XN Rising Project
Spring 2022

Northeastern University

Brief: An American multinational pharmaceutical and biotechnology corporation headquartered in Manhattan, New York City is looking to use Natural Language Processing/ Machine Learning for accurate parsing.

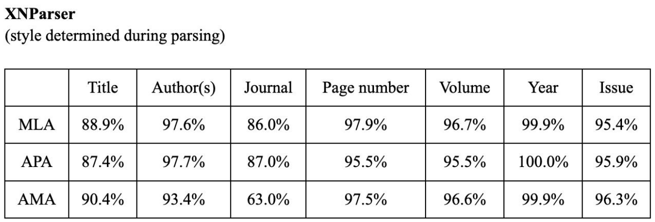
Background: Every publication has a list of references at the end that contain cited facts. There are various software tools like EndNote that help authors craft these formatted references. However, the reverse process of taking a formatted reference and parsing out the authors, title, year, publication, etc. remains a troublesome task. There have been a few different tools that sometimes work, but few achieve a reasonable accuracy. Why does this matter? There is a lot of interest in leveraging literature these days for many purposes. Some are trying to automatically verify facts (think fake news). Network analysis that connects authors and papers is increasing rapidly to identify field boundaries, key papers, identify trends, etc. Others are focused on finding collaborators, scientific funding, ranking schools, business ventures, etc. Many projects would benefit from an accurate parsing of these references.

Result: We researched, implemented, and tested four top pre-existing parsing technologies, and chose Cermine as a technology for testing, benchmarking, and as a safety net when encountering a citation style that doesn't fit in any citation standard. Our API development has been object-oriented and test-driven: we executed thousands of tests on raw references and ground truth from PubMed before implementing the actual methods to the library. Our technology contains all industry standard protocols, highly detailed documentation, and the procedures and tools requested in the project description. Given the Sponsor's preference, it's written in Python language.
The technology works as follows:

- It takes in a bibliographical reference of any style as input
- It determines what citation style it is (ama, apa, mla, etc.)
- If it's a popular citation style (ama, apa, mla), it parses the reference using our regular expression scripts
- If it's not a popular citation style, or if regex is unable to parse the reference, it parses the reference using CERMINE
- It provides methods to return individual fields of the parsed reference, namely: author, title, journal, volume, year, etc.

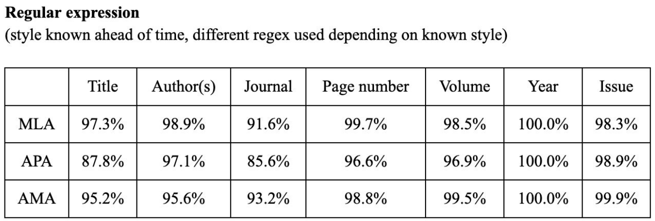
Title: Using NLP/Machine Learning for Bibliography Parsing
Sponsor: Pfizer
Team: Andrew Boylan, Juan Ceruli, Thai Khanh Huynh, Tzu-Yu (Jamie) Huang, Yutong He

Accuracy Test Result:
Tests were run on references in three styles (MLA, APA, AMA) from the first 30K references in PubMed's annual baseline dataset. With our combined parser technique approach, we were able to improve on the results of using CERMINE alone in almost every category.



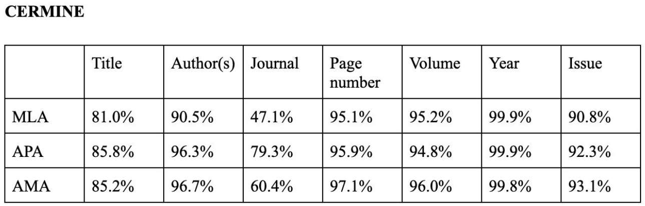
XNParser
(style determined during parsing)

	Title	Author(s)	Journal	Page number	Volume	Year	Issue
MLA	88.9%	97.6%	86.0%	97.9%	96.7%	99.9%	95.4%
APA	87.4%	97.7%	87.0%	95.5%	95.5%	100.0%	95.9%
AMA	90.4%	93.4%	63.0%	97.5%	96.6%	99.9%	96.3%



Regular expression
(style known ahead of time, different regex used depending on known style)

	Title	Author(s)	Journal	Page number	Volume	Year	Issue
MLA	97.3%	98.9%	91.6%	99.7%	98.5%	100.0%	98.3%
APA	87.8%	97.1%	85.6%	96.6%	96.9%	100.0%	98.9%
AMA	95.2%	95.6%	93.2%	98.8%	99.5%	100.0%	99.9%



CERMINE

	Title	Author(s)	Journal	Page number	Volume	Year	Issue
MLA	81.0%	90.5%	47.1%	95.1%	95.2%	99.9%	90.8%
APA	85.8%	96.3%	79.3%	95.9%	94.8%	99.9%	92.3%
AMA	85.2%	96.7%	60.4%	97.1%	96.0%	99.8%	93.1%