Topic Analysis of Textbooks Using AI

Coby Simmons

Utrecht University

cobysimmons01@gmail.com

31 January 2024

Aim and Motivation

- focus on the application of AI in identifying and categorising topics in textbooks

Aim and Motivation

- focus on the application of AI in identifying and categorising topics in textbooks
- construct a domain-dependent, fine-grained topic model

Aim and Motivation

- focus on the application of AI in identifying and categorising topics in textbooks
- construct a domain-dependent, fine-grained topic model
- enable enhanced learning experiences via learning object classification, automatic index creation, content personalisation, etc.

Overview

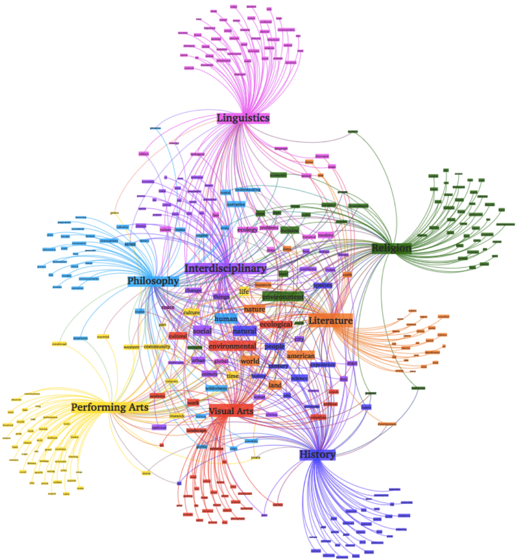
Previous Research

Modelling of textbooks

- proposal for an intelligent online learning experience: "Intextbooks"
  (Alpizar-Chacon, Hart, et al. 2020)
- extraction of knowledge models from PDF textbooks; encoding as TEI[1] files
  (Alpizar-Chacon and Sosnovsky 2020; Alpizar-Chacon and Sosnovsky 2021)
- semantic linking and enrichment of glossary & index
  (Alpizar-Chacon and Sosnovsky 2019; Alpizar-Chacon and Sosnovsky 2022)

[1] `https://tei-c.org/`

Modelling Topics



Term Frequency-Inverse Document Frequency (TF-IDF)

- statistical measure for evaluating the importance of a word in a document in relation to the document's corpus.

Term Frequency-Inverse Document Frequency (TF-IDF)

- statistical measure for evaluating the importance of a word in a document in relation to the document's corpus.
- for a term t in a document d in a corpus of size N, the TF-IDF is

$$tf_{t,d} \times \log \frac{N}{df_t}$$

  where $tf_{t,d}$ is the frequency of t in d and df_t is the number of documents containing t.
- (see Spärck Jones 1972)

Latent Dirichlet Allocation

- represents topics as a probability distribution over a fixed vocabulary; and represents documents as a mixture of topics
  (Blei, Ng, and Jordan 2003)

Latent Dirichlet Allocation

- represents topics as a probability distribution over a fixed vocabulary; and represents documents as a mixture of topics

  (Blei, Ng, and Jordan 2003)

- outperforms Apache Lucene for linking sections of textbooks

  (Guerra, Sosnovsky, and Brusilovsky 2013)

Latent Dirichlet Allocation

- represents topics as a probability distribution over a fixed vocabulary; and represents documents as a mixture of topics

  (Blei, Ng, and Jordan 2003)

- outperforms Apache Lucene for linking sections of textbooks

  (Guerra, Sosnovsky, and Brusilovsky 2013)

- faces a number of challenges:
  - need to set the number of topics as a parameter
  - struggles with topics that are tangential or overlapping

    (Ajinaja et al. 2023)

Embedding Models

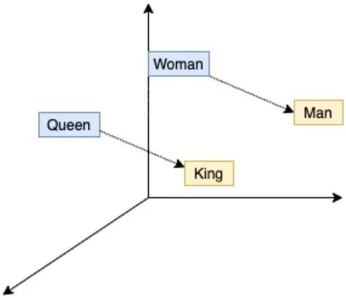
- creates dense and continuous representations of terms and documents in vector spaces

Embedding Models

- creates dense and continuous representations of terms and documents in vector spaces
- Word2Vec captures semantic and syntactic relationships between words (Mikolov et al. 2013)

