Kansas Library Association College and University Libraries Section Proceedings

Volume 14
Number 1 *Resilient Practices in an Age of Uncertainty*

Article 8

2024

# Data-driven collection development: Text mining college course catalogs

Sarah W. Sutton
*Emporia State University School of Library and Information Management*, ssutton3@emporia.edu

Hunter Tolbert
htolbert@g.emporia.edu

Khiana Harris
kharri28@g.emporia.edu

## Recommended Citation

Sutton, Sarah W.; Tolbert, Hunter; and Harris, Khiana (2024) "Data-driven collection development: Text mining college course catalogs," *Kansas Library Association College and University Libraries Section Proceedings*: Vol. 14: No. 1. https://doi.org/10.4148/2160-942X.1093

# Data-driven collection development: Text mining college course catalogs

## Abstract

Academic librarians are trained to develop and manage collections. They rely on their own subject expertise in academic disciplines, input from teaching faculty, and professional training to make informed selections to support institutional curriculum. Professional training in collection development has, in recent years, focused on evidence-based acquisition methods (Johnson, 2018, p. 134). College and university course catalogs are a potential but untapped source of evidence for identifying topics of importance to institutional curricula. Course descriptions are concise descriptions of the subjects covered in college or university courses and therefore the topics about which students may require additional sources of information. Until recently, examining course catalogs was a time-consuming prospect. The advent of data and text mining techniques, however, makes it possible to analyze course descriptions with much less time and effort expended. This article contains a brief introduction to data science in libraries; details of tools and processes used for collecting and cleaning course catalog data; and preliminary results of a project to mine course catalogs for changes in curriculum focus to benefit library collection development decisions.

## Keywords

Python, R, collection development, text mining, data science, academic libraries

## Cover Page Footnote

The authors would like to acknowledge and express gratitude to the reviewers of this article for their insight and recommendations.

## INTRODUCTION

Academic librarians are trained to develop "collections sufficient in quality, depth, diversity, format, and currency to support the research and teaching missions of the institution" (Association of College and Research Libraries, 2018). For making informed selections to support institutional curriculum, academic librarians may rely on their own subject expertise in academic disciplines, input from teaching faculty, and professional training in collection development. Professional training in collection development historically focused on evidence-based acquisition methods including patron-driven acquisitions in which materials "selection decisions are driven by library patrons' actions" (Johnson, 2018, p. 134).

Another source of evidence for identifying topics of importance to institutional curriculum are course descriptions in the institutional course catalog. Course descriptions are concise descriptions of the subjects covered in college or university courses and therefore the topics about which students may require additional sources of information. Until recently, examining course catalogs was a time-consuming prospect, often requiring more time and effort by librarians than was available. The advent of data and text mining techniques, however, makes it possible to analyze course descriptions with much less time and effort expended.

The purpose of the project this paper reports on was to mine course description data from college course catalogs to identify topics relevant to institutional curriculum and changes in those topics over time to inform college and university library collection development plans, budgeting, and support service decisions. The paper begins with a brief introduction to data science in libraries followed by details of tools and processes used for collecting and cleaning course catalog data and results of the project.

## WHAT IS TEXT MINING?

Text mining is a collection of techniques that fit within the larger domain of data science. Data science is an academic discipline that extracts knowledge from data in various fields including librarianship (Lin & Scott, 2023). "Text mining, also known as text data mining, is the process of transforming unstructured text into a structured format to identify meaningful patterns and new insights. You can use text mining to analyze vast collections of textual materials to capture key concepts, trends and hidden relationships" (IBM, 2024).

In data analysis of almost any kind, the purpose is to derive meaning that might not be clear at first glance, perhaps because the data set is very large or because the data are widely scattered or because they have not been organized (structured) in a way that makes patterns in the data visible. Textual data analysis consists of breaking chunks of text down to individual words or combinations of words in order to discover patterns and meaning. The benefit of using text mining techniques is that current computing systems are capable of performing data analysis on very, very large sets of data, so large that it would be impossible for a human to process because computers process data orders of magnitude faster than humans.

In addition to the amount of data contained in course catalogs, that they are often presented in .pdf format, can also present a barrier to their analysis. Converting a .pdf file to a "minable" text format has long been considered too difficult and time consuming a process to be a worthwhile source of information to inform collection development. Increased computational capacity and data processing techniques have simplified the process. This paper describes the application of several text mining techniques to make possible, in a timely manner, the analysis of thousands of pages of course descriptions.

