PaperParser: Text Mining for Solar Cell Literature

Christine Chang[1], Harrison Goldwyn[2], Linnette Teo[3], Neel Shah[3]

[1]Department of Materials Science and Engineering, University of Washington, Seattle, WA; [2]Department of Chemistry, University of Washington, Seattle, WA;
[3]Department of Chemical Engineering, University of Washington, Seattle, WA

paper-parser/
paper-parser

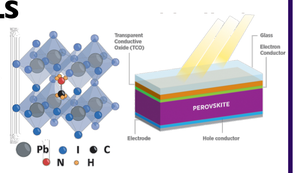
Introduction

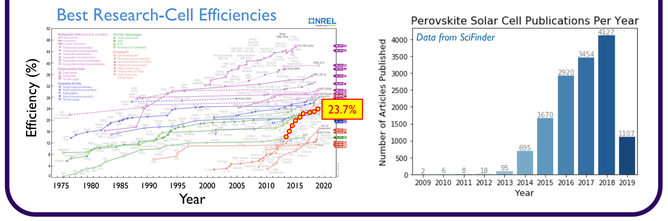
OVERVIEW

As research in a given field progresses, and the volume of literature increases accordingly, scientific advances are hindered by the difficulty of information sharing. Currently, researchers must manually read literature to extract key insights and design improvements. ***PaperParser*** is a package designed to automate this process.

PEROVSKITE SOLAR CELLS

Perovskites are a high-performance next-gen solar cell material. However, with enormous advances in performance come enormous increases in literature volume.

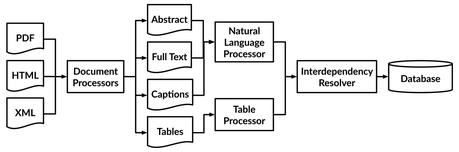




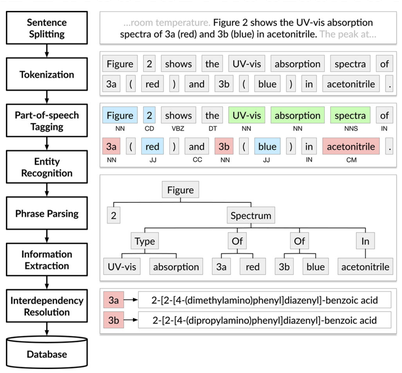
ChemDataExtractor

PaperParser is built on ChemDataExtractor, an open-source software package that extracts chemical information from scientific literature using pre-trained models

OVERVIEW

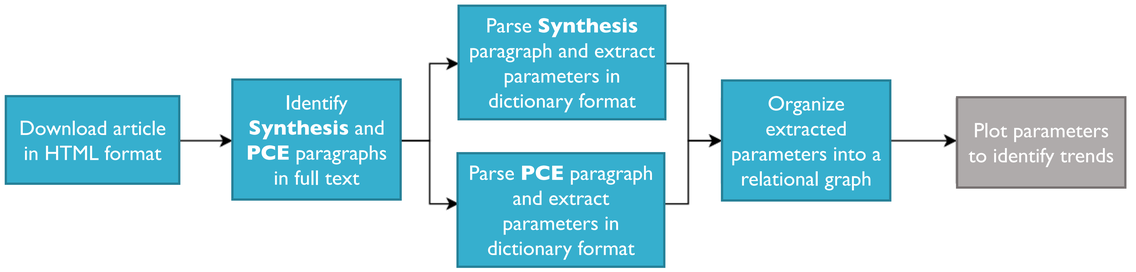


NATURAL LANGUAGE PROCESSING PIPELINE



Package Design

PACKAGE FLOWCHART



IDENTIFYING SENTENCES

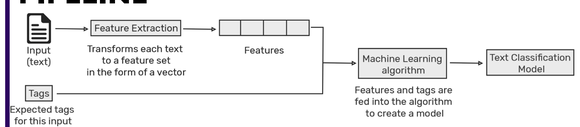
SPACY

spaCy is a powerful and industrial strength package for almost all NLP tasks. Using spaCy as a pre-processing tool to remove punctuations, stopwords and stemming words to root forms improves the accuracy of our classification model.

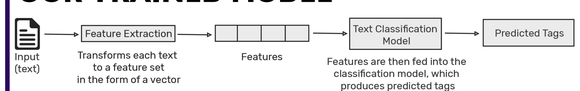
SUPPORT VECTOR MACHINE (SVM)

An SVM is a machine learning algorithm for text classification. Unlike other text classification models, SVM doesn't need much training data for accurate results. Although it needs more computational resources than Naive Bayes, SVM can achieve higher accuracy.