Embedding Models

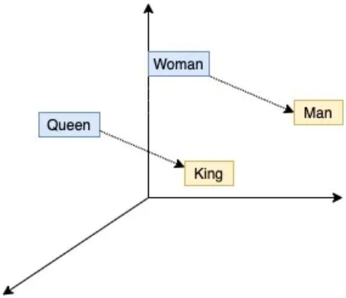
- creates dense and continuous representations of terms and documents in vector spaces
- Word2Vec captures semantic and syntactic relationships between words (Mikolov et al. 2013)



Source: `https://towardsdatascience.com/word2vec-research-paper-explained-205cb7eecc30`

Embedding Models

- creates dense and continuous representations of terms and documents in vector spaces
- Word2Vec captures semantic and syntactic relationships between words (Mikolov et al. 2013)



Source: `https://towardsdatascience.com/word2vec-research-paper-explained-205cb7eecc30`

- Doc2Vec extends Word2Vec to entire documents and outperforms LDA in textbook content linkage tasks (Thaker, Brusilovsky, and He 2018)

Bidirectional Encoder Representations from Transformers (BERT)

- uses two unsupervised tasks for pre-training:
  - masked token prediction
  - next sentence prediction

(Devlin et al. 2019)

Bidirectional Encoder Representations from Transformers (BERT)

- uses two unsupervised tasks for pre-training:
  - masked token prediction
  - next sentence prediction

  (Devlin et al. 2019)

- effectively used for automatic keyword extraction from textbooks

  (Pozzi, Alpizar-Chacon, and Sosnovsky 2023)

Recurrent Neural Networks (RNNs)

- a deep neural network with capacity to remember sequences in the data.

Recurrent Neural Networks (RNNs)

- a deep neural network with capacity to remember sequences in the data.
- can face the ‘vanishing gradient problem’, whereby the gradient becomes too small for effective learning of long-range dependencies

  (Bengio, Simard, and Frasconi 1994)

Recurrent Neural Networks (RNNs)

- a deep neural network with capacity to remember sequences in the data.
- can face the ‘vanishing gradient problem’, whereby the gradient becomes too small for effective learning of long-range dependencies
  (Bengio, Simard, and Frasconi 1994)
- Long Short-Term Memory (LSTM) networks introduced to overcome this limitation
  (Hochreiter and Schmidhuber 1997)

Semantic Similarity

- refers to the degree to which two pieces of text are alike in meaning or content

Semantic Similarity

- refers to the degree to which two pieces of text are alike in meaning or content
- can be measured through cosine similarity S_C, where documents are represented by n-dimensional vectors, $\mathbf{a}$ and $\mathbf{b}$,

$$S_C(\mathbf{a}, \mathbf{b}) := \cos(\theta) = \frac{\mathbf{a} \cdot \mathbf{b}}{\|\mathbf{a}\| \|\mathbf{b}\|}$$

Linking Multiple Textbooks

- **Topic Aggregation:** Only compute the topic vectors for the sections at lowest level in the hierarchy, and for higher levels, aggregate topic vectors by taking a weighted average of sub-topic vectors.

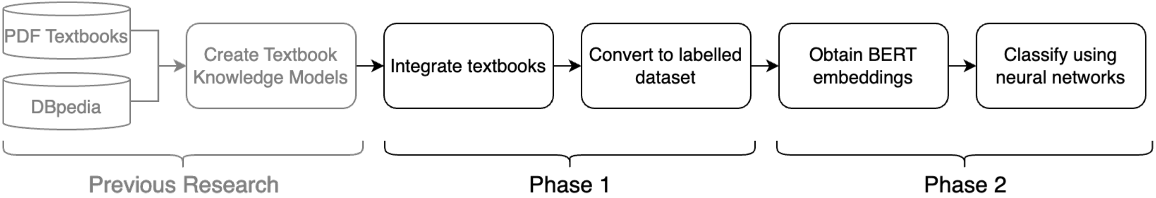
Linking Multiple Textbooks

- **Topic Aggregation:** Only compute the topic vectors for the sections at lowest level in the hierarchy, and for higher levels, aggregate topic vectors by taking a weighted average of sub-topic vectors.
- **Re-Indexing:** Compute the topic vector for every section by considering a section's 'document' to be the content of the section and all it's sub-sections.
- (see Guerra, Sosnovsky, and Brusilovsky 2013)

