PROFESSIONAL FILE

ARTICLE 139

Association for Institutional Research

MINING TEXT DATA: MAKING SENSE OF WHAT STUDENTS TELL US

John Zilvinskis
Greg V. Michalski

About the Authors
John Zilvinskis is a doctoral student in the Higher Education and Student Affairs (HESA) program at Indiana University. Greg V. Michalski is Director of Institutional Analytics and Research, Florida State College at Jacksonville.

Abstract

Text mining presents an efficient means to access the comprehensive amount of data found in written records by converting words into numbers and using algorithms to detect relevant patterns. This article presents the fundamentals of text mining, including an overview of key concepts, prevalent methodologies in this work, and popular software packages. The utility of text mining is demonstrated through description of two promising practices and presentation of two detailed examples. The two promising practices are (1) using text analytics to understand and minimize course withdrawals, and (2) assessing student understanding and depth of learning in science, technology, engineering and mathematics (STEM) (physics). The two detailed examples are (1) refining survey items on the National Survey of Student Engagement (NSSE), and (2) using text to create a learning analytics system at a community college (City University of New York [CUNY]: the Stella and Charles Guttman Community College, or CUNY Guttman). Results of this study include identification of additional item choices for the survey and discovery of a relationship between e-portfolio content and academic performance. Additional examples of text mining in higher education and ethical considerations pertaining to this technology are also discussed.

FRAMING THE ISSUE OF TEXT MINING

Students generate copious amounts of thick, rich data; however, these data are often unexamined because traditional qualitative methodologies used to examine thousands of submissions require extensive resources. Text mining (the machine coding of text with the goal of integrating converted submissions with quantitative methods) offers timely, accurate, and actionable assistance (Zhang & Segall, 2010). This article presents detailed information on how text mining can be used by staff who work in institutional research (IR) and collect, but often are forced to neglect, text-based data.

Examples of text data accessible to IR staff are

- Application essays,
- Written assignments,
- Open-ended survey responses,
- Course Management Software (CMS) postings,
- Student blogs,
- Course evaluations,
- Surveys, and
- E-portfolios.

IR professionals recognize the depth of text data that are—or potentially can be—collected, but might be uncertain how to process those data and use them for campus research informing decision-making. This article presents an overview of the concepts and strategies of text mining and text analytics, and acquaints the reader with the terminology, methodology, and software associated with this technology. Two examples using text data in higher education illustrate how mining can be carried out. It is hoped that the reader will develop a fundamental understanding of text mining, be able to suggest how it can be used to aid decision-makers, and understand the advantages as well as the constraints of this approach to data analysis.

BASIC CONCEPTS

Before describing how the researcher would approach text data, it is important to distinguish between data mining information and analytics. Data mining often implies that the researcher is employing algorithms to explore large data sets (Baker & Yacef, 2009; Fayyad, Piatetsky-Shapiro, & Smyth, 1996). Analytics incorporates data mining techniques to create "actionable intelligence," meaning information that guides decision-making (Campbell, DeBlois, & Oblinger, 2007, p. 42). This information is used to predict case-level behavior that will guide intervention (van Barneveld, Arnold, & Campbell, 2012). In education, learning analytics takes the form of collecting real-time data to measure the effectiveness of teaching practices for a particular student, and to suggest intervention in relative real-time in the case that they are not effective (Suthers & Verbert, 2013). As will be described later, the timing of when data are collected, processed, implemented, and acted on is critical from moving data mining to learning analytics (Arnold & Pistilli, 2012). This article uses similar definitions when describing text mining and text analytics.

A variety of related definitions for text mining can be found in the literature. The following definition is an adaptation of data mining that emphasizes the type of data being mined: "the discovery of useful and previously unknown 'gems' of information from textual document repositories" (Zhang & Segall, 2010, p. 626). Other definitions involve the use of specific methodological/technological approaches: "Text mining is a young interdisciplinary field which draws on information retrieval, data mining, machine learning, statistics and computational linguistics" (Singh & Raghuvanshi, 2012, p. 139).

