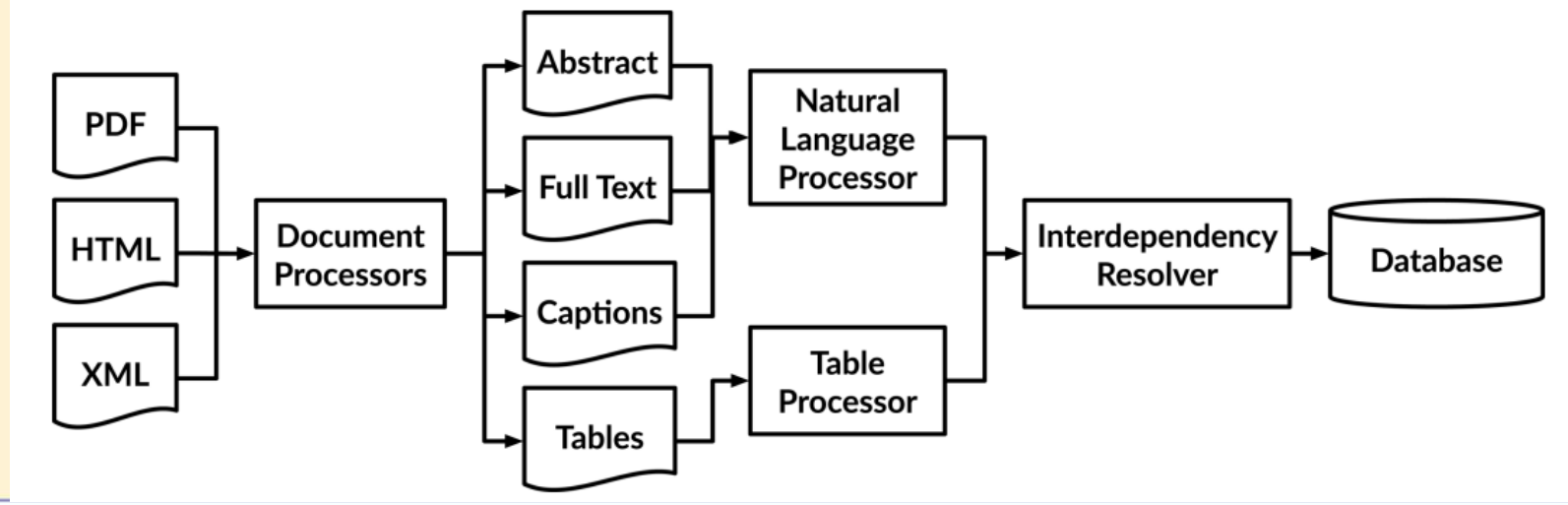
ChemDataExtractor: A Toolkit for Automated Extraction of Chemical Information from the Scientific Literature

Matthew C. Swain and Jacqueline M. Cole*

Cavendish Laboratory, University of Cambridge, J. J. Thomson Avenue, Cambridge, CB3 0HE, U.K.

"(A) toolkit for the **automated extraction of chemical entities and their associated properties**, **measurements, and relationships from scientific documents** that can be used to populate structured chemical databases."





