Deconstruct and Reconstruct: Using Topic Modeling on an Analytics Corpus

Mike Sharkey
Blue Canary
145 S. 79th St., ste. 59
Chandler, AZ 85226
602-617-4174
mike@bluecanarydata.com

Mohammed Ansari
Blue Canary
145 S. 79th St., ste. 59
Chandler, AZ 85226
480-262-3438
mohammed@bluecanarydata.com

ABSTRACT
The question posed by the 2014 LAK Data Challenge is "What do analytics on learning analytics tell us?" The authors looked to take a two-pronged approach to this challenge. First, the authors wanted to use advanced analytical techniques on the corpus to make the "eat your own dog food" point. Since many of the EDM/LAK submissions explain advanced statistical or semantic analytic approaches, we wanted to utilize those same methods for our analysis. To that end, we used two natural language processing (NLP) tools to analyze the corpus of papers. First we used Latent Dirichlet allocation (LDA) to extract clusters of terms from the content. Second, we used Turbo Topics to convert the LDA output into phrases (bi-grams and tri-grams).

The use of these NLP tools allowed us to execute the second part of our approach to the challenge. Once the corpus was aggregated as topics, we used Tableau to visually inspect the corpus for trends. In addition to standard descriptive visualizations, we were able to identify trends in the corpus topics from 2008 through 2013. Most interesting is that with both EDM and LAK, we noticed a trend of topic convergence after three years. Also, we were able to easily discern topic trends such as the increased popularity of "social" and "network," over the last three years, and the consistent appearance of 'Cognitive Tutor' related topics (e.g. intelligent tutoring, concept map). While these findings may not be unexpected, we believe that the ability to extract and visualize these outcomes is unique.

Categories and Subject Descriptors

• **Human-centered computing~Information visualization**

• *Computing methodologies~Natural language processing*

• *Computing methodologies~Topic modeling*

• *Computing methodologies~Latent Dirichlet allocation*

The ACM Computing Classification Scheme: http://dl.acm.org/ccs_flat.cfm

General Terms
Algorithms, Measurement, Design, Languages, Theory.

Keywords
Natural language, LDA, Turbo Topics, EDM, LAK, corpus, Tableau, visualization, analytics, social networks, cognitive tutor, IRT, assessment.

1. INTRODUCTION
The 2014 LAK Data Challenge is a classic meta problem. The challenge poses the question "What do analytics on learning analytics tell us?" Blue Canary chose to enter the challenge in order to contribute the analytical body of work that the EDM and LAK community members have been doing for years. As we thought about how the meta-analysis would work, we tried to use key tenets of good data analysis. One such tenet can be summarized as "automation, but with the human touch." By this we mean that we use our engineering skills to automate as many parts of the analytical process as possible, but we still rely on human intervention when required/appropriate.

The corpus of papers was comprised of papers submitted to the Educational Data Mining (EDM) conferences from 2008 to 2013 and papers submitted to the Learning Analytics and Knowledge (LAK) conferences from 2011 to 2013. Our approach to analyzing the corpus was twofold – extract topics from the corpus and then use visualizations to surface findings.

1.1 NLP and Topic Modeling
First, we used natural language processing (NLP) tools to model topics from the corpus. Precious metals processing is an appropriate analogy in this case. A gold mining operation excavates large rocks, breaks them down into ore, refines the pure gold, and then sells the gold so that designers can create jewelry. For the end user, it is the gold jewelry that is of value. Similarly, we looked at a large corpus of papers, broke it down into word vectors, aggregated those vectors into topics and aggregated again into concepts.

1.2 Visualization
Continuing to use the gold analogy, most buyers don't judge the value of a piece of jewelry by examining the quality of the gold. The value is assessed by looking at the overall presentation of the piece. For our analysis, we wanted to present data visualizations that would surface the findings and information that peers would find interesting. Additionally, though, we also wanted to allow users to 'inspect the gold' if desired. We used Tableau to create and deliver the visualizations, and we used a topic browser to let users browse topics in the context of their original papers.

2. TOPIC MODELING METHODOLOGY
The bulk of the analysis we performed was guided by the Topic Modeling work driven by David M. Blei at Princeton.[1] Specifically, we felt that using Blei's work on Turbo Topics would be the best approach to the EDM/LAK corpus. Turbo Topics builds off of single term topics and aggregates the findings into multiple term n-grams that give the user more context [1]. An n-gram (e.g. bi-grams are two-word phrases) such as "cognitive tutor" has much more meaning in this space than the terms "cognitive" and "tutor" independently.