The rationale for text mining stems from "the need to turn text into numbers so powerful algorithms can be applied to large document databases" (Miner et al., 2012, p. 30). According to Miner et al., text mining and text analytics are broad umbrella terms describing a range of technologies for analyzing and processing semi-structured and unstructured text data. Text mining is the practical application of many techniques of analytical processing in text analytics.

There are several ways that researchers can use software to access, label, and study text data. Familiar to anyone who has used an Internet search engine, information access uses recall systems common in most text-mining approaches; however, the absence of generating new information excludes this practice from true text mining status (Hearst, 1999). Another step toward text mining includes the categorization of text (such as categorizing libraries or academic journals) that can be—in its own right—a means to mine text. Similar techniques can be used in the processes of clustering documents and mining Web content. Beyond pulling specific words or clustering proximity of terms, researchers also intend to extract meaning from the text they study. Those who study computational linguistics contribute to text mining by developing algorithms (a subfield called natural language processing) to measure sentiment and meaning from text. Educational data mining can utilize text in a unique way to measure student learning of important concepts or reflecting on developmental milestones (Baker & Yacef, 2009).

Researchers familiar with qualitative research might be skeptical about the effort of those who work in natural language processing and wonder, "Why wouldn't the researcher simply perform traditional qualitative data methods with text data?" Despite the development of intelligent graders and machine learning, computer programs do not have the capacity to interpret the nuance of writing at the level of the human brain. Nonetheless, there are several reasons why text mining is a viable research technique.

First, data sets can include thousands, millions, or billions of text submissions, precluding the use of traditional qualitative techniques. Second, even if the text data were at a size that could be reviewed by researchers, hiring and training coders increases the resources (time and money) needed to complete this process. For most IR departments, hiring a team of coders might not be practical. When analytics projects are used to identify students who need support, the resulting information can come too late to provide that support if data coding, processing, and interpreting takes too long. Having coders introduces the need to account for inter-rater reliability. Third, these data are computed into numeric values so that they can be combined with other sources of data in statistical models built to predict student behavior. Text mining allows for the sizable and efficient processing of text data.

THE PROCESSES OF MINING TEXT

In their book *Practical Text Mining and Statistical Analysis for Non-Structured Text Data Applications*, Miner et al. (2012) present a comprehensive step-by-step guide for text mining. Their text-mining methodology is very similar to other types of research: the researcher has to define the purpose of the project, manage the data (seek, organize, clean, and extract model data), evaluate the results, and, finally, disseminate the results. The aspects of data management (organizing, cleaning, and extracting data) are particular to text mining.

First, the researcher generates a corpus, or a collection of documents or cases in which the desired text exists. During this phase, the researcher removes all nonessential information, such as e-mail addresses or Web links. Second, the researcher either uses an established "stop or include" word list or creates one that filters out words that lack informational value (the, a, his) while highlighting words that do have value. This phase also includes limiting the number of terms by accounting for inflection (plural vs. singular, past vs. present) or the word root (teach for teaching, teacher, teaches). Third, the researcher organizes the terms within cases. This process can be done by using a simple binary notation (1 = term is present) or by noting semifrequent but unique terms (i.e., terms that are featured in some, but not all, of the cases). Another way to organize the terms is by using singular value decomposition (SVD), which reduces the input matrix to a smaller version, representing the variability of each case (Berry & Kogan, 2010; Manning & Schütze, 1999). This latter method is important when working with large text data sets that otherwise could take a long time to process.

Once the researcher has organized the data, she can use several ways to extract important information from these cases (Miner et al., 2012):

- Classification. The researcher creates a dictionary to organize terms based on their definition and hierarchical connection; for example, she might file "classes" under the domain of "education."
- Clustering. The researcher groups terms based on the frequency and pattern of their use, compared to the number of students who use those terms.
- Association. The researcher examines the use of text in connection with some event that is occurring. For example, she might want to compare the positive or negative descriptions of faculty teaching, as reflected in course evaluations before and after final grades are posted.
- Trend analysis. The researcher measures the change of text responses to an identical prompt over time.

