The Curriculum Journal

ISSN: 0958-5176 (Print) 1469-3704 (Online) Journal homepage: http://www.tandfonline.com/loi/rcjo20

Validating curriculum development using text mining

Jason West

To cite this article: Jason West (2016): Validating curriculum development using text mining, The Curriculum Journal, DOI: 10.1080/09585176.2016.1261719

To link to this article: http://dx.doi.org/10.1080/09585176.2016.1261719

THE CURRICULUM JOURNAL, 2016
http://dx.doi.org/10.1080/09585176.2016.1261719

Routledge
Taylor & Francis Group

Validating curriculum development using text mining

Jason West

Bond Business School, Bond University, Gold Coast, QLD, Australia

ABSTRACT

Interdisciplinarity requires the collaboration of two or more disciplines to combine their expertise to jointly develop and deliver learning and teaching outcomes appropriate for a subject area. Curricula and assessment mapping are critical components to foster and enhance interdisciplinary learning environments. Emerging careers in data science and machine learning coupled with the necessary graduate outcomes mandate the need for a truly interdisciplinary pedagogical approach. The challenges for emerging academic disciplines such as data science and machine learning center on the need for multiple fields to coherently develop university-level curricula. Using text mining, we empirically analyze the breadth and depth of existing tertiary-level curricula to quantify patterns in curricula through the use of surface and deep cluster analysis. This approach helps educators validate the breadth and depth of a proposed curriculum relative to the broad evolution of data science as a discipline.

ARTICLE HISTORY
Received 19 September 2016
Accepted 10 November 2016

KEYWORDS
Text mining; natural language processing; curriculum development; hierarchical clustering; data science; machine learning

Introduction

Learning and teaching in the emerging fields of data science and machine learning requires the coordinated efforts of a range of specialist teachers in disciplines that span mathematics, statistics, computer science, information system architecture, design and visualization. As more universities consider interdisciplinary initiatives to address the growing interest in these disciplines, systematic quantitative and qualitative studies of viable structures, resources, incentives, effects and perceptions need to be devised. An interdisciplinary approach to learning and teaching in data science is critical for its evolution into disciplines in their own right.

To cope with the increasing demand for multidisciplinary learning to cater for the commercial realities of industrial practice in an increasingly creative economy, a more consilience-type of learning and teaching model with integrated operational tools that enables applications to a variety of tertiary education settings are needed. Multidisciplinary curriculum development is a critical contributor to ensure that courses are developed with true interdisciplinary objectives in mind. But the rapid pace of curriculum development in the field of data science has meant that curricula across universities has largely developed in line with the internal disciplinary strengths of each institution rather than in response to

CONTACT Jason West jwest@bond.edu.au

the needs of graduates. The result may be a set of curricula across institutions that are unintentionally biased towards existing disciplinary capability. This study aims to uncover sources of bias using text mining techniques to examine the curricula across universities offering data science degrees.

The value of text mining as a component of natural language processing is that it may improve educators' ability to construct a cohesive curriculum for rapidly evolving disciplines. Often the pace of change in fundamental ideas that underpin the discipline surpasses the development of tertiary-level curricula. Identifying meaningful patterns and trends in text data is an evolving capability that has been applied in biostatistics and engineering education to great effect (Kao & Poteet, 2007).

Text mining is an algorithmic process of extracting meaningful information from text in an effort to uncover linkages between text objects, usually documents. Unlike conventional data mining tasks that extract patterns from structured databases, text mining is intended to explore relationship among objects, generally stored in unstructured formats.

Analyses of large corpora are increasingly used to enhance ontology development. Some scholars have used automated text extraction to identify relationships among terms across domains to help build ontologies in specific fields (Feldman & Sanger, 2007; Lee, Lee, Seol, & Park, 2008). For instance, machine learning in ontology and vocabulary research has yielded credible results in building relationships based on key pharmacogenomic entities (Coulet, Shah, Garten, Musen, & Altman, 2010) as well as the discovery of abbreviations and definitions in medicine (Kuo, Ling, Lin, & Hsu, 2009) and personalized online learning through intelligent tutoring systems (Stankov, Rosic, Zitko, & Grubisic, 2008). Our approach will assist with the development of a data science ontology through text mining analysis. The text mining process not only applies a range of algorithms to detect relationships between clusters of text, but it also provides a coherent environment for file processing, text parsing, transformation, dimension reduction and document analysis.

In this context, we break the text mining process down into three steps. First, we develop a relatively structured database from unstructured text inputs, second we examine patterns and trends from the quasi-structured data, and finally we evaluate patterns and trends to quantify outcomes.