Methodology

Methodology

Architecture for topic representation and classification



Dataset Generation (Phase 1)

Base Textbook

Other Textbooks

Dataset Generation (Phase 1)

- use multiple attributes for each section
  - header
  - content
  - concept names, definitions, concept subjects
- quick, inexpensive way to generate dataset of sections and their topics

Dataset Generation (Phase 1) – Strategies

- TF-IDF
- Doc2Vec
- Clustering
- Ensemble Modelling
  - combination of TF-IDF and clustering
- TF-IDF & Doc2Vec Hybrid Approach
  - first check for matches using TF-IDF, then use Doc2Vec for uncertain matches
- Iterative Learning
  - recompute section's vector after each match

Topic Classification (Phase 2)

- goal: learn from the generated dataset to classify new content by topic

Topic Classification (Phase 2)

- goal: learn from the generated dataset to classify new content by topic
- input text is preprocessed, tokenised, and pass to DistilBERT to generate vectors

Topic Classification (Phase 2)

- goal: learn from the generated dataset to classify new content by topic
- input text is preprocessed, tokenised, and pass to DistilBERT to generate vectors
- RNNs are trained to classify sections into topics

Evaluation

Evaluation – Dataset Generation (Phase 1)

- evaluate performance using manual mapping generated by experts

Evaluation – Dataset Generation (Phase 1)

- evaluate performance using manual mapping generated by experts
- false positives can have a greater cost than false negatives
  $\Longrightarrow$ precision more important than recall for this task

Evaluation – Dataset Generation (Phase 1)

- evaluate performance using manual mapping generated by experts
- false positives can have a greater cost than false negatives
  $\implies$ precision more important than recall for this task
- therefore, use the F_β score with $\beta = 0.5$

Evaluation – Dataset Generation (Phase 1)

Table: Summary of results for all textbook integration methods

name	precision	recall	F_1	F_β
Hybrid Model	**0.8333**	0.3169	0.4592	**0.6285**
Doc2Vec	0.5714	**0.3944**	**0.4667**	0.5243
TF-IDF	0.5926	0.3380	0.4305	0.5150
Ensemble	0.4632	0.3099	0.3713	0.4215
Clustering	0.0260	0.0282	0.0270	0.0264

more detailed results available at GitHub repository:
`https://github.com/CobySim01/textbook-topic-analysis`

Evaluation – Topic Classification (Phase 2)

- expert dataset
  - 14 class labels
  - 216 data points
- small dataset
  - 32 class labels
  - 352 data points
- large generated dataset
  - 329 class labels
  - 2371 data points

Evaluation – Topic Classification (Phase 2)

Table: Summary of cross-validation performance for each dataset

dataset	concepts	accuracy		precision		recall		F_1	
		model	baseline	model	baseline	model	baseline	model	baseline
expert	true	**0.59**	0.13	**0.67**	0.07	**0.59**	0.13	0.61	0.08
expert	false	0.41	0.13	0.50	0.09	0.41	0.13	0.44	0.09
small	true	0.35	0.05	0.48	0.03	0.35	0.05	0.53	0.04
small	false	0.45	0.05	0.64	0.03	0.45	0.05	**0.62**	0.04
large	true	0.26	0.00	0.56	0.00	0.26	0.00	0.40	0.00
large	false	0.24	0.00	0.57	0.00	0.24	0.00	0.37	0.00

Discussion

Limitations

- quality issues with generated dataset limit the performance of topic classification

Limitations

- quality issues with generated dataset limit the performance of topic classification
- pre-trained model from Hugging Face is not fully tailored to our needs

Limitations

- quality issues with generated dataset limit the performance of topic classification
- pre-trained model from Hugging Face is not fully tailored to our needs
- classes are too broad, since top-level sections are used

Key Takeaways

- a wide variety of research into this area already exists

Key Takeaways

- a wide variety of research into this area already exists
- we propose a novel architecture to develop a domain-dependent and fine-grained topic model

Key Takeaways

- a wide variety of research into this area already exists
- we propose a novel architecture to develop a domain-dependent and fine-grained topic model
- initial results justify further research and investments to improve the architecture

References I

Ajinaja, Micheal Olalekan et al. (Mar. 2023). "Semantic similarity measure for topic modeling using latent Dirichlet allocation and collapsed Gibbs sampling". In: *Iran Journal of Computer Science* 6.1, pp. 81–94. ISSN: 2520-8446. DOI: 10.1007/s42044-022-00124-7.