SOFTWARE TO CONSIDER

Numerous free and commercial software programs are available for text mining. RapidMiner is an open source analytics program featuring tools that include the analysis of text data. The visual point-and-click nature of the interface allows non–computer programmers to access, clean, and analyze their data. The RapidMiner Web site includes numerous resources, and the user community has posted helpful videos on YouTube. The text extensions include easy-to-use functions such as the ability to group documents based on term frequency. Although the base-level package is free, users might want to purchase the professional package that includes more-advanced options. RapidMiner provides sizable discounts to researchers who use their software, and also provides an extension that allows for the export of data into Tableau. This is an ideal program for IR professionals who want to begin experimenting with text data.

Waikato Environment for Knowledge Analysis (WEKA) is another open source program available for download. WEKA uses a system of algorithms through Java to perform analytic functions; Keyphrase Extraction Algorithm (KEA), on the other hand, is an extension of that project focused on text mining. Both WEKA and KEA were developed at the University of Waikato, New Zealand. WEKA is a premier program for machine learning, a field of computer science that emphasizes the development of programs that can recognize patterns and self-evolve, which is relevant in text mining where programs recognize and adapt to changes in text. KEA has a clean function of recognizing text within a corpus. However, the program does not have a graphical user interface and therefore is most accessible to those with computer science backgrounds. Also, although the WEKA Web site continues to collect publications from researchers using their system, as of this writing the KEA Web site has not been updated for some time and has limited resources.

Most IR professionals are familiar with R as a program, but they might not be fluent in the programming language needed to make it work. R is a premier statistical software, in part because it is free, but also because it has been adopted by a large contingent of statisticians worldwide. There are numerous books and articles on how to use R for text mining; however, between the time it would take staff to become expert in the use of R and to teach themselves how to program the text mining, the software can lose its advantage. Nonetheless, for staff familiar with the program, it can be a great platform to begin to explore and experiment with text data.

The commercial products do distinguish themselves from the free software. IBM SPSS Modeler Premium offers text mining in addition to the analytics suite; the suite includes other data analysis processes including variable selection, various analysis applications, and visualization options. Like most SPSS platforms, the interface is extremely user friendly, probably to a fault, considering how statistical analysis is reduced to a few button clicks. However, for the purposes of text mining, the software is quite advanced. The text function has a built-in dictionary that classifies terms and allows the user to create a dictionary. This function creates a hierarchy of similar terms and places them into broad categories such as "athletics," "emotion," and "education." The application also has built-in dictionaries that include a classification scheme for terms in areas such as customer satisfaction. The natural sorting of terms along with the ease of creating a dictionary allows for comprehensive and easy classification of text data.

SAS® Enterprise Miner™ software also has a user interface and offers comprehensive tools for organizing text and clustering cases.[1] Like IBM SPSS Modeler Premium, SAS Enterprise Miner software is an analytics suite that includes some text-mining applications. However, the way this product distinguishes itself from others is that the clustering function of terms is a component of the text-mining function. Often programs will reduce terms to dichotomous values (not-present vs. present) and then employ clustering methods afterward. SAS Enterprise Miner software allows the user to augment the clustering parameters within the text-mining function and creates data visualizations that are more compelling because of their use of text-mining terms. These features within SAS Enterprise Miner software offer a comprehensive way to cluster terms for analysis.

PROMISING PRACTICES

In higher education there are numerous opportunities to mine text data to predict important student behavior. For example, application essays can be mined to predict student matriculation or even retention. These data can be incorporated into an analytics project aimed at awarding the appropriate amount of financial aid needed to secure a student's enrollment. Another way in which text mining could be used would be to measure virtual student participation in course management software, such as evaluating students' contribution in a course message board within a learning management system like Blackboard or Canvas. Already researchers are using data to model and predict students' academic performance and assign intervention, such as e-mail notification, conversations with advisors, and alerts to faculty (Arnold & Pistilli, 2012). These efforts can be enhanced with the analysis of text data.