In this study, we identify curriculum development trends and patterns in the evolving discipline of data science using the natural language processing techniques of frequency and cluster analysis. We demonstrate that this approach can assist educators identify, design, implement and improve curriculum through an integrated approach. We show that a by-product of this approach is that the curricula can be objectively tested to ensure it maintains an interdisciplinary perspective, unconstrained by single discipline bias. We further demonstrate that the inductive analyses of curriculum text data may also discover clusters of concepts that would not have been identified through the normal course of curriculum development.

Data and methodology

Statistical text mining

Statistical text mining (STM) uses inductive or data-driven algorithms that do not rely on large controlled vocabularies to examine the underlying nature of language. Controlled

vocabularies refer to the set of schemes that mandate the use of predefined and externally authorized terms that have been preselected by a scheme sponsor. For instance, libraries consist of controlled vocabularies such as subject heading systems or tags to define a catalogue. In contrast, natural language vocabularies offer no such restriction. The benefit of text mining comes with the large amount of valuable information latent in texts which is not available in classical structured data formats for a range of reasons. A key reason is that text is typically the default way of storing and relating many types of information. For example, text-based curriculum summary documents are the main form of communication to educators and students alike to gauge the breadth and depth of a specific course. Other reasons include practicality and cost constraints which prohibit the conversion of text into more structured formats (like tables). The development of automated approaches that extract specific concepts from textual summary documents that do not rely on pre-defined terminology offers a powerful alternative to either unreliable administrative data or labor-intensive intuitive manual review.

The seminal study of different systematic combinations of term frequency-based weightings, normalization terms and corpus-based statistics is Salton and Buckley (1988). Others include the BM25 ranking function (used by search engines to rank matching documents according to their relevance to a given search query) by Robertson and Walker (1994) which describes a score for each word/document pair and Wan (2007) which first decomposes each document into a set of subtopic units and then measures the effort required to transform a subtopic set into another. Recently, scholars have used a componential counting grid (Perina, Jojic, Bicego, & Truski, 2013) which merges latent Dirichlet allocation (LDA) (Blei, Ng, & Jordan, 2003) with counting grid models (Jojic, Perina, & Murino, 2010) allowing 'topics' to be mixtures of word distributions. The LDA probabilistically groups similar words into topics and represents documents as a distribution over these topics while the counting grid approach models documents as a mixture of word distributions. The combination of these two approaches exploits the strengths of each while largely eliminating their weaknesses.

Typically, STM analyses focus on maximizing the performance of prediction models. An iterative process is employed to test various models and find the one that is most predictive for the targeted outcome. Two distinct methods are commonly used to identify terms for inclusion in controlled vocabularies or ontologies. First, multimodel term scoring uses a fairly large group of different models to uncover terms, with a synthetic score ranking terms across all models. The second method, generally referred to as iterative term refinement, uses a single predictive model (selected after parameter tuning) to produce an acceptable set of terms. An iterative cycle of review is then used to remove terms (via start/stop lists) while allowing additional terms to be discovered in subsequent model building steps. We employ the second method in this analysis due to the higher stability gained for both the degree of term frequency and cluster analysis achieved through the iterative term refinement process. We represent text documents as a weighted point cloud of embedded words. The distance between two text documents is the minimum cumulative distance that words from one document need to travel to match exactly the point cloud of another document.

The aim is to arrive at a list of terms, informed by clusters, that best represents the span of subject matter for educators focused on micro-level curriculum development in the discipline of data science. We thus adapt the mining architecture to detect clusters of terms

that can be collated to assess the learnability of a curriculum as a development tool, or as a curriculum validation tool.

Method and hypotheses

We use the package 'tm' in the statistical software R for the analysis. The 'tm' package uses the concept of a so-called source to encapsulate and abstract the document input process. The software implements several algorithms for text mining, providing a supportive environment for file processing, text parsing, transformation, dimension reduction and document analysis. This allows one to work with standardized interfaces within the package without necessarily interacting with the internal structures of input document formats. We believe that the inductive analyses of the data may discover concepts and relationships that would not have been identified by the curriculum developer manually sifting through available curricula.

We hypothesize that language processing searches would better detect clusters of skills, knowledge and learning concepts than intuitive assessments of curricula for developing comprehensive and targeted programs in the field of data science and big data analysis. We also hypothesize that clusters of both surface and deep-level content within the curricula can be better detected using STM techniques.

Data

We collected textual information from academic curricula in data science, machine learning and big data courses through crawling and constructed a database for analysis. A web crawler is an automated process that commences with a narrow list sites to visit and then identifies hyperlinks from these pages to add to a list of additional sites to visit. The crawler archives the websites it copies and saves the information as it goes. The linkages between institutions offering data science programs were exploited through this process. Through the crawling process we extracted the subject objectives, overviews and descriptions from 320 university-level subject curricula through publicly available websites. We then segmented the analysis into groups for categorization. The texts used in this example are multiple websites related to data science, machine learning and big data analytics subjects that were manually identified and then automatically scraped and converted into a text document. Web scraping is application programming interface (API) to extract text data from a website for a broad range of formats. The text example was chosen because it most closely matches the approach used in other qualitative data analytics mash-up methods. Any form of data combined with other internal information and outside sources is a data 'mash-up.'