Output from processed document

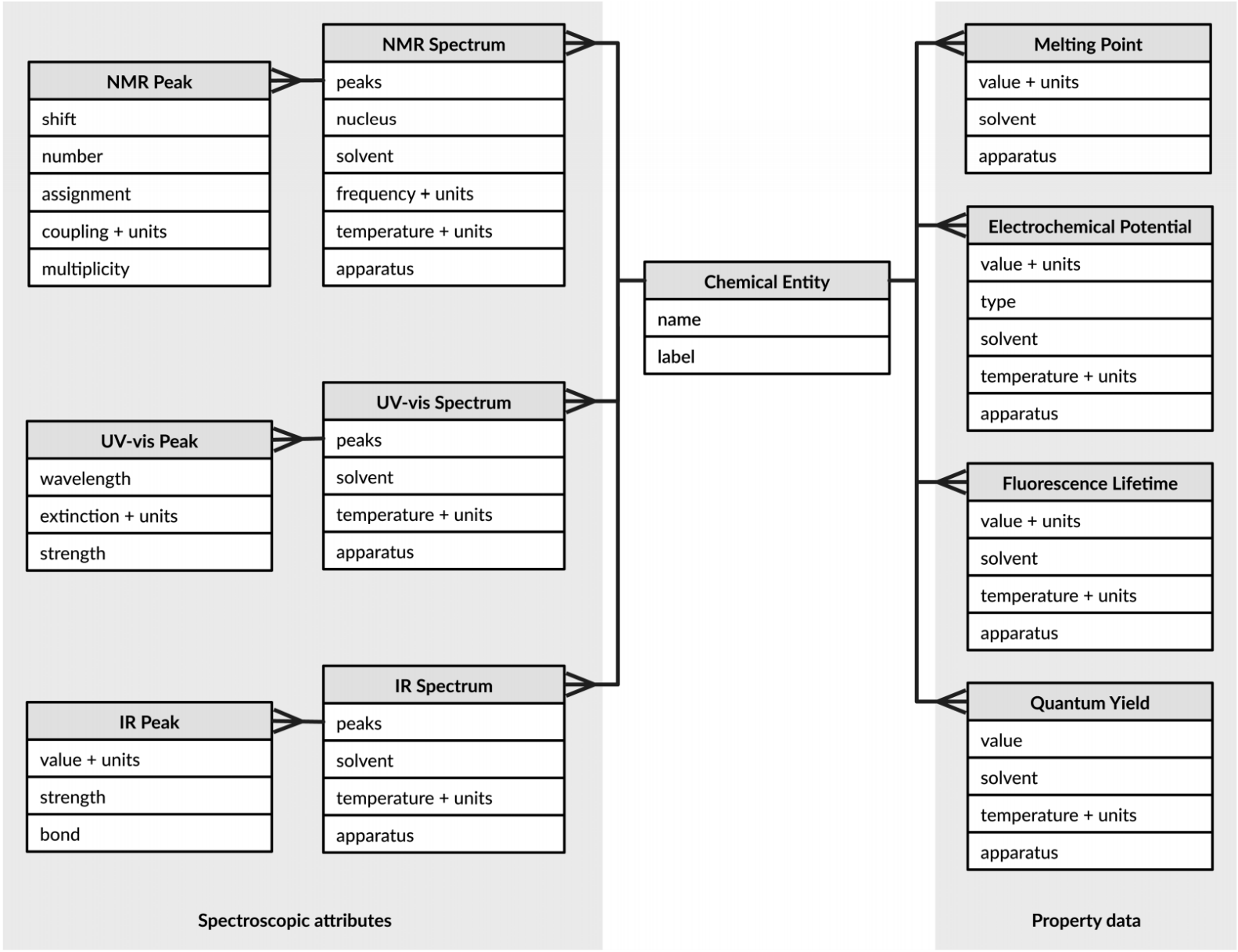


Figure 8. Data model for extracted chemical entities and their associated experimental properties and spectroscopic attributes, as currently provided by ChemDataExtractor. Users of the toolkit may extend this data model by defining their own custom parsers.

ChemDataExtractor Documentation

ChemDataExtractor is a toolkit for automatically extracting chemical information from scientific documents.

This guide provides a quick tour through ChemDataExtractor concepts and functionality.

Features

- HTML, XML and PDF document readers
- Chemistry-aware natural language processing pipeline
- Chemical named entity recognition
- Rule-based parsing grammars for property and spectra extraction
- Table parser for extracting tabulated data
- Document processing to resolve data interdependencies

Citing

If you use ChemDataExtractor as a resource in your research, please cite the following work:

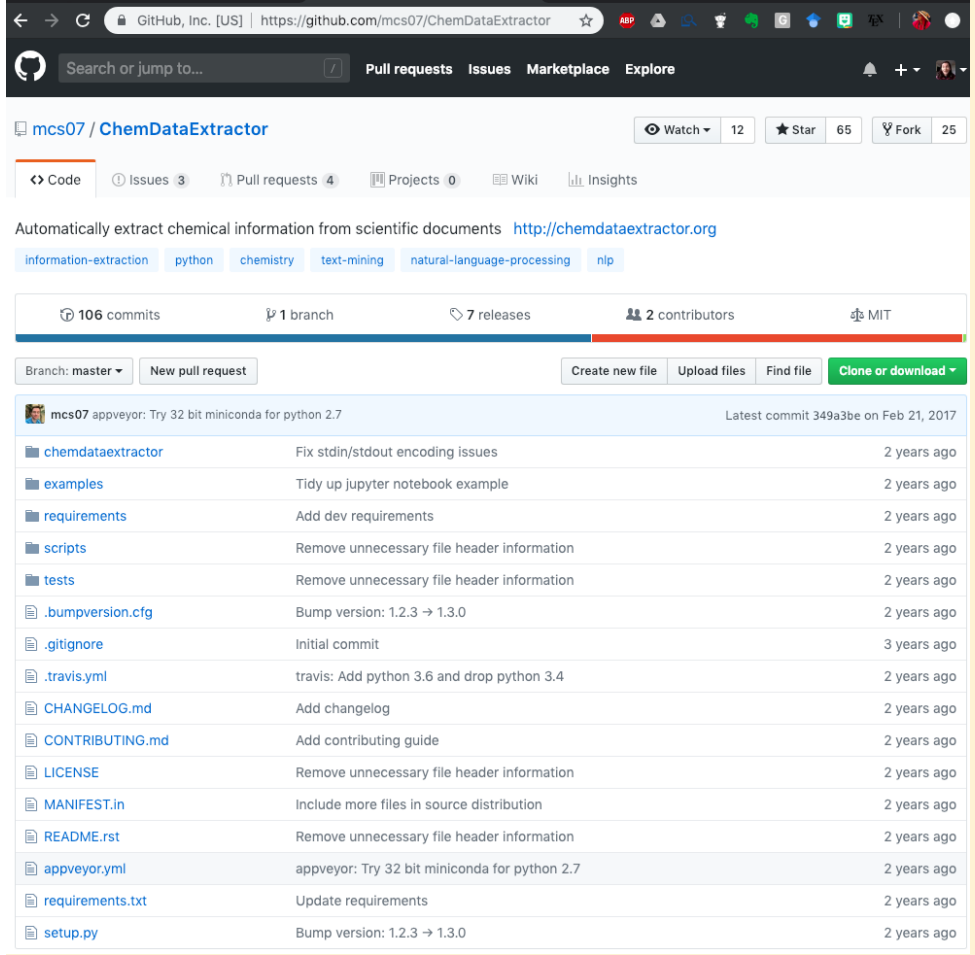
> Swain, M. C., & Cole, J. M. "ChemDataExtractor: A Toolkit for Automated Extraction of Chemical Information from the Scientific Literature", J. Chem. Inf. Model. **2016**, *56* (10), pp 1894–1904 10.1021/acs.jcim.6b00207