Using Text Analytics to Understand and Minimize Course Withdrawals

Two current and promising applications of text analytics involve its application in the areas of student course retention and course outcomes. In mining open-ended comments

[1] SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

captured from students withdrawing from college courses, Michalski (2014) produced and tested a model that succeeded in categorizing over 95% of student explanations for student course withdrawal decisions. This model includes 11 major categories for course withdrawal (i.e., reasons) and corresponds well to existing theoretical and empirical research in the area of college course withdrawal. Of the 11 categories, the top three coding categories were (1) time–schedule, (2) job–work, and (3) personal–other reasons provided by students. Other categories included finances, health, family, course/faculty negative, and online course (mentioned by students who stated their desire to take the same course via instructor-led, rather than online, delivery). Subsequent research (Michalski, 2015) further demonstrated how output from the resulting text model can be statistically analyzed using selected quantitative procedures (including Hierarchical Agglomerative Cluster Analysis, Principal Components Analysis, and Multiple Correspondence Analysis) for validation (i.e., creating clusters of terms describing course withdrawal that are both mathematically and conceptually related). These results are currently being used to develop a course Re-enrollment Assessment Online (REASON) process to minimize course withdrawals, in part by identifying and providing appropriate support, services, advising, and other assistance to encourage and assist student course reenrollment decisions.

Assessing Student Understanding and Depth of Learning in STEM

A second promising example of the application of text mining is the analysis of student responses to open-ended assessment questions at Michigan State University's (2015) Collaborative Research in Education, Assessment and Teaching Environments for the fields of science, technology, engineering and mathematics (CREATE for STEM). There, researchers analyzed student responses to open-ended test questions and related these answers to course outcomes. Within this project at Michigan State, as part of a National Science Foundation grant, Park, Haudek, and Urban-Lurain (2015) used IBM SPSS Modeler to mine the text of short-answer physics test questions about the course topic "energy." The purpose of this study was to explore the degree to which term use is associated with overall knowledge of energy-related constructs. The researchers were able to classify terms used in open-ended responses as either surface-level understanding or scientific ideas. Not surprisingly, students who wrote using scientific ideas were more likely to answer corresponding multiple-choice questions correctly. Innovative uses of text mining like these allow for a more robust understanding of student learning, and assist in test design.

TWO DETAILED EXAMPLES

This article now demonstrates the utility of text mining through presentation of two detailed examples: (1) refining survey items on the NSSE, and (2) relating e-portfolio text to student performance at a community college (CUNY Guttman).

Example One: Classifying Open-Ended Survey Questions

When creating a survey, researchers strive to develop closed-ended items that present all of the possible answers for a given question (Dillman, Christian, & Smyth, 2014). Answering open-ended questions requires more mental energy of the respondent and can lead to answers that are more difficult than closed-ended items for the researcher to process. It can be challenging for institutional researchers to develop a list of all the possible survey responses to a particular question. What do researchers do if they are not sure if all possible responses are presented? A simple answer is to create a text box next to an "other" option where a respondent could write in an answer. Using text mining, the researcher can organize these write-ins to create a more comprehensive list of options.

In the 2014 administration of the NSSE, an experimental item set was developed to identify types of leadership positions held by students. In the development of this survey item set, researchers wanted to know how often students said they had held a formal leadership position in the core survey, as compared to how often they identified serving in a specific

leadership role later in the survey. Of the students who responded to the items, 1,482 of 4,836 (31%) wrote in a leadership position not listed in the core survey item. Text mining was used to determine (1) what new leadership roles should be added to the survey set and (2) the degree to which responses to the leadership item in the core survey related to answers in the experimental item set.
Write-in entries were classified using the text-mining capability of IBM SPSS Modeler Premium. Of the 1,482 students who entered a response, 830 (56%) entries were summarized into eight categories. For example, the concept "teaching assistant" included variations in term, such as "teaching assistant," "teachering assistant," and "teacher's assistant." The top eight leadership positions written into the open-ended text box and their percent of total written comments are shown in Table 1.

