Prepare for Class 04

John Glendenning

Prepare for Class

1. Assigned reading (case study)

- Read the articles in the special volumes:
  - "Mining Scientific Papers: NLP-enhanced Bibliometrics"
  - "Mining Scientific Papers, Volume II: Knowledge Discovery and Data Exploitation"
- Summarize what you find (it could be out of your domain knowledge but just give your general impression)
- Write three questions from the assigned reading for class discussion. For example:
  - "What is knowledge mining (before reading)?" vs. "What is knowledge mining (after)?"
- Write a presentation (e.g. Powerpoint, Beamer) to summarize your findings and invite discussions in class.

Very Short Summaries, Volume I-1

“Editorial: Mining Scientific Papers: NLP-enhanced Bibliometrics”

This editorial introduces the concept of NLP-enhance Bibliometrics. It gives an overview of the tools and techniques introduced in the seven other papers in this series. An emphasis is given to the full-text analysis to enhance metadata models.

Very Short Summaries, Volume I-2

"The NLP4NLP Corpus (I): 50 Years of Publication, Collaboration, and Citation in Speech and Language Processing"

This research introduces the NLP4NLP corpus. This corpus is an NLP analysis of the NLP literature. A dataset of about 65,000 over 50 years (1965-2015) analyzed trends in research collaboration, citation patterns, and the evolution of research topics.

While the results were interesting, the conclusions pointed out that this research would not be possible without access to the underlying writings. This can only be achieved with open access and machine-readable copy.

"The NLP4NLP Corpus (II): 50 Years of Research in Speech and Language Processing"

This is a follow up paper to the previous one. This paper examined how different techniques can be applied to entire bodies of work. It addressed some issues about "consistent and uniform identification of entities", which has made it troublesome to properly connect different pieces of research, and identify "self plagirism."

Very Short Summaries, Volume I-4

“Resolving Citation Links With Neural Networks”

This paper explores the use of neural networks to identify citation links. It discusses using both a triplet model and a binary model to identify the linkages. The paper explored some of the differences in the methods and identified areas of better / worse performance.

Very Short Summaries, Volume I-5

"Temporal Representations of Citations for Understanding the Changing Roles of Scientific Publications"

This study looked at the citations over time. The experimental was built around identifying citations, and then evaluating the statistics about those citations over time. Since they are using papers from Pub Med Central which are written using MeSH (Medical Subject Headings), there may have been a more controlled use of language to facilitate the analysis.

This paper developed a methodology that could be extended to other corpora in other fields to examine the citations of papers over time.

Very Short Summaries, Volume I-6

"Deep Reference Mining From Scholarly Literature in the Arts and Humanities"

This paper has developed methods to handle bibliographic references including "detection, extraction, and classification." The paper claims that at that time, the techniques in use were from the previous decade.

A model was developed to extract and classify bibliographic references anywhere they exist in a source from text to footnotes to endnotes, etc. In the conclusion, they discussed some of the issues dealing with older humanities papers including faulty OCR and other inconsistencies. Because of this, they recommend that future activities be supervised or semi-supervised.

"The Termolator: Terminology Recognition Based on Chunking, Statistical and Search-Based Scores"

This paper discussed Termolator, a terminology extraction tool. It combines both a knowledge-based approach with statistical analysis to achieve the best results. This model works best with many documents, and not as well with single documents in isolation.

Termolator is compared with Termostat, the only other tool at the time that can perform the same functions and is readily available to researchers. Termolator seems to have performed better, especially when working with domain-specific terms.

Very Short Summaries, Volume I-8

"Is the Abstract a Mere Teaser? Evaluating Generosity of Article Abstracts in the Environmental Sciences"

"The main objective of this paper is to introduce a new metric called GEM to measure the "generosity" or representativeness of an abstract. Schematically speaking, a generous abstract should have the best possible score of similarity for the sections important to the reader."

There were differences in GEM scores between publishers as well as between subject areas. Later works have had increasing GEM scores, meaning that they are more informative about the work and less "teasers".

