Using Topic Modeling to Assess Middle School Science Discourse based on Next Generation Science Standards

Erin Shaw[1,a], Richard Phillips[1,b], Anne Kao[1,c], and Robert Torres[1,d]

[1]University of Southern California, Los Angeles, CA, USA

[a]erinshaw@usc.edu, [b]rhphilli@usc.edu, [c]akao@usc.edu, [d]rtorres@usc.edu

Keywords: K-12 education assessment, discourse analysis, Next Generation Science Standards, natural language processing, topic modeling.

Abstract. In this paper we apply natural language understanding and topic modeling methods to create an instructional tool for assessing science discourse in interactive multimedia presentations that are created, shared and discussed by students. The resulting assessment prototype will enable teachers to follow student discourse in real time, to understand gaps in student learning, and to improve teaching practices. Middle School Life Sciences and Physical Sciences standards from the Next Generation Science Standards (NGSS) serve as a basis for assessing domain relevance. Domain knowledge and standards are modeled, and student discussions are analyzed and graphed.

Introduction

Participatory cultures are defined as cultures with relatively low barriers to artistic expression and civic engagement, strong support for creating and sharing one's creations, and some type of informal mentorship [1]. Participation fosters new media literacies that build on traditional literacy skills taught in the classroom but focus on social skills developed through collaboration and networking, such as play, performance, simulation, appropriation, distributed cognition, and judgment. New educational practices like *Playground,* an implementation of the *Participatory Learning and You!* (*PLAY!*) framework, have the potential to produce "radical and transformational shifts" in learning [2][3]. Playground is a social, multimedia development environment that encourages users to experiment with ideas. It embraces participatory learning by implementing the four C's of participation in the learning process: creation, circulation, collaboration and connection.

For new practices to become accepted, however, they must fit into the curriculum, be aligned to state standards and have appropriate assessments [4]. The goal of our research was to enable instructors who use Playground as a teaching platform to assess student discourse in two science domains, the middle school subjects of Life Sciences and Physical Sciences, with respect to U.S. national science standards. Each domain and corresponding standard was modeled, and the models were used to analyze student discourse within Playground presentations, or *canvases*. The research was the principle component of a U.S. National Science Foundation *Research Experiences for Undergraduates* program, and built upon a pipeline for assessing the discourse of students who created Minecraft canvases [5]. The work presented here is based on authentic K-12 textbooks and core standards, as compared to Minecraft documentation and engineering practices. The approach has not been previously published.

Methodology

The preprocessing and real-time processing steps of the development pipeline are shown in Figure 1. Preprocessing steps included creating corpora for each science domain by performing natural language processing (NLP) and topic modeling on two science textbooks and Next Generation Science Standards (NGSS) descriptions [6]. Real-time processing steps included creating six Playground canvases – three on life sciences and three on physical sciences; participating in discussions about the science content; and then analyzing and graphing the results.

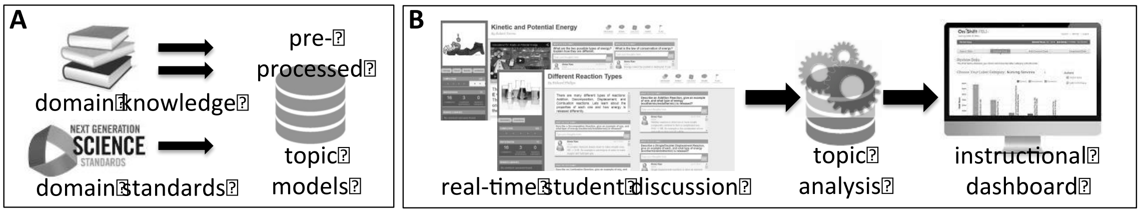


Figure 1. (A) Preprocessing steps. (B) Real-time processing steps.