The researchers then calculated how well these positions represented the respondents' understanding of a "formal leadership position." In the core survey, respondents were asked if they had completed (or were in the process of completing) a formal leadership position. By comparing the number of students who identified one of these positions to the number of students who said they had participated in a formal leadership position, the researchers had a better understanding of positions that constitute "formal leadership" from the perspective of the respondents. For example, students who wrote in "secretaries," "treasurers," and "editors" were more likely to have said they had completed a leadership role.

Table 1. Counts and Percentage of Write-ins of Additional Leadership Positions Added Through Text Mining

Position	n	%
Tutoring	145	9.8%
Teaching assistant	87	5.9%
Research assistant	60	4.0%
Secretary	55	3.7%
Treasurer	57	3.8%
Mentor	54	3.6%
Member	51	3.4%
Editor	25	1.7%

As a result of this study, the response options "Instructor or Teaching Assistant," "Tutor," and "Editor" were added as leadership positions for the second administration of this experimental item set. Researchers considered installing skip logic that would allow only affirmative responses to the formal leadership item on the core survey to access the experimental item set. However, since a large number of respondents who held positions (both the original and write-in options) had not reported completing a formal leadership experience in the core survey, the skip logic was not included. In this instance, text mining allowed the researchers to expand the item bank and improve the survey design.

Example Two: Clustering E-Portfolio Submissions

As part of a Bill & Melinda Gates Foundation–funded project, researchers with CUNY Guttman were interested in mining first-year e-portfolio introductions to see if student text data could predict academic performance. All students were required to attend a summer bridge program prior to their first semester. During that program, students began their e-portfolio by writing introductions of themselves. Some submissions were informal, reading like an online "shout out" instead of an academic piece of work. Other submissions described ambitions and backgrounds.

Using SAS Enterprise Miner, terms from 163 student e-portfolio introductions were clustered to predict student outcomes such as the number of credit hours earned and GPA. The terms were grouped using a K-cluster analysis within the Enterprise Miner suite; groups of terms were given names by researchers, based on their seemingly shared concept. These concepts, which aligned with student development theory, emerged from the terms; for example, worrying about making friends (e.g., shy, person, friend, know, quiet) and commitment to society (e.g., social, worker, work, believe, help). These results are summarized in Table 2.

Table 2. Results from Text Clustering of Student E-Portfolio Introductions

Concept	Clustered Terms
Connection to family	family, york,* high school, college, child
Learning	class, teacher, art, math, subject
Everyday	know, day, love, life
College participation	high school, school, attend, guttman
Gaming	game, movie, favorite, watch, video
Worrying about making friends	shy, person, friend, know, quiet
Recreation	art, basketball, play, sport, travel
Commitment to society	social, worker, work, believe, help
Technology integration	technology, information, art, health, mind
Aspirations to work in business	guttman, business, manhattan, administration, graduate

Note: * CUNY Guttman is located in Manhattan, so it might be common for students who are describing their connection to family and education to include their connection to New York City.*

Variables were coded dichotomously based on whether the term was contained in a concept's cluster. In an ordinary least squares (OLS) regression analysis, after controlling for college preparation (SAT and writing proficiency scores) and age, a relationship (p = 0.06) was found between the concept "connection to family" and credit hours earned that fall. Students who used terms within this concept earned fewer credit hours than students who did not. Besides identifying the topics that students write about in their e-portfolio introductions, the results of this research identified one concept—connection to family—that predicted the number of credit hours earned. In this case, students who wrote about connection to family earned fewer credit hours than students who did not. This finding is consistent with the Theory of Student Departure, in which Tinto (1987) argues that family obligations can lead to student attrition. The ultimate goal of this project was to explore whether text mining could be used in learning analytics at CUNY Guttman. A limitation of this study is that this institution has a small enrollment, so many of the principles of big data (specifically a large number of cases or students) will not work. However, by identifying which types of information (such as e-portfolio data) are predictive of student success, researchers at this institution can create analytic models that are accurate, using few variables.