Curriculum summaries were transformed into structured data using regular expressions that recognized strings of text, such as 'Bayesian analysis' and 'linear regression', but did not account for the syntax in which the term was identified. Text documents were also mapped to phrase and sentence strings allowing inclusion in the rules string searches of colloquial terms or ordering of expressions not yet recognized by the language processing tool vocabulary. Search queries were also constructed using structured data from university subject databases.

Upon loading the documents into the text database a degree of pre-processing is needed for the texts to remove numbers, capitalization, common words, punctuation,

gaps and other fundamental edits for analysis. This was time consuming but essential for high quality analysis.

We then perform a stemming operation on the corpus which is the process of erasing word suffixes to retrieve their radicals. This technique reduces complexity without any severe loss of information for our application (i.e. a form of 'bag-of-words' analysis). We apply the general form of a stemming algorithm from Porter (1980) which removes common morphological and inflectional endings from English words. The R packages 'Rstem' and 'Snowball' are used to implement this stemming algorithm in combination with the 'tm' package.

Finally, we removed all remaining spaces, converted the text to lower case and eliminated pertinent stopwords (common words in English with information value is almost zero, as measured by their low entropy value to the corpus).

Initial results

Frequency analysis

To assess word frequency, we use text mining to initially assess the most and least frequently occurring words. First, we use count-based evaluation (terms with the highest occurrence frequencies) which is a simple approach to derive a corpus that can be interpreted easily and is computationally inexpensive (Davi et al., 2005). The use of a term-document matrix (Shawe-Taylor & Cristianini, 2004) is then exported from the corpus and used as a bag-of-words mechanism, where the order of tokens (i.e. reference terms) is irrelevant. The result is a matrix with document identifiers as rows and terms as columns, with matrix elements being the term frequencies. A histogram of the frequency of words and phrases for the tertiary-level data science and machine learning curriculum corpus is provided in Figure 1.

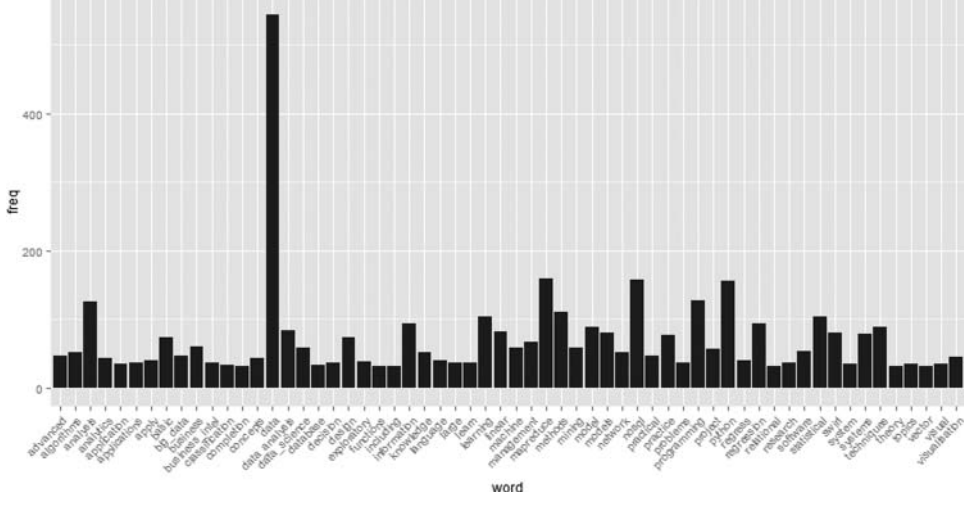

Figure 1. Histogram of text-mining for tertiary-level data science curricula filtered by frequency >40 (frequency normalized to one-tenth of actual frequency).

Unsurprisingly the most common word throughout the corpus is 'data' followed by a relatively broad set of terms that define the main analytical techniques that align to a common data science ontology. In contrast however, the broad array of programming languages common to data science and machine learning courses dilutes the frequency result somewhat with very few registering common usage throughout all tertiary-level courses. In our analysis, we observe that programming languages appear far less frequently than principal analytical techniques, and even appear less frequently that some of the emerging techniques on the fringe of data analytics. The choice of coding language thus appears to be a relatively flexible element in existing data science curricula.