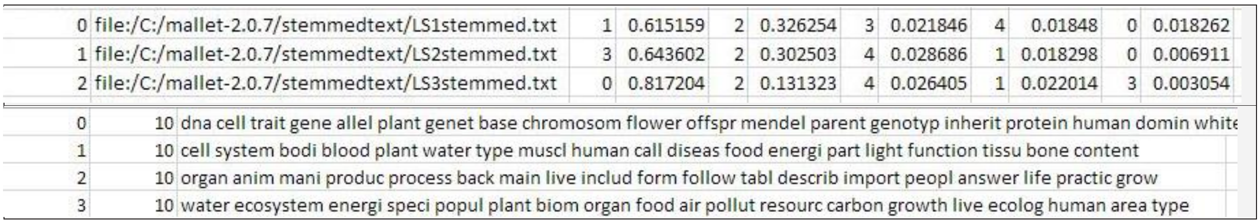
Corpora Creation: The subject domains were based on the six NGSS *core ideas*, three from the life sciences (LS1, LS3, LS4) and three from the physical sciences (PS1, PS2, PS3), shown in Table 1 (top). These, together with NGSS *scientific and engineering practices* and *crosscutting concepts*, support middle school students in developing useable knowledge across the science disciplines [6]. Two online textbooks were used to build a domain corpus for the life sciences and two were used to build a second corpus for the physical sciences. Online PDF files were converted to ASCII text files and each chapter was marked as corresponding to an NGSS core idea, or unit, sometimes provided by the publishers themselves. Text corresponding to each unit was transferred to one of four large text files for each unit. For example, LS4 contained material from five chapters of one textbook and one chapter of another. This created enough material for each standard and substandard, for example, LS1-1 and LS1-3, shown in Table 1 (bottom). The same procedure was followed to create small corpora for each of the two science domain standards based on NGSS documentation.

Table 1. Top: Core Ideas of LS and PS domains. Bottom: Sub-standards for the Life Sciences.

Core Ideas from Life Sciences (LS)	Core Ideas from Physical Sciences (PS)
LS1: Molecules to Organisms: Structures & Processes LS3: Heredity: Inheritance &Variation of Traits LS4: Biological Evolution: Unity and Diversity	PS1: Matter & its Interactions PS2: Motion & Stability: Forces & Interactions PS3: Energy

LS1-1	Conduct an investigation to provide evidence that living things are made of cells; either one cell or many different numbers and types of cells.
LS1-3	Use argument supported by evidence for how the body is a system of interacting subsystems composed of groups of cells.

Topic Modeling: The domain and standards corpora were preprocessed using NLTK [7]: Stop words, numbers and punctuation were removed, and the words were stemmed. The glossary and indexes of one of each of the textbooks were added to the domain dictionaries to support a larger science vocabulary [8]. Mallet [9], a language processing tool, was then used to identify key topics in each corpus, resulting in a list of topic words for each core idea and for each subject. For example, in Figure 2, we see that *LS1stemmed.txt*, at top, which contains the LS1 corpus, is correlated with topic 1 (62%) and topic 2 (33%), below.