IR professionals often have access to text data such as admissions essays, answers to online tests, and e-portfolio submissions, which are often not incorporated in data analysis because they are too difficult to synthesize compared with basic metrics such as SAT score or GPA. However, clustering terms from these corpuses can help to identify themes within these documents. Although clustering terms requires more mathematical training and the naming of the concepts is subjective (as it is in factor analysis), the mathematical grouping of terms provides insight into what students are writing about, while providing units of analysis that can be included in models to predict student success. Clustering student text provides a means to harness these often-overlooked data.

CONSIDERATIONS FOR IMPLEMENTATION

Text mining has the possibility of being a featured tool in analytics projects; however, institutional stakeholders need to be intentional in how they collect and process these data so they can use information in a timely manner. Processes for data collection, distribution, project implementation, and analysis of results must be coordinated. In any analytic project, timing is important. Text data need to be cleaned, processed, and incorporated into predictive models in time for interventions to occur; because of the quick turnaround, most projects include automation.

Another aspect to consider is the data source or origin. The two examples in this article illustrate different sources of text data. In the first example, asking respondents to submit one "other" leadership position seems straightforward and easy to sort; however, there can be issues with classifying even these basic data. For example, some responses do not easily or naturally fit into a category, while others might be placed in more than one category. In the second example, researchers were mining students' introduction statements for specific information, but the prompts might have been too vague. Alternatively, prompts can be too prescriptive. Contrast the prompts, "What do you think it takes to succeed in college?" with "Who will you ask to mentor you so you will be successful in college?" Similar to survey item design, the creation of prompts for text mining might develop into its own science. (For information regarding the intersection of text mining and survey question design, see George et al., 2012.)

Another aspect to consider when using text data for predictive models is the use of sentiment analysis: "Is it only important to know what a student is writing or is it also important to understand how a student feels about a particular topic?" There might be text sources, such as student course evaluations or satisfaction surveys, where it would be important to consider sentiment. In these cases, the sentiment analysis component of customer satisfaction software can be implemented to measure student attitude when describing certain phenomena as positive, negative, or indifferent. IBM SPSS Modeler Premium has a built-in library for measuring customer satisfaction, allowing for a more refined and adjustable view of individual likes and dislikes. Measuring sentiment might be crucial for stakeholders (e.g., administrators and faculty); therefore, for researchers implementing text-mining projects, the analysis of sentiment might be an important aspect when trying to garner buy-in to approve text-mining projects.

ETHICAL CONSIDERATIONS

As we continue to evolve in this era of analytics, it is not uncommon to hear concerns about how data are being collected and used. Students might believe that their intellectual property or even identity has been exploited when researchers use their text data, some of which can be very personal (Slade & Prinsloo, 2013). Because there is much at stake in terms of student response to analytics, institutions using these technologies need to develop sophisticated policies regarding the data ownership, transparency, and security of data (Pardo & Siemens, 2014). Furthermore, institutions are sometimes ill-equipped to grapple with the more complex issues around the use of data within analytics, such as what to do when unexpected outcomes occur, such as predictive models that misclassify student potential (Willis, 2014). All of these factors reflect a campus culture in which administrators act ethically in using data to improve student success, to ensure that students and faculty are valued partners with technologies, and to create an environment that views these technologies as advancing those aims (Willis, Campbell, & Pistilli, 2013). These concerns over analytics also relate to the use of text mining.