We next find associations for given terms which are simply a further form of count-based evaluation. This is especially interesting when analyzing a text to assess the relative interaction between data science techniques and approaches to machine learning. Analytically, we obtain this by computing correlations between terms in the term-document matrix using a threshold level for correlation of 0.98 in order to identify valid associations.

For certain terms that are of particular importance to the analysis of data science curricula, this approach will help identify the words that most highly correlate with certain terms. If words always appear together then the correlation will equate to 1.0. Table 1 highlights the most highly related terms in the text documents that derive the corpus. These terms represent a loose association of subject offerings that generally appear in all curricula that was reviewed in the data collection phase. The generic nature of most of these terms suggests that a large proportion of data science and machine learning subject offerings embedded in the curricula are relatively generic and few specialist technical methodologies are defined in course overviews. The word count-based analysis of data science and machine learning curricula highlights a general degree of similarity among the majority of courses on offer, suggesting that courses are replicating each other despite the lack of an accrediting industry body in these emerging fields of inquiry.

We next use the term frequency-inverse document frequency (TF-IDF statistic) tool for extracting keywords from an individual document (i.e. a curriculum published by an institution) by considering all documents from our corpus. The results are illustrated in Figure 2. For the TF-IDF statistic, a word is important for a specific document if it shows up relatively often within that document but rarely appears in other documents of the corpus (Salton & Buckley, 1988). However, this approach to quantification of a term's relevance is limited. For instance, a term may show up in many documents of the corpus and also be a central term in a single document. Alternatively, a subject may be covered in several

Table 1. Terms with correlation exceeding 0.98 across documents in the tertiary-level data science curricula corpora (terms are in alphabetical order and unrelated to column location).

Highest correlated terms across curricula			
advanced	data_analysis	methods	python
algorithms	data_science	mining	R
analysis	design	model	regression
analytics	information	models	regression_tools
apply	knowledge	network	software
basic	learning	nosql	statistical
big_data	linear	practical	swirl
business	machine	practice	systems
concepts	management	programming	techniques
data	mapreduce	project	visualization

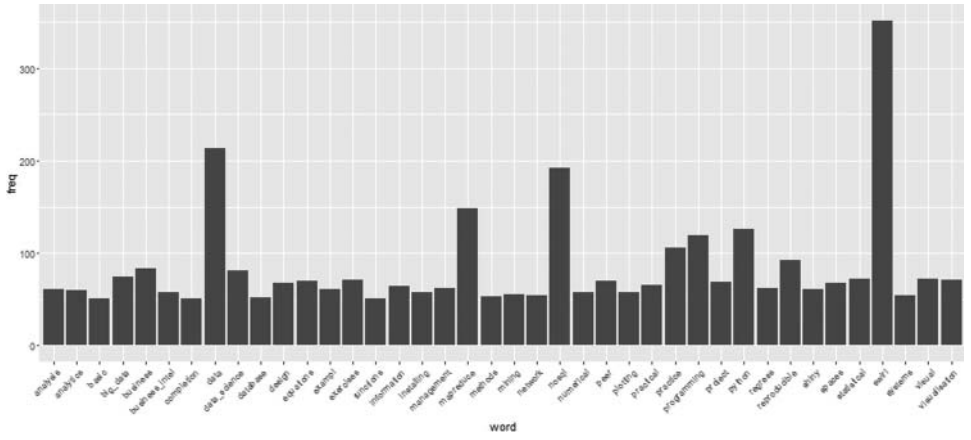

Figure 2. Histogram of text-mining for tertiary-level data science curricula using the term frequency-inverse document frequency approach filtered by frequency >40 (frequency normalized to one-tenth of actual frequency).

documents because it is an important term, but this algorithm penalizes the term for that very reason.

Comparing the results in Figures 1 and 2 reveals that data science and machine learning applications (i.e. software, platforms, networks, etc.) are quite specific to institutions while analytical techniques (i.e. statistical approaches, database design, graphing methods) are more generic and common across curricula. The fundamental structure of analysis across curricula is thus fairly common while the tools and platforms used to conduct the analysis are unique to each curriculum. Institutions therefore appear to develop curricula aligned with other institutions while the actual hands-on tools used for analysis are specific to the academic strengths and preferences of each faculty.

Cluster analysis

To differentiate and assess the level of grouping of both surface and deep-level content within curricula, a cluster analysis is needed. There are two main types of clustering methods available for text mining applications. First is a hierarchical method manifest in the form of a dendrogram which is a tree where the leaves are the data objects and the branches are defined as clusters. More formally, a dendrogram is a nested list of lists with attributes. Second is a partitioning method that divides the data into *k* non-overlapping clusters, so that objects of the same cluster are close to each other and objects of different clusters are dissimilar. The output from the partitioning method is simply a list of clusters and their objects, and for the purposes of testing our hypotheses is the preferred method. While this approach can offer a richer array of interactions, the results are generally difficult to interpret.