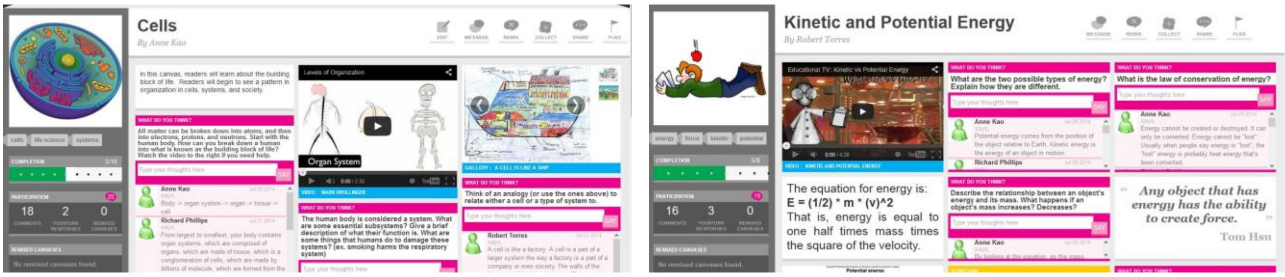


0	file:/C:/mallet-2.0.7/stemmedtext/LS1stemmed.txt	1	0.615159	2	0.326254	3	0.021846	4	0.01848	0	0.018262
1	file:/C:/mallet-2.0.7/stemmedtext/LS2stemmed.txt	3	0.643602	2	0.302503	4	0.028686	1	0.018298	0	0.006911
2	file:/C:/mallet-2.0.7/stemmedtext/LS3stemmed.txt	0	0.817204	2	0.131323	4	0.026405	1	0.022014	3	0.003054

0	10	dna cell trait gene allel plant genet base chromosom flower offspr mendel parent genotyp inherit protein human domin white
1	10	cell system bodi blood plant water type muscl human call diseas food energi part light function tissu bone content
2	10	organ anim mani produc process back main live includ form follow tabl describ import peopl answer life practic grow
3	10	water ecosystem energi speci popul plant biom organ food air pollut resourc carbon growth live ecolog human area type

Figure 2. Top: Topics and percentages for LS standards. Bottom: Words corresponding to topics.

Playground Canvases: Students created six canvases on *Cells*, *Genetic Traits*, *Mechanisms of Evolution*, *Different Reaction Types*, *Gravity and Electromagnetism*, and *Kinetic and Potential Energy*. Two examples are shown in Figure 3. Students used prompts like "*What do you think?*", which were integrated into Playground, to start discussions in which each of the other students then participated, resulting in multiple discussion threads per canvas. The project leader, in the role of the teacher, also participated in the canvas discussions. The final canvases included photos, videos, and links to websites. The threads were extracted from the Playground database, cleaned with NLTK and processed with MALLET's *train-topics* and *infer-topics* commands to identify the relevant topics and enable us to analyze how frequently students spoke of domain concepts. The results were pushed to a web-accessible database with information about the author and canvas. The analysis for the prototype was performed offline due to access restrictions on the database, but with the preprocessed topics, the analysis and dashboard display can be easily performed in real time.



Cells
By Anne Kao

In this canvas, readers will learn about the building block of life. Readers will begin to see a pattern in organization in cells, systems, and society.

WHAT DO YOU THINK?
All matter can be broken down into atoms, and then into electrons, protons, and neutrons. Start with the human body. How can you break down a human into what is known as the building block of life? Watch the video to the right if you need help.

Anne Kao says: Body -> organ system -> organ -> tissue -> cell

Richard Phillips says: From largest to smallest, your body contains organ systems, which are comprised of organs, which are made of tissue, which is a conglomeration of cells, which are made by billions of molecule, which are formed from the

VIDEO : MARK DROLLINGER

WHAT DO YOU THINK?
The human body is considered a system. What are some essential subsystems? Give a brief description of what their function is. What are some things that humans do to damage these systems? (ex. smoking harms the respiratory system)

GALLERY : A CELL IS LIKE A SHIP

WHAT DO YOU THINK?
Think of an analogy (or use the ones above) to relate either a cell or a type of system to.

Robert Torres says: A cell is like a factory. A cell is a part of a larger system the way a factory is a part of a company or even society. The walls of the

COMPLETION 5/10
PARTICIPATION 20
18 COMMENTS, 2 YOURTURN RESPONSES, 0 REMIXED CANVASES
REMIXED CANVASES
No remixed canvases found.

Kinetic and Potential Energy
By Robert Torres

VIDEO : KINETIC AND POTENTIAL ENERGY

The equation for energy is:
E = (1/2) * m * (v)^2
That is, energy is equal to one half times mass times the square of the velocity.

WHAT DO YOU THINK?
What are the two possible types of energy? Explain how they are different.

Anne Kao says: Potential energy comes from the position of the object relative to Earth. Kinetic energy is the energy of an object in motion.

Richard Phillips

WHAT DO YOU THINK?
What is the law of conservation of energy?