[1] http://www.cs.princeton.edu/~blei/topicmodeling.html

2.1 Analysis Process

We followed a specific process in order to deconstruct the corpus down to topics and then aggregate the topics back up to meaningful n-grams. Figure 1 below shows the steps involved:

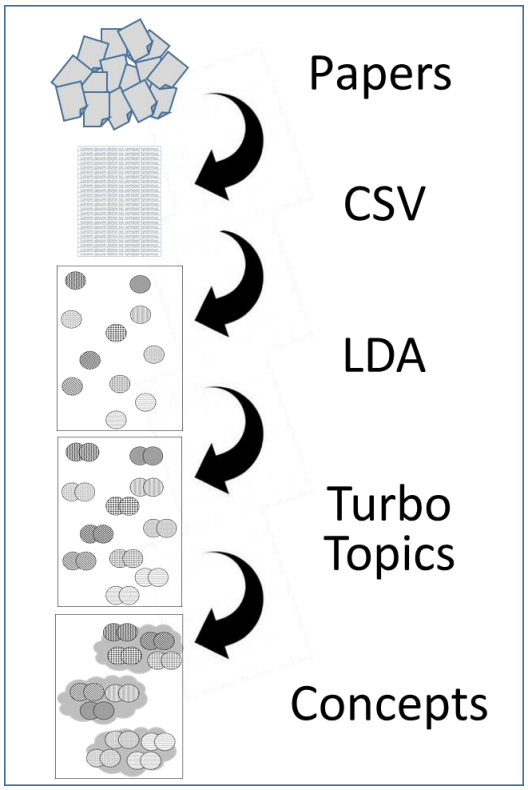

Figure 1. Blue Canary topic modeling process

The papers started in XML format[2] thanks to the work done by Taibi & Deitze [2]. We had to convert that into a format that was more suitable for our NLP pipeline. The papers were converted into a CSV file where each paper was a line in the CSV. Once the corpus was more readily machine readable, we used LDA to process the first round of topic aggregation. LDA assumes that there are latent underlying topics in a corpus and that the topic has a number of correlated words [3].

The output of LDA was a series of vectors, with each vector corresponding to an assumed underlying topic. These vectors then became the input for Turbo Topics – a process that would ingest the LDA results and create a series of n-grams that are relevant to the topics in the corpus. Examples of such n-grams from this corpus included "classification algorithms", "intelligent tutoring", and "decision tree".

2.2 Human Intervention

While Blue Canary attempted to systematize as much of the analysis as possible, we realize that there is still a need for human intelligence to guide the topic modeling.

The first step of human intervention was in selecting stop words. These are words that should be excluded from the analysis because their frequency doesn't add value to the observations. In the CSV step, we experimented with different upper and lower bound settings for stop word limitations. We settled on only including words that appeared more than 50 times but less than 200 times – running LDA with these limits created a set of topic vectors that we observed to be optimal.

A second example of human intervention was in the final step of grouping topics into concepts. This was purely a manual process that involved browsing the approximately 80 n-grams, looking for analytic themes, and grouping them accordingly. For example, topics such as "error rate", "feature selection", and "activity sequences" were grouped as Machine Learning while "discussion forums", "natural language", and "topic words" were grouped as Semantic/Text. The entire list of topics and concepts can be seen in the "Concepts and Terms" tab of the accompanying Blue Canary LAK site[3].

3. VISUALIZING THE RESULTS

In the analytics space, visualizing ones data is an effective way of divining trends and patterns in the underlying set. For the LAK challenge, we had the topics and concepts as the core results from our analysis. We created two metrics that we could use to frame the observation of these results.

The first was frequency – how many times does the topic appear in the corpus. Since both the number of papers per year and the number of words per paper varied from 2008 to 2013, we had to normalize this metric. We chose to index the frequency on the most frequent term. To illustrate, we look at the corpus in 2011 where the most frequent topic was 'social network' with 112 appearances. The next most frequent topic was 'item difficulty' with 90 appearances. In our analysis, we gave 'social network' a frequency of 1.0 and 'item difficulty' a score of 0.8.

The second metric was breadth – the number of different papers containing a given topic. Again, we had to normalize since the number of papers increased from 27 to 144 over the 6-year timeframe. We normalized breadth by using percent of documents in which the topic appeared. To remove outliers, we set a rule that a topic must appear in at least 5% of the papers to be included in our analysis.