However, one way around this limitation is the use of a graphical display to better describe the objects with their interrelations while simultaneously defining the clusters. This allows one to picture the size and shape of the clusters as well as their relative position. Following Kaufman and Rousseeuw (1990), we construct a cluster plot of the data. We then perform a hierarchal clustering analysis by first calculating distance between

words and then clustering them according to similarity. The choice of the distance measure significantly influences the outcome of hierarchical clustering algorithms. Common similarity measures in text mining are metric distances, cosine measure, Pearson correlation and extended Jaccard similarity (Strehl, Ghosh, & Mooney, 2000). We use a generic custom distance function for each of the term-document matrices.

Fundamentally, hierarchical cluster analysis is a multivariate statistical method uses a series of nested correlation calculations (in the form of distance measures) to reorder a dataset such that 'clusters' of data patterns are closest to each other in a list (Rencher, 2002). The hierarchical clustering analysis is visualized in the form of a dendrogram where several groups can be simply categorized. For each iteration of the algorithm, the cluster tree 'grows' as the first level of clusters are connected to other clusters hierarchically. This continues until the entire dataset is represented as a single cluster. Within a cluster tree, clusters of cases and clusters of clusters can quickly be identified by the closeness of lines corresponding to cases and linked to other cases. The unit length of the horizontal line indicates similarity of patterns, the distance in the data space between the two clusters is in the units of the measure, with a shorter line denoting higher similarity.

We initially use an algorithmic solution to identify clusters but a different number of groups can easily be implemented. When extracting the attributes from the dendrogram's nodes the algorithm conducts a 'depth-first search' which traverses the tree data structure from the root and explores as far as possible along each branch before backtracking. The analysis shows larger versus smaller aggregations of terms.

We also use a *k*-means clustering method to cluster words into groups. This is measured as the sum of squared distances between individual words and one of the group centers. Using the classical *k*-means algorithm (Pison, Struyf, & Rousseeuw, 1999) we use the term-document matrix representation of the collected texts to provide valid input for the analysis. We used a classical linear *k*-means clustering with $k = 5$ due to this being the optimal number of clusters derived from aggregating clusters in the term document matrix using an optimization function within the 'tm' package.

A dendrogram of tertiary-level curricula in data science and machine learning is provided in Figure 3. The clusters differentiate between the general reference to 'data' and the application of analysis to data. Furthermore, advanced analytics methods relating to Bayesian analysis, logistic regression and support vector machines are differentiated from simpler analytical techniques and applications using algorithms and other learning methods. Interestingly, programming languages and dedicated technologies (e.g. database methods) again featured less prominently in the branch and leaf analysis. This suggests that the analytical methods to manipulate data are well established in many curricula while the choice of programming language, platforms and other algorithmic methods is less well-defined, but for this analysis its utility is in illustrating the proportion of clusters.

From this representation, we can see the grouping of analytic methods into sub-groups of basic and advanced techniques, applications, model processes and systems of learning.

A second dendrogram for tertiary-level curricula in data science and machine learning is provided in Figure 4. Here, the clusters are grouped in a fanned fashion to highlight the aggregation of various topics within well-defined clusters as well as the unbiased distance between each notable topic across curricula. The circular layout can be a more efficient way of visualizing a large amount of information, but for this depth of analysis both approaches yield equally-valid insights. Again, the grouping of analytic methods into

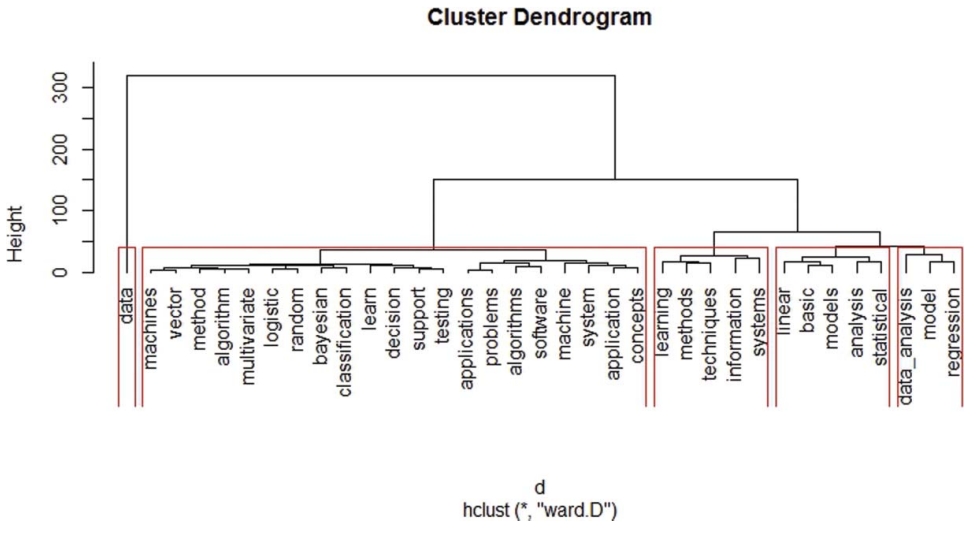

Figure 3. Cluster dendrogram of tertiary-level data science curricula.