Very Short Summaries, Volume II-1

"NLP4NLP+5: The Deep (R)evolution in Speech and Language Processing"

Published in 2022, this paper analyzed trends in speech and NLP research from 2016-2020. The corpus now includes nearly 100,000 documents covering 55 years with nearly 70K authors. The results of this study have reinforced the previous two studys in the last group of papers.

This study still had problems with the identification of entities because of a lack of similar structures between publications. This paper continued its warning on reuse or plagirism stating that manual checking is required. According to the data and charts, reuse peaked about about 25% in 2010 and has been decreasing.

Very Short Summaries, Volume II-2

"Editorial: Mining Scientific Papers, Volume II: Knowledge Discovery and Data Exploitation"

This is the second paper of "Mining Scientific Papers." It discussed the other 6 papers in Volume II.

It goes through items like tasks, corpus, objects and methods of the other papers in this series.

Very Short Summaries, Volume II-3

"Visual Summary Identification from Scientific Publications via Self-Supervised Learning"

This paper outlined a methodology for identification of Graphical Abstracts from scientific papers utilizing supervised learning. This was performed on copora from both PubMed (PMC) and their own computer science dataset. For PMC, a single figure was associated with the abstract as representative. With the CS dataset, three were used.

This self-supervised approach was effective at identifying the primary figure that represented the paper.

Very Short Summaries, Volume II-4

“SYMBALS: A Systematic Review Methodology Blending Active Learning and Snowballing”

In this paper, a methodology called SYMBALS is proposed, which combines the traditional technique of backward snowballing with automated literature screening. This allowed for a 6 times increase in title and abstract screening.

When compared to FAST2, another similar method, it was faster by 6%.

"Enhancing Knowledge Graph Extraction and Validation from Scholarly Publications Using Bibliographic Metadata"

Bibliometric-Enhanced Information Retrieval (BIR) can used metadata to enhance the retrieval of knowledge graphs from scholarly publications. The title, abstract, controlled keywords and other textual clues can be used to create the knowledge graph to extract information from publications.

The paper proposed the further use of AI and machine learning to create these knowledge graphs.

Very Short Summaries, Volume II-6

"Large Scale Subject Category Classification of Scholarly Papers with Deep Attentive Neural Networks"

This paper proposed a Deep Attentive Neural Network (DANN) for classifying papers into subject categories by using only their abstracts. The team classified nine million abstracts into 104 different groups. The median F1s were about 0.75 under the best settings.

The primary issue in classification was that the categories were not necessarily mutually exclusive. Future work will attempt to classify into a multilevel schema instead of the flat 104 that was used.

Very Short Summaries, Volume II-7

“Language Bias in Health Research: External Factors That Influence Latent Language Patterns”

According to the paper, the “language used in scientific reporting can misrepresent research findings in a manner analogous to falsifying or incorrectly analyzing numeric data.” This paper has attempted to analyze the language and framing in scientific papers, focusing on health research.

While this paper showed the change of language over time, and brought up some interesting questions, it was very light on conclusions.

Questions & Expected Answers

1. Can language be studied over time? How?

- Yes, there are tools to do this.

2. Can RDF or semantic knowledge graphs be handled by NLP?

- Hopefully there is some methodology to parse RDF graphs.

3. Can images be part of the knowledge graph of a paper?

- I expect that rudimentary

Questions & Realized Answers

1. Can language be studied over time? How?

- The corpus has to be manually separated or grouped before analysis.

2. Can RDF or semantic knowledge graphs be handled by NLP?

- Nope, SPARQL, then analysis.

3. Can images be part of the knowledge graph of a paper?

- Not really, but the primary image associated with the abstract can reliably be extracted.

Project

Statement of Purpose

Review the literature over time to see if there was a change in sentiment about the structure of the Roman Economy.

From Today's Readings

- Review the bibiliography from "Language Bias in Health Research: External Factors That Influence Latent Language Patterns"
- Can the techniques from any of the papers be applied to this?
- What are the optimal methods for extracting text from images? Tesseract in R?
- How can a corpus be built using snowballing or similar techniques? McDermott?