A Theory of Action for Classifying and Implementing Online Text-Based Curriculum Materials Using Natural Language Processing

Aaron M. Kessler, Anindya Roy, and Daniel Seaton
kesslera@mit.edu, anindyar@mit.edu, dseaton@mit.edu
Massachusetts Institute of Technology

Abstract: The purpose of this work is to propose a theory of action for how educators can use Natural Language Processing (NLP) as a way to explore, classify, and implement online text-based curricular materials. Within the context of open and free online text-based curricular materials, this work forwards and operationalizes the theory of action for using NLP in the context of higher education physics instruction. Results demonstrate the feasibility of such a model with implications for the importance of educators being "in the loop" instead of black boxing such processes.

Significance of work

Research has repeatedly shown the impact of curricular materials on students' formal learning experiences depends on a number of factors associated with the educator (teacher, instructional designer, faculty member) and the materials themselves (Stigler & Hiebert, 2004; Forbes & Davis, 2010). The increasing number of online curricular materials, especially open and free materials, presents a unique opportunity to think about and consider how educators organize and implement such distributed curriculum resources. Building on a number of frameworks from across mathematics and science education (Cartier et al., 2013; Cohen, Raudenbush, & Ball, 2003), we present a Theory of Action (TOA) that describes a mechanism by which educators can utilize trusted resources to train natural language processing (NLP) algorithms in order to better classify other sets of online resources that can be implemented by the educator. In this manuscript we propose how previous research can guide the use of NLP and online materials to assist educators in developing learning opportunities.

A theory of action for exploring curriculum using NLP

The TOA presented in Figure 1 is predicated on a set of four assumptions. 1) The use of the term educators is intended to be very broad in the proposed TOA. Educators may be teachers, instructors, instructional designers, learning engineers, and others. 2) Educators know their context and the requirements needed of curricular materials beyond the one size fits all model of traditional curriculum. 3) Resources and processes that allow educators to create better learning opportunities can improve student learning outcomes. 4) Educators may not be able to interrogate and improve their work when the processes associated with searching and categorizing curricular materials are black boxed from them. Based on these assumptions, Figure 2 presents a visual representation of the TOA for how educators can utilize NLP tools to train a model to classify online curricular resources in order to organize and implement such resources.

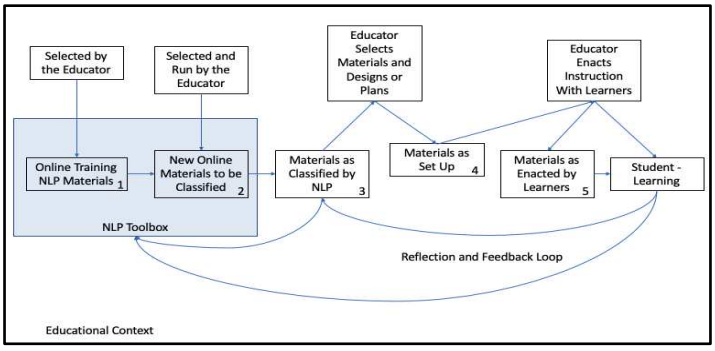

Figure 1. Theory of action for educators using NLP to classify and enact online-text based materials.

Methodological approach

We grounded this exploration of applying NLP to resource classification in the content area of Physics, as the field has a well-defined set of OER and two of the paper authors have content backgrounds in physics. We trained a topic model based on the OpenStax physics textbook "University Physics Vol. 1", to demonstrate how such a

model could be used to identify topics from a set of OER with unknown topics. Then, as an example, we applied the OpenStax topic model to classify the following OER on circular motion: 1) The Wikipedia article on circular motion, 2) A pdf document from an MIT Open Learning Library (MIT OLL) physics course (Mechanics), and 3) A transcript file of a 80-second video from the same course. We use an implementation of Latent Dirichlet Allocation (LDA) (Blei et al., 2003) based on the popular NLP package gensim (Rehurek et al., 2010).

Results

Once we trained our model with the OpenStax physics textbook, we applied the resulting 94-topic model on the 3 unseen documents specified earlier. In Figure 2 we show the results schematically: the model associates the Wikipedia article titled "Circular Motion" with 2 topics, and the MIT OLL introductory text and the subtitle file on *circular motion* with one topic. The Wikipedia article is associated primarily with 2 topics, #9 and #88, with topic-distribution probabilities 0.54 and 0.25 respectively, while the text and the subtitle file from MIT OLL shows a very strong association with the topic #9 (a probability of 0.89 and 0.90, respectively).

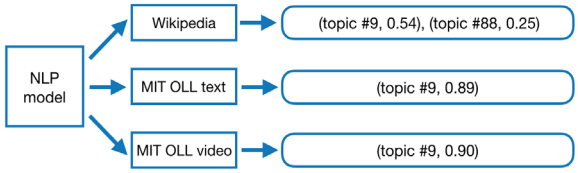

Figure 2. Application of the NLP model on 3 specific resources and the inferred topics with associated topic-distribution probability.

Conclusions and implications

By making the TOA explicit this work hopes to provide the NLP and Learning Sciences communities with a frame for thinking about the next steps in exploring, understanding, and testing how educators can and should engage with NLP as a way to find and classify online text-based curricular resources. In grounding this work in research related to curriculum material implementation, we hope to point researchers and educators to critical portions of the TOA that demand attention in support of learners and learning outcomes. Most importantly, the education community must move beyond the black box approach to curriculum in order to make explicit how we conduct the work of finding, organizing, and enacting distributed curricular materials. More research is needed in order to build effective NLP toolkits for educators and to understand how educators implement such tools.

References

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, *3*(Jan), 993-1022.

Cartier, J. L., Smith, M. S., Stein, M. K., & Ross, D. K. (2013). *5 Practices for Orchestrating Productive Science Discussions*. Reston, VA: National Council of Teachers of Mathematics and Corwin Press.

Cohen, D. K., Raudenbush, S. W., & Ball, D. L. (2003). Resources, instruction, and research. *Educational evaluation and policy analysis*, *25*(2), 119-142.

Forbes, C. T., & Davis, E. A. (2010). Curriculum design for inquiry: Preservice elementary teachers' mobilization and adaptation of science curriculum materials. *Journal of Research in Science Teaching*, *47*(7), 820-839.

Rehurek, R., & Sojka, P. (2010). Software framework for topic modelling with large corpora. In *In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*.

Stigler, J. W., & Hiebert, J. (2004). Improving mathematics teaching. *Educational leadership*, *61*(5), 12-17.