3.1 Tableau

The first tool used for visualization was Tableau. Tableau's advanced visual features makes it an ideal tool for exploring our results.

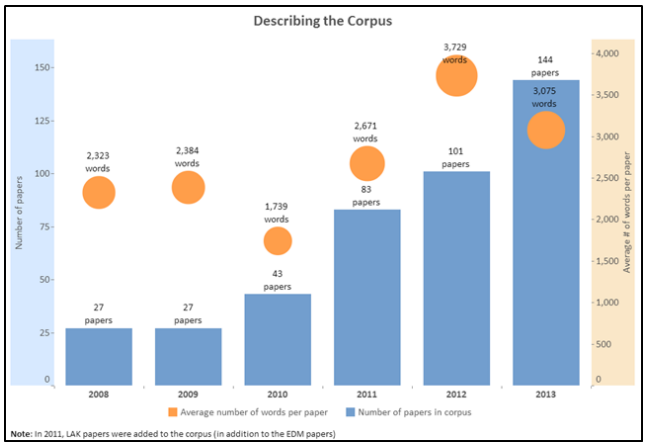

Figure 2. Describing the Corpus

[2] http://meco.l3s.uni-hannover.de:9080/wp2/?page_id=16

[3] http://lak14.bluecanarydata.com

All Tableau visualizations can be found at the Blue Canary LAK site (http://lak14.bluecanarydata.com). The visualizations can range from simple descriptive charts (like Figure 2 showing the size and breadth of papers in the corpus) to interactive trend charts (like Figure 3 where the user can find the top N topics from any of the six years of the corpus).

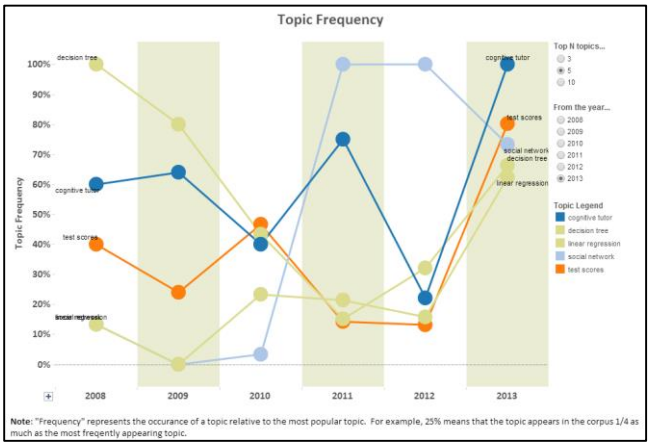

Figure 3. Top Topics per Year

In the context of the LAK Data Challenge, Tableau was the most useful tool utilized by the Blue Canary team. We did not approach the task with a specific hypothesis to be proved or disproved. Rather, we took the challenge more literally and asked 'What do the data have to say?' Tableau helped us find answers to that question.

3.2 Topic Browser

A second, more detailed way to view the output is using a web-based topic browser. Researchers have developed different tools to accomplish this task, including tools such as TopicExplorer [4] and Topic Model Visualization Engine [5]. Blue Canary chose to use Topic Model Visualization Engine (as shown in Figure 4). The topic browser can also be found on the Blue Canary LAK site (http://lak14.bluecanarydata.com).

Figure 4. Topic Model Visualization Engine

This topic browser allows the user to explore the occurrences of any topic in the context of its native paper. This explorer is useful when trying to decipher irregularities of the topic modeling output. For example, the topic "free fall" showed significant presence in the 2013 papers. It turns out that two different papers used a Physical Sciences class as the backdrop for their analysis and "free fall" was one of the course concepts.

4. FINDINGS

After processing the data and looking for trends, the Blue Canary team found two things that we believe are of interest to the analytics community.

4.1 Topic Convergence

We created a scatter plot of all topics year by year. We excluded topics that appeared in less than 5% of the papers for the year and we plotted against our two main metrics (topic frequency and breadth).