Anne Kao says: Energy cannot be created or destroyed. It can only be converted. Energy cannot be "lost". Usually when people say energy is "lost", the "lost" energy is probably heat energy that's been converted.

WHAT DO YOU THINK?
Describe the relationship between an object's energy and its mass. What happens if an object's mass increases? Decreases?

Anne Kao says: By looking at the equation, as the mass

"Any object that has energy has the ability to create force." Tom Hsu

COMPLETION 5/8
PARTICIPATION 19
16 COMMENTS, 3 YOURTURN RESPONSES, 0 REMIXED CANVASES
REMIXED CANVASES
No remixed canvases found.

Figure 3. Student canvases on Cells and Energy, respectively, with corresponding discussions.

Results

Two charts were created to visualize the results. The bar chart shown in Figure 4, which was created with charts.js, shows the science standards that most closely align (topically) with the discussions. Teachers can select individual users or all users (shown). Instructors might use this chart to gauge the overall level of participation and infer student interest, understanding and/or teaching gaps. For example, the topic Matter may have been discussed less because relatively less content was taught or, perhaps, because the topic was less interesting, or understood, than others.

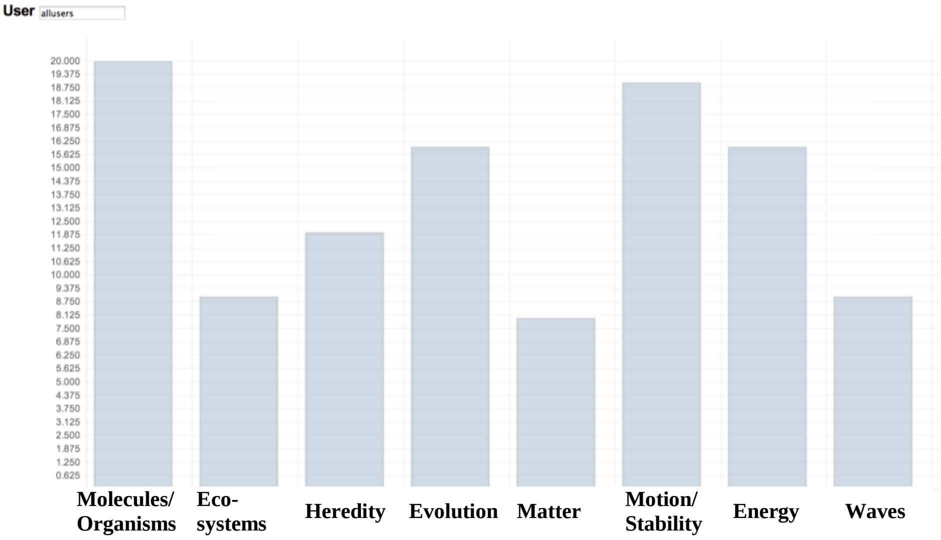


Figure 4. A topic-model based analysis of student discourse, per user or for all users (shown).

In Figure 5, which was created with Google Charts, the grouped bar chart displays the percentage of student discourse relevant to each topic. Instructors might use this chart to compare student activity within and between subjects, to see who is participating and in what they are participating. For example, the results for Students 1, 2 and 3, who created canvases on Evolution and Motion (blue), Heredity (red), and Molecules and Matter (yellow), respectively, show a spike for their own topics, which was expected. Analogously, the results for Student 4 (green), who participated in each canvas but did not create her own, do not include any spikes.