Figure 4. Cluster fan dendrogram of tertiary-level data science curricula.

sub-groups of basic and advanced techniques, applications, model processes and systems of learning is apparent.

Cluster plot

While bivariate datasets are relatively easy to be represented in cluster formats, more complex datasets have more than two variables or they come in the form of inter-object dissimilarities. To circumvent this limitation, we employ the CLUSPLOT graphical display

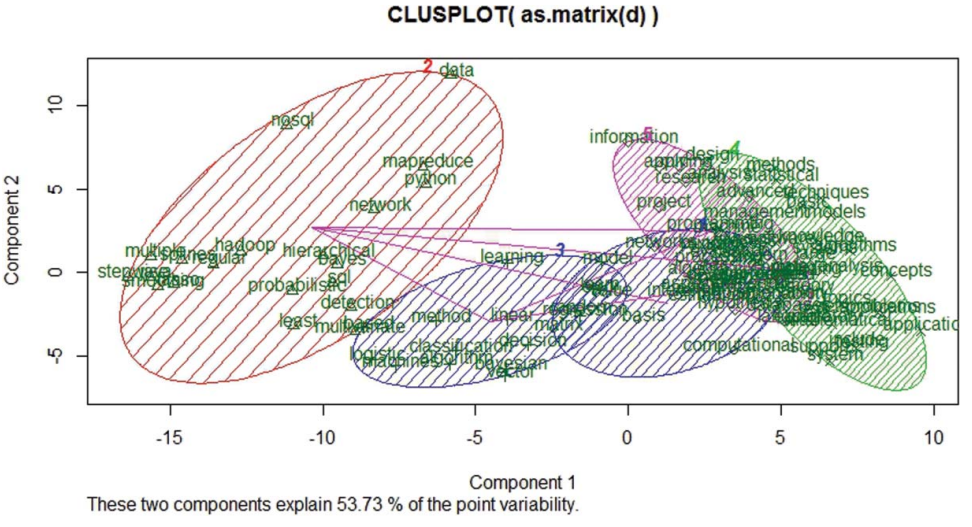

Figure 5. CLUSPLOT of tertiary-level data science curricula.

to represent objects as points in a bivariate plot along with clusters being represented as ellipses of various dimensions (Pison et al., 1999). The ellipses are based on the average and the covariance matrix of each cluster, and their size is such that they contain all the points of their cluster. This construction is confirmed by the presence of an object on the boundary of each ellipse. It is also possible to draw the spanning ellipse of each cluster, i.e. the smallest ellipse that covers all its objects. However, for the purposes of this analysis, we have excluded this mechanism since we are searching for general clusters of curriculum objects rather than well-defined families of objects.

The CLUSPLOT in Figure 5 provides a two-dimensional representation of the objects and the spanning ellipses of the clusters. The boundary of a spanning ellipse always contains at least two objects and the distance between two clusters can be represented as a line connecting the cluster centers.

This analysis reveals that almost 55% of the point variability is explained by the first two principal components, which means that the plot is a relatively faithful representation of the four-dimensional data. Each cluster is provided with a 'cluster number.' The figure shows the five clusters obtained by the *k*-means method for $k = 5$, which was chosen because it represented the best compromise between cluster breadth and compactness, and it aligns with the dendrogram results for direct comparison.

The CLUSPLOT approach was initially devised as a diagnostic tool for the validity of a cluster of concepts or objects. In this study, a similar diagnostic analysis can be performed. The distance between clusters indicates the dissimilarity between groups in an n-dimensional space while the compactness of each cluster indicates the homogeneity within groups.

The clusters reveal that certain elements of curricula can be arranged into a meaningful taxonomy into which homogenous groups from curricula summaries are related and co-occur. The commonality of applications, simple analytical tools, advanced analytical tools and programming methods indicates that some curricula among universities are relatively homogenous and that curricula among institutions appear to largely mimic each other.

The CLUSPLOT also reveals the overlap areas between clusters which represent areas where cluster membership is ambiguous. For instance, there is a great deal of ambiguity between the advanced analytics cluster and the application cluster around certain terms like databases, probability theory, software, statistics and algebra.