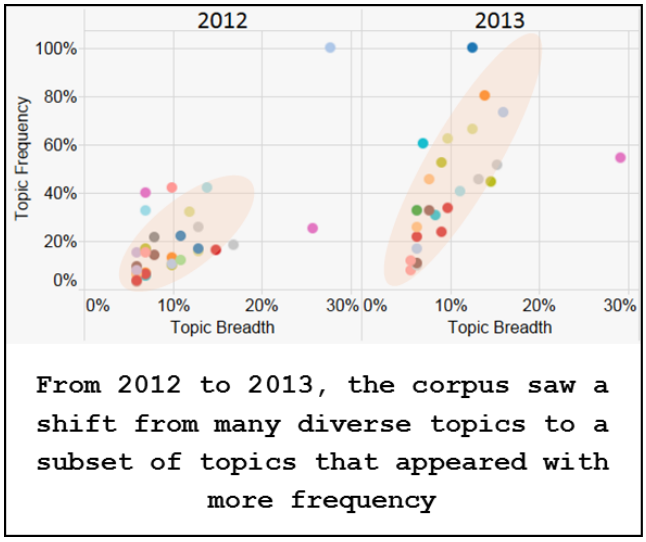

Figure 5. Scatter Plots

The resulting scatter plots (as shown in Figure 5. and on the 'Convergence' report on the blue Canary LAK site) show an interesting trend. The slope of the line reflects the homo/heterogeneity of the paper topics. A shallow slope indicates that few topics dominate the overall corpus conversation. A steeper slope indicates that there are more topics that are frequently mentioned across more papers. The purpose of the regression is to help highlight the trend. It is not meant to comment on the strength of the fit.

Looking at the patterns in the scatter plot, we see that in the first three years of the LAK papers, the topics trended towards a more concentrated set. The implication here is that at the start of the conferences, papers tended to be more diverse. However, after a few years, submissions started to address a similar set of topics. One possible explanation of this is that after two years, the topics become more accepted in the space and therefore get adopted/used more frequently.

One caveat is that this trend might be specific to the conference (EDM vs. LAK). Additional work to split the data by conference might shed more light on this trend.

4.2 Concept Trends

The most obvious output of our LDA to Turbo Topics to Concept process is to look at the popularity of the overall concepts over time. The 'Concept Frequency' report shows that starting in 2011 (when the LAK papers were introduced to the corpus), the topics

associated with 'Social and Networks' tended to dominate in popularity. This is not surprising as this concept includes topics such as 'interaction network', 'online communities', and 'network structure'.

A second observation about concept frequency is the consistent presence of topics associated with 'Cognitive Tutors". There is a close link between work done by researchers in this area and both the EDM and LAK communities so it's not too surprising to see this outcome. What makes the presence of this concept more striking is that there are two other concept groups about related fields ("Item Response Theory" and "Assessment Topics"). Even with these topics being spread across three different concept groups, their high frequency still shows up in the reports.

4.3 Top Topics

For reference purposes, the following tables list the top 3 topics that appeared in the corpus from 2008 to 2013. This information is also available at the Blue Canary LAK site (http://lak14.bluecanarydata.com) under the 'Topic Frequency' report tab:

Table 1. Top 3 Topics by Year

Year	Top Topics	Relative %
2008	Decision tree	100%
	Feature selection	73%
	Logistic regression	60%
2009	Correct answer	100%
	Decision tree	80%
	Statistically significant	80%
2010	Final exam	100%
	Predictive accuracy	90%
	Standard deviation	77%
2011	Social network	100%
	Item difficulty	80%
	Cognitive tutor	75%
2012	Social network	100%
	Logistic regression	42%
	Final grade	42%
2013	Cognitive tutor	100%
	Test scores	80%
	Social network	73%

'Relative %' in Table 1 refers to the frequency with which a topic appears in the corpus relative to the most frequently appearing topic.

5. COMPARISON TO OTHER ANALYSES

Blue Canary is by no means the first to apply NLP techniques to the corpus of analytic papers in an attempt to extract meaning. Prior LAK Data Challenge entrants have taken similar approaches.

5.1 LAK13 Ontology Learning

In the 2013 LAK Data Challenge, Zouaq et. al. [6] used an ontology learning tool to extract concepts and concept maps from the corpus. The researchers presented the top ranked concepts from both the EDM and LAK papers, and from the EDM and LAK abstracts. The resulting table showed that most of the top concepts were unigrams such as "student", "datum", "model", "learner", and "result".

Contrasting this to Blue Canary's Topic Modeling approach, we see a natural progression from these unigrams to bigrams (as can be seen in Table 1). This progression is a good example of how one body of research can build upon previous works in order to add more clarity for the audience.

