IRI Presentation May 2018

Name: Avirudh Theraja
Research Topic: Metadata Extraction from Scientific Publications

previous

next

Research Areas

- Parsing and restructuring PDF documents into structured XML and text formats
- Extracting useful information from these formats using Machine Learning & Text Extraction techniques

Main Ideas

- Build an extraction pipeline to support all scientific publications, primary focus on software related publications
- Metadata found on journal websites is inadequate; publications are too dense to read
- Use principles of Machine Learning and Information Extraction to enhance the metadata

Research Process

- Coded entire pipeline using the Python programming language
- Researched into various machine learning models for information extraction, finalized open source library called GROBID
- Added support for institutions, grant information, source code repository data etc.

Next Steps

- What more metadata can we fetch? Categorize papers into domains based on the content
- Expand focus to non software related publications, generate metadata based on the domain of the publication
- Curate this metadata to build a website for students and researchers

Pipeline Steps

01

Download the publication

PDF file for most publications can be downloaded at their journal's website.

02

Restructure document

Convert PDF file into standard XML and text formats.

03

Extract information

Using machine learning via 3rd party software and libraries, extract useful metadata.

04

Save metadata in JSON format

Use JSON (Javascript Object Notation) so data can be easily stored in a database or shown on a webpage.

< previous

next >

What metadata can it fetch?

- Title, Summary, Abstract & Keywords
- Institution & Grant/Funding Information
- Authors & Acknowledgements
- Links including source code links
- Repository data from Github & SourceForge
- Technologies used & Number of citations

< previous

next >

Research Areas

- Parsing and restructuring PDF documents into structured XML and text formats
- Extracting useful information from these formats using Machine Learning & Text Extraction techniques

Main Ideas

- Build an extraction pipeline to support all scientific publications, primary focus on software related publications
- Metadata found on journal websites is inadequate; publications are too dense to read
- Use principles of Machine Learning and Information Extraction to enhance the metadata

Research Process

- Coded entire pipeline using the Python programming language
- Researched into various machine learning models for information extraction, finalized open source library called GROBID
- Added support for institutions, grant information, source code repository data etc.

Next Steps

- What more metadata can we fetch? Categorize papers into domains based on the content
- Expand focus to non software related publications, generate metadata based on the domain of the publication
- Curate this metadata to build a website for students and researchers

Thank you!