Discussion

Strengths and weaknesses of a text mining approach

There are obviously a range of techniques that can derive 'meaning' from text. Currently, the most accurate is the use of a skilled human being simply reading the text and interpreting its meaning. While human capability is accurate and possesses the ability read deeper meaning into the intent of narrative, it is also the slowest and most costly method. Importantly, it is also subject to the risks of bias and misinterpretation. This limitation cannot be completely eliminated but it can be minimized through natural language processing. Text mining offers a better return on investment than a human interrogator due to its ability to analyze and quantify clusters in text data rapidly in some circumstances it may only deliver a crude indicator of truth and trends. The quality of the results generated through automated text analysis can be augmented through manual verification.

While natural language processing offers an objective perspective of curriculum breadth and depth, it is unlikely to fully replace the engagement of experts to interpret and consolidate their thoughts on curriculum design. Current text mining capabilities in this context would be better served to validate the human interpretative approach rather than to adjudicate curriculum quality alone.

The approach does suffer from other limitations. Methods using phrase-matching algorithms typically have a higher sensitivity but lower specificity than an approach using natural language processing in medical studies (Luther et al., 2011). This may be applicable to the identification of clusters of learning concepts to offer a richer suite of terminology for framing curriculum development.

An alternative is a kernel-based clustering method which, like kernel *k*-means, uses an implicit mapping of the input data into a high dimensional feature space defined by a kernel function. The primary advantage of this approach is that the algorithmic processing is far less computationally expensive than if operating directly in the feature space which permits analysis to be conducted with high-dimensional spaces, including natural texts, and can contain several thousand term dimensions. A richer understanding of the underlying corpora across courses related to data science may thus be possible to assist with validating the comprehensiveness of a curriculum.

Evolving disciplines

Shulman (1986) argues that teachers must not only be capable of defining for students the accepted truths in a domain, the must also be able to explain why a particular proposition is deemed warranted, why it is worth knowing, and how it relates to other propositions. Notwithstanding industry accreditation body requirements for many courses, curriculum development is a necessary component of this process. The view of what is

meant by big data, machine learning or data science is fragmented due to its emergence as an evolving discipline. Its construction, by necessity, draws upon the expertise of several existing disciplines to advance new ideas and techniques. New undergraduate and postgraduate programs are introduced regularly, and they have their own notions of what is meant by those terms. They also have their own notions of what students need to know to be proficient in data-intensive work. As such there is no agreement on what should comprise the core subjects in data science. For instance, it is clear that training in data science and machine learning requires a multidisciplinary foundation that includes at least computer science, data analytics, statistics and mathematics. But exactly how they should be developed and delivered in a coherent framework to develop graduates who can offer pragmatic solutions to future employers remains unclear.

There is some agreement between leading researchers in the field that extracting meaning from big data using data science and machine learning applications requires skills in at least three key areas: (1) computing and software engineering, (2) machine learning, statistics and optimization and (3) product understanding and experimentation. However, as we have shown in this analysis, it is very difficult to formulate a curriculum that adequately and coherently addresses all three areas simultaneously, especially when choosing a common set of tools, software and platforms upon which to conduct the analysis. Indeed, it is similarly difficult to identify professionals and academics who maintain expertise and skills in all three of the same areas. Thus, the problems with curriculum development will persist until a common set of widely held principles is established.

In the context of this analysis, we consider some discriminative terms known a priori (e.g. a representative set of terms from the texts), and find that there is less overlap between diverse fields than is required to deliver coherent data science and machine learning programs.

Quantifying curricula gaps

A natural language processing-based approach offers several advantages over more qualitative or intuitive-based strategies to identify curriculum gaps. First is the flexibility of the approach to meet the individual institutional needs. Once documents have been processed, different approaches and query strategies to identify a specific outcome can be implemented at a relatively low programming effort using standard database query applications. Second, as opposed to a more qualitative approach, search strategies can be monitored on a prospective basis, which can potentially identify gaps in the curriculum while students are still completing their program. This could greatly facilitate near real-time quality assurance processes. A natural language processing-based search strategy is far more scalable than manual abstraction, potentially allowing surveillance over the entire curriculum population rather than on smaller samples.

This approach has been adopted to identify curricula gaps in data science courses at several US and Australian institutions. A basic web-based keyword analysis tool using the above approach has been made available at www.pluridisciplinary.com.au for course developers to use to help quantify the contribution of pre-defined categories to assist with curriculum validation.

Conclusion

Emerging careers in technology-focused interdisciplinary fields such as data science coupled with necessary graduate outcomes mandate the need for a truly cross-disciplinary pedagogical approach. This study introduced a text mining approach as an initial step for the development of an integrated curriculum for students engaged in interdisciplinary programs. We introduced a framework for the text mining of curricula for purposes of subject development as well as various functions for managing curriculum design through the abstraction of document manipulation and use of heterogeneous text formats. Through the integration of a database, we are able to minimize memory demands, as well as cater for advanced metadata management for collections of text documents.

