CS 6460: Project Proposal
Patrick Miller (pmiller42)

Project Proposal

I will use NLP to develop a concept map from existing textbooks. This will be a development project that starts with building a prototype using just a few textbooks in specific subjects, and if successful, will then expand into the larger task of using hundreds of textbooks to represent the knowledge links contained within and across them. This sort of knowledge graph can be used in Learning Management Systems and by Intelligent Tutors to create prerequisite readings and adjust learning plans for different types of students.

Knowledge graphs seek to build networks of information that are structured around the relationships between the entities that it identifies. Concept maps are a subset of knowledge graphs that incorporate a directional structure. This is particularly useful in education, as many concepts exist in a prerequisite structure.

Knowledge graphs and concept maps have generally been built by subject matter experts or through crowd-sourcing (e.g. Wikipedia, Khan Academy and Metacademy). Furthermore, textbooks have generally not been included in these concept maps due to copyright issues. My goal in this project is to create an automated framework for concept extraction and linking through text processing on textbooks.

Existing Solutions

There are a few research projects aimed at using NLP to link textbooks together. Guerra et al. [1] examines linking multiple textbooks using probabilistic topic models. Their approach is limited because their exploration study looks at a total of 9 textbooks in 2 categories. From their experiment they conclude that LDA should be able to successfully link textbooks.

In Wang et al. [5 and 6], they present the idea of using Wikipedia as a base of knowledge for extracting concepts from textbooks. They build a concept hierarchy by considering both “local relatedness” and “global coherence”, which examines both chapter by chapter concepts and the similarity of concepts throughout. Huang et al. [3] extend on this approach through a variety of different models and combine student learning patterns to analyze personalized learning. Meng et al. [4] use a semantic-based approach to knowledge component extraction from textbooks instead of just relying on key terms.

This project seeks to combine the best attributes of each of these approaches, and leverage the author’s access to a large corpus of textbooks at Macmillan Learning.

Product

The concept map extraction tool will be a NLP pipeline that takes a list of textbook resources (PDFs and ePubs) and outputs a concept map that links concepts included in these textbooks in a prerequisite directed graph structure. The visualization of this concept map is TBD, but it will likely exist in multiple forms:

- A directed graph of concepts
- A list of major concepts contained within each textbook
- A graph of textbooks, detailing how closely they are related to each other

Tools and Resources

I will be using Python to build the concept map extraction tool. The major NLP-related packages within Python that I will use are SpaCy and scikit-learn. SpaCy deals with large-scale natural language processing and links up well with scikit-learn. Scikit-learn is a mature set of libraries for a variety of machine learning tasks, including topic modeling. I will attempt to link the topics from the textbooks to DBpedia using the DBpedia Spotlight API. DBpedia extracts structured content from Wikipedia and provides it to users through an API. If utilizing DBpedia is not successful, I will resort to using the unsupervised topics discovered by LDA, and will manually classify them for the purposes of visualization.

In order to process the textbook files, which are mostly PDFs, I will likely use pdfminer. For ePub files, I will use the epub module in Python. I have not yet decided on how I will visualize the final concept map, but there are many options out there.

Work Completed

- Initial subjects have been chosen: Psychology and Biology.
- Textbooks have been selected and downloaded in PDF and/or ePub formats:

title	author	imprint
INTRODUCING PSYCHOLOGY	DANIEL L SCHACTER	FREEMAN/WORTH
LIFE: THE SCIENCE OF BIOLOGY	DAVID E SADAVA	FREEMAN/WORTH
EXPLORING PSYCHOLOGY	DAVID G MYERS	WORTH PUBLISHERS
PSYCHOLOGY	DAVID G MYERS	FREEMAN/WORTH
PSYCHOLOGY IN EVERYDAY LIFE	DAVID G MYERS	WORTH PUBLISHERS
BIOLOGY: HOW LIFE WORKS, VOLUME 2	JAMES R MORRIS	W.H. FREEMAN
WHAT IS LIFE? A GUIDE TO BIOLOGY	JAY PHELAN	FREEMAN/WORTH
MOLECULAR BIOLOGY: PRINCIPLES AND PRACTICE	MICHAEL M COX	W. H. FREEMAN
BIOLOGY OF PLANTS	PETER H RAVEN	FREEMAN/WORTH
PSYCHOLOGY: A CONCISE INTRODUCTION	RICHARD A GRIGGS	WORTH PUBLISHERS
ABNORMAL PSYCHOLOGY	RONALD J COMER	WORTH PUBLISHERS

- I created the development environment with many of the software tools I plan to use: Docker, Python, SpaCy, scikit-learn. I still need to figure out the Python package I will use to parse PDFs.
- I completed my research on directed concept maps, and have concluded that I should try to incorporate a very simple version of prerequisite structure in my knowledge graph. Anything complex would be out of the scope of this project.

