

Content analytics for curriculum review: A learning analytics use case for exploration of learner context

Leah P. Macfadyen

The University of British Columbia

Curriculum analysis is a core component of curriculum renewal. Traditional approaches to curriculum analysis are manual, slow and subjective, but some studies have suggested that text analysis might usefully be employed for exploration of curriculum. This concise paper outlines a pilot use case of content analytics to support curriculum review and analysis. I have co-opted Quantext – a relatively user-friendly text analysis tool designed to help educators explore student writing – for analysis of the text content of the 17 courses in our online master's program. Quantext computed descriptive metrics and readability indices for each course and identified top keywords and ngrams per course. Compilation and comparison of these revealed frequent curricular topics and networks of thematic relationships between courses, in ways that both individual educators and curriculum committees can interpret and use for decision-making. Future Quantext features will allow even more sophisticated identification of curricular gaps and redundancies.

Keywords: content analytics, text analysis, curriculum

Introduction and context

Analysis and mapping of curriculum have become routine in curriculum renewal processes in higher education, and are usually undertaken to allow evaluation, analysis and/or improvement of the scope or quality of a unit's educational offerings (Richards & Ashbourne, 2017). Curriculum analysis calls for examination of a program as a whole, and examination of individual courses to understand how/if each contributes to student learning outcomes. Does the curriculum continue to effectively represent the current state of the field? Do curricular gaps exist? Are there redundancies, with the same outcomes being taught in multiple courses? What is out of date, what is missing? Where might innovation be needed?

I am undertaking just such an analysis of the curriculum of our fully online master's degree program. In 2015, a rigorous external review concluded that our program had "high enrolment, satisfactory completion rates, and overall good course evaluations," but recommended a review and updating of curriculum, including mapping "current course offerings to program goals and objectives... eliminating courses or content...that overlap." In response, we embarked on an ongoing project of program review and renewal.

Content analytics and curriculum analysis

A variety of frameworks and guidelines for conducting curriculum analysis or curriculum mapping can now be found on university learning centre websites and in the educational literature (see, for example, Dyjur & Kenny, 2015; Richards & Ashbourne, 2017; University of Saskatchewan; Wolf, 2007). As Gottipatti & Shankararaman noted in 2014, however, "curriculum analysis has been mostly a manual process", which they describe as "tedious and painstaking work". Perhaps more problematically, most established curriculum mapping and review processes rely solely on course *syllabi* for documentation of course content and learning outcomes, even though syllabus quality, accuracy and currency may be variable or limited.

In their 2017 chapter in the *Handbook of Learning Analytics*, Kovanović, Joksimović, Gašević, Hatala, and Siemens (2017) reviewed a decade of work that they classified as 'content analytics':

Automated methods for examining, evaluating, indexing, filtering, recommending, and visualizing different forms of digital learning content...with the goal of understanding learning activities and improving educational practice and research. (p. 78)

While most learning analytics research and implementation studies to date have focused on ‘data about learners’, Kovanović et al. (2017) describe content analytics of learning resources as an area of learning analytics work that has sought to investigate ‘data about the learning context’. These authors point to Moore’s insight (1989) that interaction between learners and learning content is central to any educational endeavour, and note that educational content is most commonly ‘written material’, even though the landscape of educational content has expanded and diversified in the digital era. Indeed, our fully online web-based courses make heavy use of text to introduce concepts and structure learning activities. I therefore wondered whether content analytics might usefully support curriculum analysis for our purposes. Could analysis of the text of actual course *content* (rather than potentially limited syllabi) facilitate identification of dominant themes in each course, and features of course design or structure that may have educational implications?

Kovanović et al. (2017) describe that so far, most content analytics of learning materials has been focussed on developing systems to recommend relevant learning-related content to learners, to organise or classify instructional materials, or to assess the quality of available instructional materials and how they impact learning outcomes. A small collection of studies have proposed that text analysis methods might be used to explore or develop curriculum and have demonstrated application of methods such as topic modelling and cluster analysis to course materials (Fiallos & Ochoa, 2019; Gottipatti & Shankararaman, 2014; Kawintiranon, Vateekul, Suchato, & Punyabukkana, 2016; West, 2017; Wu, Yu, & Wang, 2018). It is notable, however, that these studies have been undertaken by scholars in computer science and data science programs – individuals whose area of expertise means they are confident in designing, implementing and interpreting output from highly technical analytic methods.