Alpizar-Chacon, Isaac, Max van der Hart, et al. (July 2020). "Transformation of PDF Textbooks into Intelligent Educational Resources". In: *iTextbooks 2020*. Proceedings of the Second International Workshop on Intelligent Textbooks 2020.

Alpizar-Chacon, Isaac and Sergey Sosnovsky (2019). "Expanding the Web of Knowledge: One Textbook at a Time". In: *Proceedings of the 30th ACM Conference on Hypertext and Social Media*. HT '19. Hof, Germany: Association for Computing Machinery, pp. 9–18. ISBN: 9781450368858. DOI: 10.1145/3342220.3343671.

— (2021). "Knowledge models from PDF textbooks". In: *New Review of Hypermedia and Multimedia* 27.1-2, pp. 128–176. DOI: 10.1080/13614568.2021.1889692.

— (2020). "Order out of Chaos: Construction of Knowledge Models from PDF Textbooks". In: *Proceedings of the ACM Symposium on Document Engineering 2020*. DocEng '20. Virtual Event, CA, USA: Association for Computing Machinery. ISBN: 9781450380003. DOI: 10.1145/3395027.3419585.

References II

Alpizar-Chacon, Isaac and Sergey Sosnovsky (2022). "What's in an Index: Extracting Domain-Specific Knowledge Graphs from Textbooks". In: *Proceedings of the ACM Web Conference 2022*. WWW '22. Virtual Event, Lyon, France: Association for Computing Machinery, pp. 966–976. ISBN: 9781450390965. DOI: 10.1145/3485447.3512140.

Bengio, Y., P. Simard, and P. Frasconi (1994). "Learning long-term dependencies with gradient descent is difficult". In: *IEEE Transactions on Neural Networks* 5.2, pp. 157–166. DOI: 10.1109/72.279181.

Blei, David M, Andrew Y Ng, and Michael I Jordan (2003). "Latent dirichlet allocation". In: *Journal of machine Learning research* 3.Jan, pp. 993–1022.

Devlin, Jacob et al. (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. arXiv: 1810.04805 [cs.CL].

Guerra, Julio, Sergey Sosnovsky, and Peter Brusilovsky (2013). "When One Textbook Is Not Enough: Linking Multiple Textbooks Using Probabilistic Topic Models". In: *Scaling up Learning for Sustained Impact*. Ed. by Davinia Hernández-Leo et al. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 125–138. ISBN: 978-3-642-40814-4. DOI: 10.1007/978-3-642-40814-4_11.

Hochreiter, Sepp and Jürgen Schmidhuber (Dec. 1997). "Long Short-term Memory". In: *Neural computation* 9, pp. 1735–80. DOI: 10.1162/neco.1997.9.8.1735.

References III

Mikolov, Tomas et al. (2013). *Efficient Estimation of Word Representations in Vector Space*. arXiv: 1301.3781 [cs.CL].

Pozzi, Lorenzo, Isaac Alpizar-Chacon, and Sergey Sosnovsky (July 2023). “Harnessing Textbooks for High-Quality Labeled Data: An Approach to Automatic Keyword Extraction”. In: *iTextbooks 2023*. Fifth Workshop on Intelligent Textbooks.

Spärck Jones, Karen (Jan. 1972). “A Statistical Interpretation of Term Specificity and its Application in Retrieval”. In: *Journal of Documentation* 28.1, pp. 11–21. ISSN: 0022-0418. DOI: 10.1108/eb026526.

Thaker, Khushboo Maulikmihir, Peter Brusilovsky, and Daqing He (2018). “Concept Enhanced Content Representation for Linking Educational Resources”. In: *2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*, pp. 413–420. DOI: 10.1109/WI.2018.00-59.

Table: Selected parameters for each dataset

dataset	concepts	model type	batch size	dropout rate	units
expert	true	SimpleRNN	128	0.40	125
expert	false	LSTM	128	0.40	200
small	true	SimpleRNN	32	0.90	100
small	false	SimpleRNN	64	0.90	100
large	true	SimpleRNN	64	0.90	125
large	false	SimpleRNN	64	0.90	100