This approach can assist educators design, implement and adapt the breadth and depth of a curriculum to the required level unconstrained by single discipline bias. It also serves as a validation tool for students to distinguish between courses and subjects across academic institutions, as well as comparing university-housed courses against online-delivery courses in this field.

There are still areas open for further improvement such as the use of more common methods in linguistics like latent semantic analysis to provide a better scan of subject materials and patterns in text data across subjects.

Acknowledgments

The author would like to acknowledge the support of the Australian Office of Learning and Teaching through OLT grant FS15-0252 and two anonymous reviewers for insightful feedback.

Disclosure statement

No potential conflict of interest was reported by the author.

Funding

Australian Office of Learning and Teaching through OLT [grant number FS15-0252].

Notes on contributor

Jason West has over 25 years of both academic and industry experience. He received the PhD degree in quantitative finance from the University of Technology, Sydney. He has published two books and over 40 journal articles on the use of quantitative techniques for solving complex financial, energy and climate change adaptation issues.

References

Blei, D.M., Ng, A.Y., & Jordan, M.I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research, 3*, 993–1022.

Coulet, A., Shah, N.H., Garten, Y., Musen, M., & Altman, R.B. (2010). Using text to build semantic networks for pharmacogenomics. *Journal of Biomedical Informatics, 43*(6), 1009–1019.

Davi, A., Haughton, D., Nasr, N., Shah, G., Skaletsky, M., & Spack, R. (2005). A review of two text-mining packages: SAS textmining and wordstat. *The American Statistician, 59*(1), 89–103.

Feldman, R., & Sanger, J. (2007). *The text mining handbook: Advanced approach in analyzing unstructured data*. Cambridge: Cambridge University Press.

Jojic, N., Perina, A., & Murino, V. (2010). Structural epitome: A way to summarize one's visual experience. *Advances in Neural Information Processing Systems, 23*, 1027–1035..

Kao, A., & Poteet, S.R. (2007). *Natural language processing and text mining*. London: Springer.

Kaufman, L., & Rousseeuw, P.J. 1990. *Finding groups in data*. New York, NY: Wiley.

Kuo, C.-J., Ling, M.H.T., Lin, K.-T., & Hsu, C.-N. (2009). BIOADI: A machine learning approach to identifying abbreviations and definitions in biological literature. *BMC Bioinformatics, 10*(Suppl 15), S7.

Lee, S., Lee, S., Seol, H., & Park, Y. (2008). Using patent information for designing new product and technology: Keyword based technology roadmapping. *R&D Management, 38*, 169–188.

Luther, S., Berndt, D., Finch, D., Richardson, M., Hickling, E., & Hickam, D. (2011). Using statistical text mining to supplement the development of an ontology. *Journal of Biomedical Informatics, 44*, S86–S93.

Perina, A., Jojic, N., Bicego, M., & Truski, A. (2013). Documents as multiple overlapping windows into grids of counts. In C. Burges (Ed.), *Proceedings of the 2013 Conference on Neural Information Processing Systems* (pp. 10–18). Tahoe, NV.

Pison, G., Struyf, A., & Rousseeuw, P.J. (1999). Displaying a clustering with CLUSPLOT. *Computational Statistics & Data Analysis, 30*(1999), 381–392.

Porter, M.F. (1980). An algorithm for suffix stripping. *Program, 14*(3), 130–137.

Rencher, A.C. (2002). *Methods in multivariate analysis* (2nd ed.). Hoboken, NJ: John Wiley & Sons.

Robertson, S.E., & Walker, S. (1994). Some simple effective approximations to the 2-Poisson model for probabilistic weighted retrieval. In W. Bruce Croft & C. J. van Rijsbergen (Eds.), *Proceedings ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 232–241). New York, NY: Springer-Verlag.

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing and Management, 24*(5), 513–523.

Shawe-Taylor, J., & Cristianini, N. (2004). *Kernel methods for pattern analysis*. Cambridge: Cambridge University Press.

Shulman, L. (1986). Those who understand: Knowledge growth in teaching. *Educational Researcher, 15*(2), 4–14.

Stankov, S., Rosic, M., Zitko, B., & Grubisic, A. (2008). TEx-Sys model for building intelligent tutoring systems. *Computers & Education, 51*, 1017–1036.

Strehl, A., Ghosh, J., & Mooney, R. (2000). *Impact of similarity measures on web-page clustering*. Austin, TX: Workshop on Artificial Intelligence for Web Search (AAAI 2000).

Wan, X. (2007). A novel document similarity measure based on earth movers distance. *Information Sciences, 177*(18), 3718–3730.