Across the wider educational landscape, however, technical and data literacy skills among educators and educational leaders are much more variable, and this lack of capacity is a persistent barrier to greater integration of learning analytics methods. When educators and educational leaders lack foundational technical and data literacy skills that would allow them to either carry out analyses or interpret the results, the goal of developing and share actionable data are difficult to achieve (Wolff, Moore, Zdrahal, Hlosta, & Kuzilek, 2016). A number of studies confirm this challenge (see, for example, Herodotou, Rienties, Verdin, & Boroowa, 2019; Tsai et al., 2018) and have reported that a lack of analytic and technical literacy among academic staff and leadership, together with concerns about workload and resistance to change, remain key barriers to the uptake of analytic approaches that might offer greater insight. Unless analytic methods used are easy to apply, and the results generated easy to interpret, they will not be embraced (Wei, Cutler, Macfadyen, & Shirazi, 2019).

User-friendly analysis of course content with *Quantext*?

The question guiding this pilot study, then, was whether an intuitive, ‘educator-friendly’ text analysis tool may offer sufficient insight to assist with curriculum analysis. A number of existing text analysis applications (each with different attributes) potentially meet these criteria: see, for example, AntConc¹, Voyant Tools², NetLytic³ or InfraNodus⁴, but we are unaware of any efforts to apply their analytic power to curricula. (Elsewhere, we (Rashtian, Hashemi, & Macfadyen, 2020) have recently reported early work developing a user-friendly curriculum review tool that harnesses the natural language processing power of IBM Watson (IBM, 2011); this remains a work in progress). In this manuscript, I report a simple test of the online *Quantext*⁵ platform. McDonald, Moskal, Elgort, and Gunn (2018) originally designed *Quantext* to help educators extract insights from student writing. Their explicit goal was a user-friendly and accessible analytic tool that could be adopted and used relatively easily by educators, even if they lack well-developed technical or analytic skills or experience. *Quantext* extracts text features from a text dataset, and presents findings for further exploration: readability indices, most frequent words, bigrams and trigrams, keyword-in-context displays and associated wordtree visualisations. Jargon is avoided.

This concise paper outlines a small content analytics use case using *Quantext* for analysis of curriculum. It directly addresses practitioners (educators and educational leaders), and seeks to contribute to the small but growing literature that bridges the gap between learning analytics research and development, and meaningful

¹ AntConc: <https://www.laurenceanthony.net/software/antconc/>

² Voyant Tools: <https://voyant-tools.org/>

³ NetLytic: <https://netlytic.org/home/>

⁴ InfraNodus: <https://noduslabs.com/infranodus/>

⁵ Quantext: <https://quantext.org/>

application of learning analytics tools and methods in real learning contexts.

Method

Data extraction and cleaning

I extracted a complete corpus of course content (text) from each online course in our master's degree program. For courses fully or partly hosted in the Canvas LMS, I generated a zipped 'course content export package' from a current copy of each course, and then employed a custom Python script to extract the text content from each course content package, which was saved to a csv. Similarly, for courses hosted partly or entirely in WordPress, I made use of a second custom Python script to scrape text from each WordPress course site, remove html code, and save to csv. For courses making use of both platforms, I manually combined content from each platform. This generated a final corpus of 17 course-content text documents, written in Canadian English, with an average length of 28,510 words. Each document was cleaned by spell-checking and standardisation of spelling (using MS Office built-in spell-checking features), correction of any incorrectly transcribed international characters (e.g., é, ù, Ø) and removal of any stray HTML code.