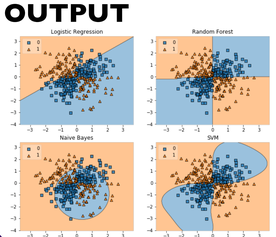
PIPELINE



OUR TRAINED MODEL

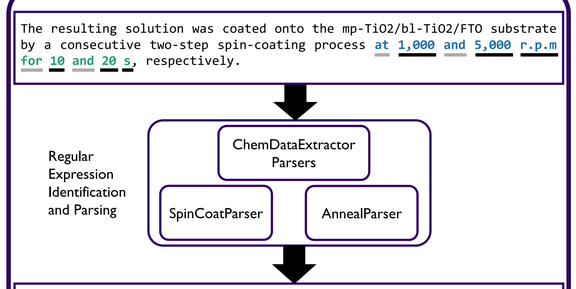


OUTPUT

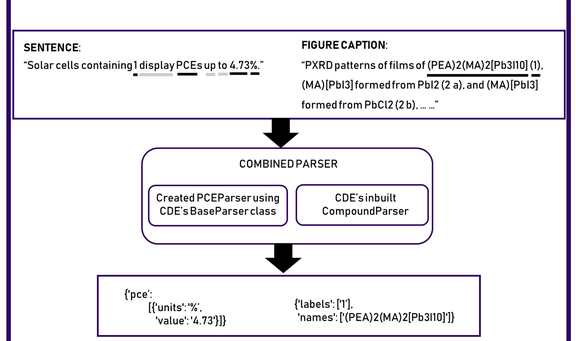


Classifier	Accuracy
Logistic Regression	0.9835
Random Forest	0.9835
Naive Bayes	0.9247
SVM-SVC	0.9858

SYNTHESIS EXTRACTION

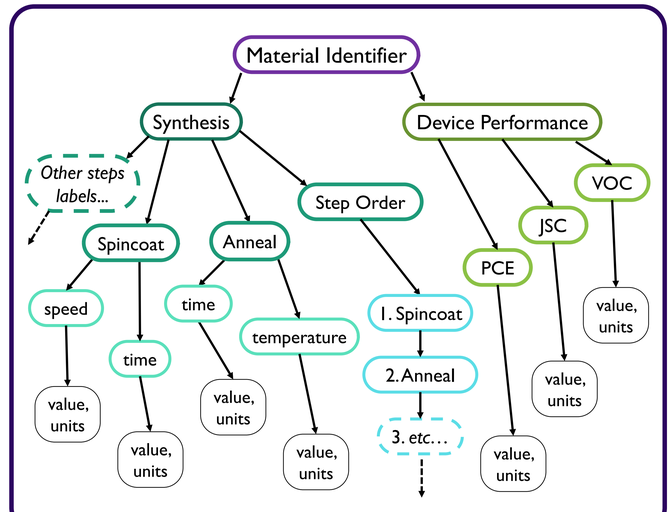


PERFORMANCE EXTRACTION



Results

OUTPUT DATA TREE



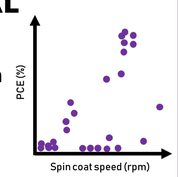
Conclusions and Future Work

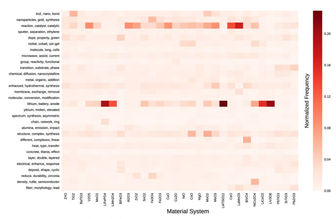
POSSIBLE IMPROVEMENTS

- Increase training dataset for SVC model to improve accuracy
- Extend extraction methods to include more synthesis actions and performance metrics

OUR BIGGER PICTURE END GOAL

- Integrate publisher APIs to download large number of papers based on user input search queries
- Generate a large data set and identify trends





Olivetti group example: Heatmap of topics across material systems from mining the synthesis parameters of oxide materials

References

1. Perovskite solar cells: materials and devices. *United States Department of Energy (DoE)*.
2. Best Research Cell Performance Chart. *National Renewable Energy Laboratory (NREL)*.
3. Swain, M.C. and Cole, J.M., 2016. *ChemDataExtractor: a toolkit for automated extraction of chemical information from the scientific literature*. Journal of chemical information and modeling, 56(10), pp.1894-1904.
4. Kim, E., Huang, K., Tomala, A., Matthews, S., Strubell, E., Saunders, A., McCallum, A. and Olivetti, E., 2017. *Machine-learned and codified synthesis parameters of oxide materials*. Scientific data, 4, p.170127.

Acknowledgments

The authors acknowledge the University of Washington DIRECT program, including instructional staff (Dr. David Beck, Theodore Cohen, Chad Curtis, Torin Stetina, and Caitlyn Wolf) for support. Only open-source packages were used in this work; all documentation can be found on our Github repository (github.com/paper-parser/paper-parser/).

UNIVERSITY of WASHINGTON

CLEAN ENERGY INSTITUTE

DIRECT Data Intensive Research Enabling Clean Technologies