One of the benefits of text mining is that it presents a bridge between the atheoretical process of data mining/analytics and the more theoretically driven approaches for research used in higher education. It can be unnerving for campus stakeholders, including faculty, administrators, and students, to rely on the black box of analytics. Text mining raises ethical concerns about the results of analytics processes. What is the implication when an algorithm predicts that a student is less likely to pass a course or persist within a particular major? What aspects of a student's background and institutional environment are either being overlooked or considered in making these predictions? Furthermore, institutional policy is traditionally built on empirical studies of students that reference some theoretical framework, or at least a plausible hypothesis. For example, first-generation college students are disadvantaged, and therefore administrators argue that they should receive additional resources as a matter of policy. Gathering stakeholders to support this one aspect of identity is easy because it is an understandable concept: "Students who did not grow up in a home where at least one parent completed college are at a disadvantage compared with students who did." However, in an analytic project, first-generation status might be one variable among many to predict student success. These complex models replace an understandable narrative such as the one related to first-generation status with a narrative that is reliant on numerous factors. On the other hand, the results of text mining can be easily interpreted. In the second example of this study, there was an inverse relationship between a student mentioning connection to family and earning credit hours. An educator meeting with students and reviewing e-portfolio introductions would find it easy to act on this information.

There are other institutional cultural aspects to consider when implementing text mining programs on campus: Students could react to the common knowledge that all of their submissions on the campus management system are being mined. Students could change their responses once they realize that their text is being mined. In the second example of this study, it might be the case that upper-class students will warn incoming first-years, "Make sure not to mention connection to your family in your e-portfolio introduction; otherwise, you'll have to schedule an additional meeting with your advisor."

When considering equity, institutional stakeholders need to take the increasingly heterogeneous nature of student demography into consideration. Researchers need to account for diverse levels of familiarity with English, resources, and prior education when working with student text data. Simply relying on text data without regard to student background variables can mislead. Students who have better educational preparation or more resources may be advantaged in the types of phenomena they describe in text. If so, text mining could be used to reinforce inequity within higher education. These aspects, though not the focus of this article, ask whether the researcher should mine text data. Researchers need to consider that question at length before they consider whether they are able to mine text data.

PUTTING TEXT MINING TO USE

This article describes the ways in which text mining can be implemented in the work of those institutional researchers who are asked to create analytics programs on their campuses. The article described two promising application areas, and detailed two examples of text mining projects. Important considerations for any text-mining project are the context and timing of text collection. Although text mining offers several promising avenues within higher education, ethical considerations should provoke deep questions about the appropriateness of these methods balanced with the needs of the institution.

Like many projects within an IR shop, successful text mining requires multiple areas of expertise (Haight, 2014). First, expertise in data acquisition and management is needed to procure and then clean data (e.g., to remove extraneous text) while pairing data with other sources of student information. Second, someone must mine and model text data into predictive schemes that are statistically sound. Third, expertise in graphic design or data visualization is needed to successfully communicate the connection between data source, usable text, and student outcomes. Fourth, as described earlier, the use of automation (i.e., creating systems to automatically collect, clean, and process data) is paramount in any analytics process, so a team member needs to be in charge of this aspect of the project. Finally, text-mining projects need a team champion to advertise and demonstrate the value of this technology and to promote it to campus stakeholders. These skills are essential to integrating text mining into campus decision-making. Text mining presents a microcosm of other IR processes while offering an exciting way to make use of dense text data waiting to be unlocked. As a field, IR is in a unique position because text data have been collected for decades and the complex decision-making on college campuses can only be further informed by the inclusion of those data.

REFERENCES

Arnold, K. E., & Pistilli, M. D. (2012, April). Course signals at Purdue: Using learning analytics to increase student success. In *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge* (pp. 267–270). doi:10.1145/2330601.2330666.

Baker, R.S.J.D., Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. *Journal of Educational Data Mining*, 1(1), 3–17.

Berry, M. W., & Kogan, J. (Eds.). (2010). *Text mining: Applications and theory*. Hoboken, NJ: Wiley & Sons.

Campbell, J. P., DeBlois, P. B., & Oblinger, D. G. (2007). Academic analytics: A new tool for a new era. *EDUCAUSE Review*, 42(4), 40–57.