5.2 LAK13 Dynamic Topic Modeling

Another 2013 LAK Data Challenge entrant, Derntl et. al. [6], used an approach that was more similar to what Blue Canary did with Turbo Topics. The researchers used Dynamic Topic Modeling, a precursor to the Turbo Topics technique developed by Blei. One key difference was that the Blue Canary work tried to make the topics more understandable. That is, a grouping of keywords forms a topic, but that topic needs to be something palatable to the reader. Derntl et. al. labelled their topics as an amalgam of the keywords (e.g. students – data – courses – system). While this is descriptive of the content, it is less relatable in context. Blue Canary's use of bi-grams and concept labelling helps to better bridge the context gap.

5.3 Google Trends

As a litmus test for the topic trends, Blue Canary also looked at a Google Trends chart of the popularity of some of the LAK/EDM topics (http://bit.ly/P23CRn). While interesting to look at (Figure 6.), this avenue doesn't provide much insight into the trajectory of the LAK/EDM topics. Google Trends takes its popularity metrics from a wider swath of sources so the ratings shouldn't be expected to be correlated with the work from the corpus. As an example, the topic "social network" was left off of the Google Trends search. The popularity of the 2010 movie made the scale of that term dwarf all others.

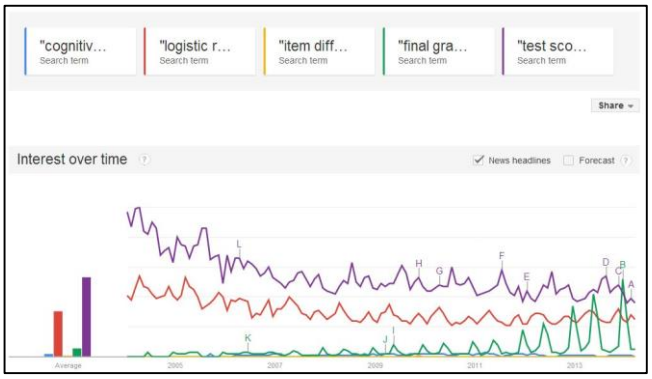

Figure 6. Google Trends topic chart

6. CONCLUSION

Blue Canary went in to the LAK Data Challenge assuming that we could accomplish two goals. First, that we could use our engineering expertise and analytical knowledge to efficiently process the corpus. Second, that we could visualize the processed results and uncover findings that would be of interest to the EDM and LAK communities. We believe that the steps outlined in this paper combined with the visualizations created with the output (http://lak14.bluecanarydata.com) prove that we have successfully accomplished our goals.

Perhaps the most salient takeaway is what's referenced in the title of this paper. The process that Blue Canary used to analyze topics was to deconstruct the corpus to a more atomic level and then to reconstruct the findings into contextual parts. It is this reconstruction that we believe has the most value. This paper built off of previous researchers who did a similar job of deconstructing the papers. What makes this paper different, though, is that Blue Canary reconstructed the findings to a more coarse level that allows others to better understand the topics discussed in the corpus

Blue Canary took this approach as a way to stress the fact that the analytics must be usable by others in order for the work to have some tangible value beyond pure research. The research furthers the state of the art, and then the application of the research is what's used by institutions and businesses to help students and customers. We reconstructed keywords into topics and concepts, and we also created a companion web application (http://lak14.bluecanarydata.com) that allows users to browse and drill into the findings. This is a good example of how the research can be extended to an applied solution that can derive value from analytics.

7. ACKNOWLEDGEMENTS

Deepest thanks to Andy Allen and the rest of the Blue Canary team members who contributed directly and indirectly. Papers like this are a great example of what smart people can do when they work collaboratively.

Additional thanks to the EDM and LAK communities for continually fostering a culture of innovation around data and analytics in higher education.

8. REFERENCES

[1] Blei, D. M., & Lafferty, J. D. (2009). Visualizing topics with multi-word expressions. *arXiv preprint arXiv:0907.1013*.

[2] Taibi, D., & Dietze, S. (2013). Fostering analytics on learning analytics research: the LAK dataset.

[3] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2001). Latent dirichlet allocation. In Advances in neural information processing systems (pp. 601-608).

[4] Hinneburg, A., Preiss, R., & Schröder, R. (2012). TopicExplorer: Exploring document collections with topic models. In Machine Learning and Knowledge Discovery in Databases (pp. 838-841). Springer Berlin Heidelberg.

[5] Chaney, A. J. B., & Blei, D. M. (2012, March). Visualizing Topic Models. In ICWSM.

[6] Zouaq, A., Joksimović, S., & Gašević, D. Ontology Learning to Analyze Research Trends in Learning Analytics Publications.

[7] Derntl, M., Günnemann, N., & Klamma, R. (2013). A Dynamic Topic Model of Learning Analytics Research.