Plan

Intermediate Milestone – 7/2/17

The first 2 weeks will be dedicated to parsing the textbooks and making sure that the text is in a form that is readily processed through NLP methods.

Tasks

- Finish doing research into the existing solutions and consolidate knowledge on their technical approaches
- Assemble existing code bases on GitHub for using DBpedia to extract knowledge
- Learn pdfminer and write code to parse textbooks in PDF form
- Extract section headers and table of contents, if possible
- Perform basic NLP using SpaCy

Deliverables

- Proof that the textbooks have been successfully parsed: number of words, most common words
- Most common words (entities?) by textbook
- Structure of each textbook as parsed by the pipeline (table of contents or chapter headings) [optional]

Intermediate Milestone – 7/16/17

The following 2 weeks will focus on performing NLP modeling on the already processed text. If it is possible, I will also integrate DBpedia into the pipeline for a more structured approached to identifying concepts.

Tasks

- Build a topic modeler using LDA or NMF to extract topics for entire textbooks
- Build a topic modeler using LDA or NMF to extract topics for individual chapters
- Use DBpedia Spotlight to integrate textbook topics with DBpedia linked data

Deliverables

- Most common LDA topics by textbook (by chapter)
- Most common DBpedia concepts by textbook
- Linkage of most common topics to DBpedia concepts by textbook

Final Project – 7/30/17

The final 2 weeks will be focused on finishing up the NLP work and creating the visualizations. The key here will be creating a heuristic for linking topics in the textbooks in a directed prerequisite structure.

Tasks

- Create links between topics in a prerequisite structure using textbook progression
- Visualize the concept map and textbook relation structure

Deliverables

- Visualization of concept map
- Graph of textbooks
- Final paper

References

[1] Guerra, Julio et al. "When One Textbook Is Not Enough: Linking Multiple Textbooks Using Probabilistic Topic Models." EC-TEL 2013.
[2] Weikum, Gerhard and Marin Theobald. "From Information to Knowledge: Harvesting Entities and Relationships from Web Sources." PODS 2010.
[3] Huang, Yun et al. "A Framework for Dynamic Knowledge Modeling in Textbook-Based Learning." DOI 2016. ACM 2016.
[4] Meng, Rui et al. "Knowledge-Based Content Linking for Online Textbooks." IEEE/WIC/ACM 2016.
[5] Shuting Wang, Chen Liang, Zhaohui Wu, Kyle Williams, Bart Pursel, Benjamin Brautigam, Sherwyn Saul, Hannah Williams, Kyle Bowen, and C. Lee Giles. "Concept Hierarchy Extraction from Textbooks." ACM Symposium on Document Engineering 2015.
[6] Wang, Shuting; Ororbia, Alexander; Wu, Zhaohui; Williams, Kyle; Liang, Chen; Pursel, Bart; and C. Lee Giles. "Using Prerequisites to Extract Concept Maps from Textbooks." CIRM 2016.
[7] R. Agrawal, S. Gollapudi, A. Kannan, and K. Kenthapadi. Data mining for improving textbooks. ACM SIGKDD Explorations Newsletter, 13(2):7–19, 2012.
[8] R. Mihalcea and A. Csomai. Wikify!: linking documents to encyclopedic knowledge. In CIKM, pages 233–242. ACM, 2007.
[9] L. Ratinov, D. Roth, D. Downey, and M. Anderson. Local and global algorithms for disambiguation to wikipedia. pages 1375–1384, 2011
[10] Pedregosa et al. "Scikit-learn: Machine Learning in Python." JMLR 12 2011.
[11] Matt Honnibal. "Introducing Spacy." https://explosion.ai/blog/introducing-spacy.
[12] Joachim Daiber and Max Jakob and Chris Hokamp and Pablo N. Mendes. Improving Efficiency and Accuracy in Multilingual Entity Extraction. Proceedings of the 9th International Conference on Semantic Systems (I-Semantics) 2013.