This project has been funded by the UK Engineering and Physical Sciences Research Council (EPSRC).

Installation »

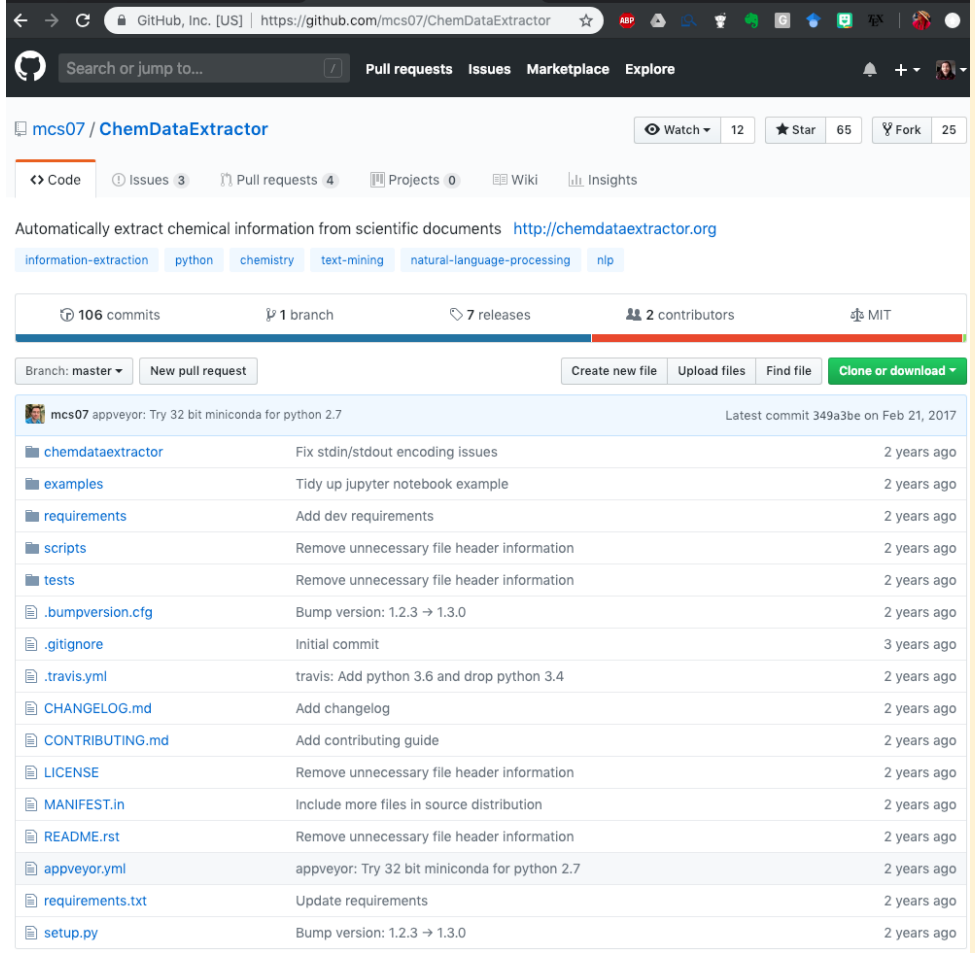
Table of Contents

Introduction
Installation
Getting Started
Reading Documents
Chemical Records
Tokenization
Part-of-speech Tagging
Chemical Named Entities
Lexicon
Abbreviation Detection
Command Line Interface
Scraping Structured Data
Contributing



Sounds great! But should we use it?

- Clean GitHub repo.
  - Issues: 3 open / 6 closed. stagnant since Jan 2018
  - Pull requests: 4 open / 10 closed.



Sounds great! But should we use it?

- Clean GitHub repo.
  - Issues: 3 open / 6 closed. stagnant since Jan 2018
  - Pull requests: 4 open / 10 closed.
- Written in python, compatible with python 3
- Jupyter notebook examples

Performance

- F-scores: 93.4% (chemical identifier extraction), 86.8% (spectroscopic attribute extraction), and 91.5% (chemical property attributes)
  - CHEMDNER chemical name extraction challenge: F-score of 87.8%, vs. high scores of 87-88%