Dillman, D. A., Christian, L. M., & Smyth, J. D. (2014). *Internet, phone, mail, and mixed-mode surveys: The tailored design method*. Hoboken, NJ: Wiley & Sons.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI Magazine*, 17(3), 37–54.

George, C. P., Wang, D. Z., Wilson, J. N., Epstein, L. M., Garland, P., & Suh, A. (2012, December). A machine learning based topic exploration and categorization on surveys. In *11th International Conference on Machine Learning and Applications* (vol. 2, pp. 7–12). doi:10.1109/ICMLA.2012.132

Haight, D. (2014). *The five faces of analytics*. Dark Horse Analytics. http://www.darkhorseanalytics.com/blog/the-five-faces-of-analytics

Hearst, M. A. (1999, June). Untangling text data mining. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics* (pp. 3–10). doi:10.3115/1034678.1034679

Manning, C. D., & Schütze, H. (Ed. (1999). *Foundations of statistical natural language processing*. Cambridge, MA: MIT Press.

Michalski, G. V. (2014). In their own words: A text analytics investigation of college course attrition. *Community College Journal of Research and Practice*, 38(9), 811–826.

Michalski, G. V. (2015). *Using analytics to minimize student course withdrawals*. Paper presented at the Association for Institutional Research (AIR) Annual Forum, Denver, CO, August. http://www.forum.airweb.org/2015/Documents/Presentations/1095_6bd9881f-0d68-457f-8557-8da3d450cbb8.pdf

Michigan State University. (2015) *Home page*. CREATE for STEM Institute. http://create4stem.msu.edu/

Miner, M., Delen, D., Elder, J., Fast, A., Hill, T., & Nisbet, B. (2012). *Practical text mining and statistical analysis for non-structured text data applications*. Waltham, MA: Academic Press.

National Survey of Student Engagement (NSSE). 2014. NSSE 2014 experimental items codebook formal leadership. http://nsse.indiana.edu/pdf/exp_items/2014/NSSE%20 2014%20Exp_FOL_Codebook.pdf

Pardo, A., & Siemens, G. (2014). Ethical and privacy principles for learning analytics. *British Journal of Educational Technology*, 45(3), 438–450.

Park, M., Haudek, K., & Urban-Lurain, M. (2015). Computerized lexical analysis of students' written responses for diagnosing conceptual understanding of energy. In *National Association for Research in Science Teaching (NARST) 2015 Annual International Conference*, April. http://create4stem.msu.edu/publication/3361

Singh, P. D., & Raghuvanshi, J. (2012). Rising of text mining technique: An unforeseen-part of data mining. *International Journal of Advanced Research in Computer Science and Electronics Engineering*, 1(3), 139–144.

Slade, S., & Prinsloo, P. (2013). Learning analytics ethical issues and dilemmas. *American Behavioral Scientist*, 57(10), 1510–1529.

Suthers, D., & Verbert, K. (2013). Learning analytics as a middle space. In Proceedings of the *Third International Conference on Learning Analytics and Knowledge* (pp. 1–4). Leuven, Belgium, April 8–12.

Tinto, V. (1987). *Leaving college: Rethinking the causes and cures of student attrition*. Chicago: University of Chicago Press.

van Barneveld, A., Arnold, K. E., & Campbell, J. P. (2012). Analytics in higher education: Establishing a common language. *EDUCAUSE Learning Initiative, 1,* 1–11.

Willis III, J. E. (2014). *Learning analytics and ethics: A framework beyond utilitarianism*. EDUCAUSE Review. http://er.educause.edu/articles/2014/8/learning-analytics-and-ethics-a-framework-beyond-utilitarianism

Willis, J. E., Campbell, J. P., & Pistilli, M. D. (2013). *Ethics, big data, and analytics: A model for application*. EDUCAUSE Review. http://er.educause.edu/articles/2013/5/ethics-big-data-and-analytics-a-model-for-application

Zhang & Segall, 2010 Zhang, Q., & Segall, R. S. (2010). Review of data, text and web mining software. *Kybernetes*, 39(4), 625–655.