My analytic process was guided by the *Quantext User Guide*⁶. I compiled descriptive metrics for each course (number of pages, average words per page, and average number of sentences per page) and the various readability indices computed by *Quantext* (lexical density, LD, TTR, SMOG, Gunning Fog, Flesch reading ease, Flesh-Kincaid, and MTL⁷). I then selected a standardised set of analytic parameters for analysis of the text from each course: I elected to show 20 keywords and ngrams for each, to use the Students t-distribution (STUDT) statistical measure for calculation of the most common ngrams, to use a window of 4 words to identify bigrams and trigrams, and to exclude the standard blacklist of common words, filter punctuation, normalise contractions, filter numbers and lemmatise words. I developed a custom blacklist to remove words or ngrams that arose frequently but offered no insights into course themes (for example, textbook publisher names and locations, and generic course-related words such as 'course', 'module', 'page', 'quiz', etc.). For each course I identified the top 20 set of course-specific keywords, and then drilled down to discover the top 20 bigrams and trigrams containing the most common keyword. I used Tableau desktop to analyse, visualise and compare descriptive metrics and readability indices.

Results and discussion

Course-specific analysis (selected example)

Bosnic, Verbert, and Duval (2010) have proposed that automated keyword extraction from course content offers a starting point for useful content analytics. I therefore focussed on content of each of our courses in turn, and used *Quantext* to identify top keywords and ngrams (I found bigrams most consistently useful). For example, the course *ETEC 500: Research methodology in education* was redesigned in 2018 with the goal of increasing the focus on critical thinking about educational research. Keyword analysis of ETEC 500 content offered rich insights into dominant themes in this revised course. Figure 1, for example, shows the top 20 ETEC 500 keywords and bigrams. These findings tend to confirm that the desired focus on research and critical thinking in this course had been achieved.

Whole-curriculum analysis

To compare all courses in our program, I examined course-specific metrics and readability indices generated by *Quantext* for each course. This revealed other programmatic features that prompted critical reflection on program design and expectations of our learners. For example, Figure 2 shows that total word count per course ranges dramatically from ~12,000 to ~59,000 words. If the average reading speed of an adult is ~200 words per minute, this clearly has implications for learner workload (ranging from 5-25 hours/week) - particularly for our audience of busy working professional learners.

The various readability indices computed for our courses meanwhile do not necessarily correlate with course size (that is, 'smaller' courses are not necessarily easier to read). Figure 2 charts the SMOG index against word

⁶ Quantext User Guide: <https://workbench.quantext.org/>

⁷ For readability indices, see: https://en.wikipedia.org/wiki/List_of_readability_tests_and_formulas

count per course. Since SMOG metrics are understood to indicate ‘years of formal education needed for this reading level’, it is evident that the language of our courses ranges from ‘accessible to a high school student’ (10.5) to ‘requires university level education’ (14.0). Findings like these should prompt further reflection by our course authors and curriculum committee on the scholarly level of our courses.

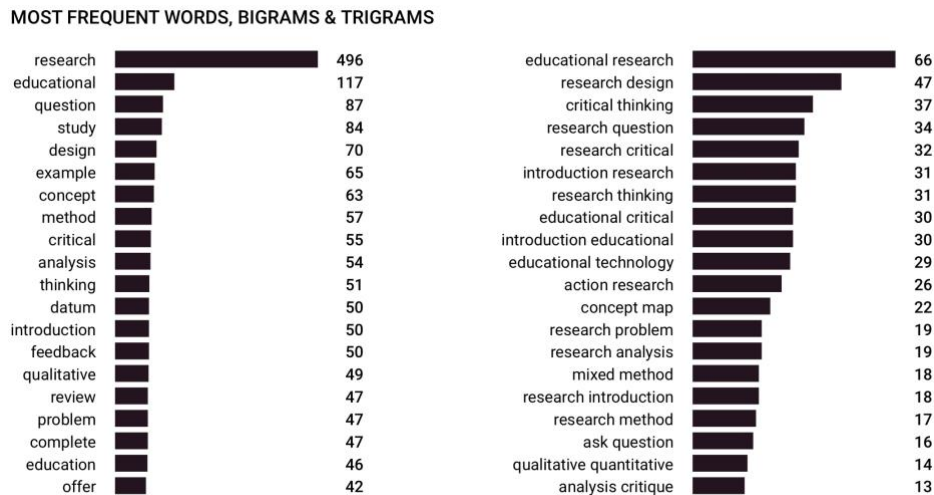


Figure 1. Top 20 keywords and ngrams in course *ETEC 500: Research methodology in education*

Quantext for curriculum analysis?

