Mining Scientific Papers

Knowledge Mining – Prepare for Class 4 & 5

Summarizing Volume I

- NLP to be used for large-scale analysis of texts and citations, as well as structures of scientific papers, descriptions about authors and more
- Recent papers have more detailed abstracts although there is no correlation between the length of abstracts and how frequently papers are cited
- Termolator tool used for detecting specialized key-words
- Neural network models allow for approximating citation link targets by looking at the area around the true target than its exact location
- Various challenges due to papers lacking metadata, different languages, inconsistent author identification, different formats, etc.

Summarizing Volume II

- Some techniques that can be used to help categorize papers
  - Deep attentive neural network (DANN) to train based on abstracts
  - SYMBALS to use backward snowballing with active learning
  - Self-supervised learning approach that matches texts with figures
- Need for international collaboration to help standardize the process to make it easier for the algorithms to sort out the papers

Three questions for class discussion

- What are the international standards that everyone must follow and who gets to set them?
- When using NLP to analyze all the scientific papers, do we also take data privacy into consideration? And who is doing the analyzing?
- Can NLP help identify bias or misinformation in those scientific papers?