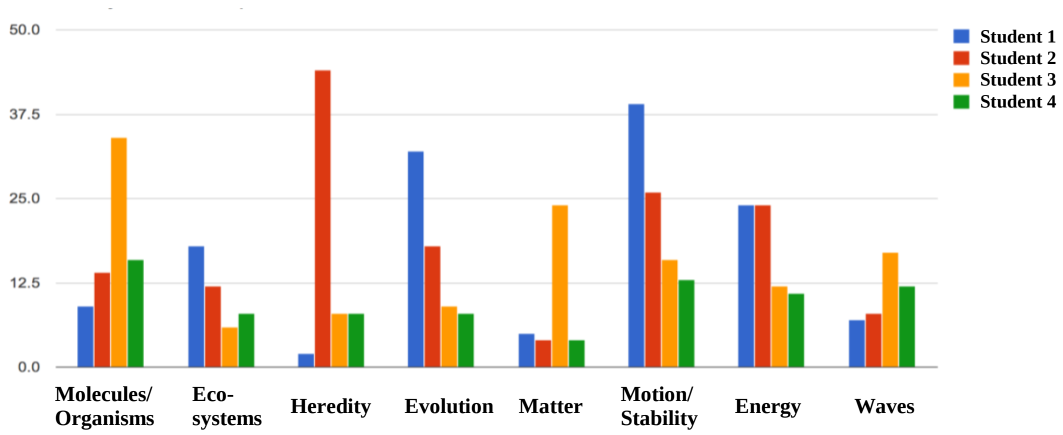


Figure 5. A topic-model based analysis of student discourse for each subject, for each student.

Summary

In this paper, we have presented a new model of student discourse-based assessment. Machine learning and topic modeling were used to build a preprocessed model of two science domains and their standards, and a real-time model of student discourse based on these standards. The observed and expected results of the analysis were in correspondence, and teachers may use the resulting graphs to assess student discussion within the selected domains. While any textbook may be used to build a corpus, differing school curricula may emphasize different standards and make it difficult to adapt the program nationally. Future work includes expanding MALLET's keyword results to encompass a greater range of the science domain and finding the ideal level for topic modeling. Aligning standards with reading levels was also a challenge and was addressed in a related paper.

Acknowledgements

The authors thank Creative Director Erin Reilly and Software Developer Aninoy Mahapatra of USC's Game Innovation Lab for their support and assistance. This work was supported by a National Science Foundation Research Experiences for Undergraduates supplement (#1008747).

References

[1] Jenkins, H., Clinton K., Purushotma, R., Robinson, A. J. & Weigel, M. (2006). Confronting the challenges of participatory culture: Media education for the 21st century. Chicago, IL. Online at http://www.macfound.org/media/article_pdfs/JENKINS_WHITE_PAPER.PDF.

[2] Reilly, E., Jenkins, H., Felt, L.J. & Vartabedian, V. (2012) *Shall we PLAY?* Online at http://www.annenberglab.com/sites/default/files/uploads/Shall_We_PLAY_final_small.pdf.

[3] McLoughlin, C. & Lee, M. J. (2007). Social software and participatory learning: Pedagogical choices with technology affordances in the Web 2.0 era. In *ICT: Providing choices for learners and learning.* In *Proceedings of ASCILITE*, Singapore, 2007, pp. 664-675.

[4] Nobori, M. (2013). A Step-by-Step Guide to Best Projects: Discover a project-based learning model that motivates students to pursue knowledge and drives academic achievement, Edutopia. Online at http://www.edutopia.org/stw-project-based-learning-best-practices-guide.

[5] Shaw, E., La, M. Phillips, R. & Reilly, E. (2014). PLAY Minecraft! Assessing secondary engineering education using game challenges within a participatory learning environment. In *Proceedings, American Society for Engineering Education (ASEE) 2014.*

[6] Next Generation Science Standards (2013). Online at http://www.nextgenscience.org.

[7] Loper, E. & Bird, S. (2002). NLTK: Natural Language Toolkit. In *Proc., ACL-02 Workshop on Effective tools and methodologies for teaching natural language processing and computational linguistics*. Vol. 1, Association for Computational Linguistics, Stroudsburg, PA, 2002, pp. 63-70.

[8] Bird, S., Loper, E. & Klein, E. (2009). Natural Language Processing with Python. O'Reilly Media Inc.

[9] McCallum, A.K. (2002). MALLET: A Machine Learning for Language Toolkit (mallet.cs.umass.edu)