Overall, analysis of course content using *Quantext* shows promise as a relatively user-friendly approach to curriculum analysis, and permits analysis of ‘whole course content’ (rather than just syllabi). Data preparation can be undertaken in widely available and familiar applications (e.g., MS Excel). Metrics generated are well established and defined in the literature (and already in use in education research). And visual presentations of

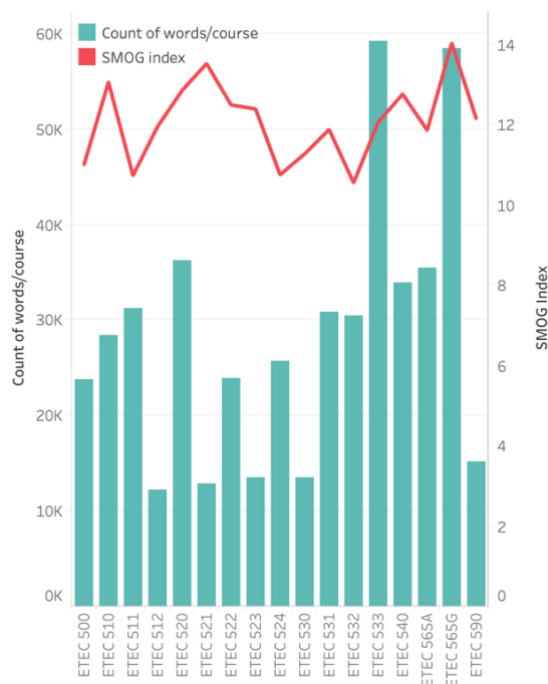


Figure 2. Word count and SMOG index per course

data are clear and easy to interpret. At present, per course results can be valuably compiled, compared and further analysed (albeit manually), to give a big picture overview of frequent curricular topics and networks of relationships between courses, in ways that both individual educators and curriculum committees can interpret and use for decision-making. Two further Quantext analytic features still in development promise further benefits. First, the capacity to assess semantic similarity of course content will allow between-course comparison of content for topic redundancy. Most critically, implementation of the ability to analyse text against a 'reference corpus' will permit comparative analysis of course content against, for example, a definitive textbook in this field of study, or a compilation of published research articles, to allow identification of curricular gaps.

In summary, this pilot study does not describe development of yet another 'state of the art' learning analytics tool or method. Instead, it highlights the novel application of an existing and user-friendly analytics tool to a common educational challenge: analysis of curriculum. Future testing of other existing text analysis tools may expand the potential for educator-led content analytics further.

References

- Bosnic, I., Verbert, K., & Duval, E. (2010). Automatic keywords extraction - a basis for content recommendation. *Proceedings of the 4th International Workshop on Search and Exchange of e-learning Materials (SE@M'10)*, Barcelona, Spain (pp. 51–60).
<http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.369.7493&rep=rep1&type=pdf>
- Dyjur, P., & Kenny, N. (2015). Analyzing curriculum mapping data: Enhancing student learning through curriculum redesign. *Proceedings, University of Calgary Conference on Postsecondary Learning and Teaching*, Calgary, AB). University of Calgary.
- Fiallos, A., & Ochoa, X. (2019). Semi-automatic generation of intelligent curricula to facilitate learning analytics. *Proceedings of the Ninth International Conference on Learning Analytics & Knowledge*, Tempe, AZ, USA (pp. 46–50). New York, USA: ACM. <https://doi.org/10.1145/3303772.3303834>
- Gottipatti, S., & Shankararaman, V. (2014). Learning analytics applied to curriculum analysis. *Proceedings, 2014 AIS SIGED: IAIM International Conference on Information Systems Education and Research*, 2).
<http://aisel.aisnet.org/siged2014/2>
- Herodotou, C., Rienties, B., Verdin, B., & Boroowa, A. (2019). Predictive learning analytics 'at scale': Guidelines to successful implementation in higher education based on the case of the Open University UK. *Journal of Learning Analytics*, 6(1), 85–95. <https://doi.org/10.18608/jla.2019.61.5>
- IBM. (2011). DeepQA project: FAQ. <http://www.research.ibm.com/deepqa/faq.shtml>
- Kawintiranon, K., Vateekul, P., Suchato, A., & Punyabukkana, P. (2016). Understanding knowledge areas in curriculum through text mining from course materials. *Proceedings, 2016 IEEE International Conference on Teaching, Assessment, and Learning for Engineering (TALE)* (pp. 161–168). IEEE.
- Kovanović, V., Joksimović, S., Gašević, D., Hatala, M., & Siemens, G. (2017). Content analytics: The definition, scope, and an overview of published research. In *Handbook of learning analytics* (1st ed., pp. 77–92). <https://www.solaresearch.org/hla-17/hla17-chapter7/>
- McDonald, J., Moskal, A. C. M., Elgort, I., & Gunn, C. (2018). Analysing student responses: Early lessons from a pilot study. *Companion Proceedings of the Eighth International Conference on Learning Analytics & Knowledge*, Sydney, Australia, March 2018). <https://www.solaresearch.org/core/companion-proceedings-of-the-8th-international-learning-analytics-knowledge-conference-lak18/>
- Moore, M. G. (1989). Editorial: Three types of interaction. *American Journal of Distance Education*, 3(2), 1-7. <https://doi.org/10.1080/08923648909526659>
- Rashtian, H., Hashemi, A., & Macfadyen, L. P. (2020). Harnessing natural language processing to support curriculum analysis. *Proceedings, iCERi 2020: 13th Annual International Conference of Education, Research and Innovation*, Seville, Spain). Spain: International Academy of Technology, Education and Development (IATED).
- Richards, J., & Ashbourne, D. (2017). *A guide to curriculum renewal at the University of Toronto*. Toronto, Canada. <https://teaching.utoronto.ca/wp-content/uploads/2017/04/A-Guide-to-Curriculum-Renewal-at-the-University-of-Toronto-2017.pdf>
- Tsai, Y.-S., Moreno-Marcos, P. M., Jivet, I., Scheffel, M., Tammets, K., Kollom, K., & Gašević, D. (2018). The SHEILA framework: Informing institutional strategies and policy processes of learning analytics. *Journal of Learning Analytics*, 5(3), 5–20. <https://doi.org/10.18608/jla.2018.53.2>
- University of Saskatchewan. Curriculum alignment tool (CAT).
<https://teaching.usask.ca/curriculum/curriculum-alignment-tool.php>

- Wei, J., Cutler, F., Macfadyen, L. P., & Shirazi, S. (2019). Implementing learning analytics: Instructor perspectives. *Companion Proceedings, Ninth International Conference on Learning Analytics & Knowledge*, Tempe, AZ, USA, March 2019 (pp. 56–61). <https://www.solaresearch.org/core/companion-proceedings-of-the-9th-international-learning-analytics-and-knowledge-conference-lak19/>
- West, J. (2017). Validating curriculum development using text mining. *The Curriculum Journal*, 28(3), 389-402. <https://doi.org/10.1080/09585176.2016.1261719>
- Wolf, P. (2007). A model for facilitating curriculum development in higher education: A faculty-driven, data-informed, and educational developer-supported approach. *New Directions for Teaching and Learning*, 2007(112), 15–20. <https://doi.org/10.1002/tl.294>
- Wolff, A., Moore, J., Zdrahal, Z., Hlosta, M., & Kuzilek, J. (2016, 2016). Data literacy for learning analytics. *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge - LAK '16*. ACM. <https://doi.org/10.1145/2883851.2883864>
- Wu, P., Yu, S., & Wang, D. (2018). Using a learner-topic model for mining learner interests in open learning environments. *Journal of Educational Technology & Society*, 21(2), 192–204.

Macfayden, L. P. (2020). Content analytics for curriculum review: A learning analytics use case for exploration of learner context. In S. Gregory, S. Warburton, & M. Parkes (Eds.), *ASCILITE's First Virtual Conference*. Proceedings ASCILITE 2020 in Armidale (pp. 42–47). <https://doi.org/10.14742/ascilite2020.0102>

Note: All published papers are refereed, having undergone a double-blind peer-review process. The author(s) assign a Creative Commons by attribution licence enabling others to distribute, remix, tweak, and build upon their work, even commercially, as long as credit is given to the author(s) for the original creation.

© Macfayden, L. P.2020