## TEXT MINING IN THE LIS LITERATURE

While data science has long been a topic of interest, text mining has been applied only rarely in library and information science (LIS) literature. For example, using data to inform library practices include the analysis of reference question data to enhance library outreach and create targeted guides (Finnell & Fontane, 2010). Another group employed the WEKA software suite to implement the Apriori algorithm, identifying borrowing patterns from self-service library stations in Taipei (Tu et al., 2021). Most closely related to the present project was the use of transaction log analysis at the University of Punjab to assess e-book popularity, informing future purchasing decisions (Rafique et al., 2023). To the best of the authors' knowledge, however, our project along with Been, Thompson, and Weber's (2023, 2024), is one of the first to apply text mining techniques to library collection development.

## DATA COLLECTION AND PREPARATION IN THE CURRENT PROJECT

For the current project, raw data in the form of .pdf course catalogs were gathered from publicly available web pages of a medium sized regional university in the Midwest United States. The .pdfs were converted into .docx files and extraneous text removed leaving only course descriptions. Course description data in text files were imported into RStudio and Google Colab for further cleansing. Data cleansing is the process of cleaning data in each variable to streamline analysis. It includes breaking sentences into individual words (called tokenizing), transforming text into lowercase, removing some or all punctuation, and removing stopwords. Stopwords are words in any language that are deemed to not add meaning to the text being analyzed.

Data cleansing resulted in variables describing the academic department code, course number, course title, course description, and year. These were captured in data frames and stored in .csv format. A data frame is a data structure in which, similar to a spreadsheet, data are arranged in rows that represent variables and columns that represent cases. Final preparation of the full data set was accomplished using regular expressions, delimiters, and manual editing of .csv files in Excel.

## Data Analysis Techniques

Data analysis was performed using Python and R. Data analysis steps varied slightly by language and results compared. Detailed steps are described in the R markup and Python notebooks in the Github repository associated with this article (https://github.com/sarahwsutton/Data-driven-collection-development-Text-mining-college-course-catalogs). Since this project was exploratory, we applied several text mining techniques to data analysis: n-grams, word clouds, and topic modeling. Each author selected one or more academic disciplines whose course descriptions they would examine across 10 years of catalogs.

N-grams are sequences of *n* adjacent words within a corpus of texts ("*n*-gram", 2024). They are sometimes referred to using the number of words used in combination, for example "bi-grams" is a combination of 2 words and a "tri-gram" is a combination of 3 words. Google n-grams (https://books.google.com/ngrams/info) is a well-known example of the use of n-grams to identify trends in word use over time. For this project, tri-grams were determined to be the most useful unit of analysis for identifying course topics.

Word clouds are used as a preliminary step in topic modeling to determine the effects of data preprocessing (Kapadia, 2022), but we also found them to be useful for identifying themes in course descriptions. Topic modeling is "a type of statistical language [modeling] used for uncovering hidden structure in a collection of texts" (Kapadia, 2019, sec. "Introduction").

## Results

In this section results are reported for each individual technique and discipline and then comparisons across techniques are made.

### N-gram Frequency in Python

The process of identifying the frequently occurring tri-grams in course descriptions for English began with isolating English course descriptions into a single data frame. Course descriptions were cleansed by removing standard stop words and punctuation and then transformed to lowercase. Using the Natural Language Toolkit (NLTK), a suite of programs and libraries in Python that help to process natural language as used by humans into code readable by computers, a list of trigrams was developed from the data. After the initial list of trigrams was created it became clear that there were additional words that were not contributing to overall meaning. An additional list of stop words customized to the English curriculum being examined was created, the words removed from the data, and a final list of trigrams developed. The most frequent are included in Table 1.

**Table 1.**
*Frequent Tri-grams Identified in English Course Descriptions*

| Tri-gram | Frequency |
|---|---|
| Creative, writing, literary | 24 |
| Superior, college, preparation | 22 |
| Sonnets, epic, poems | 22 |
| Young, adult, literature | 22 |
| Author, author, studied | 22 |

## TRI-GRAM FREQUENCY IN R

Similar to Python, using several R packages in conjunction (readr, tidytext, dplyr, and tidyr) allowed for the creation of data frames containing the tri-grams and removal of standard and custom stop words within Sociology, Psychology, and Anthropology course descriptions. Unlike Python, these R packages include the step of ignoring punctuation and casing so it does not have to be included as a separate step. Once the tri-grams were tabulated and stored in a data frame, they were exported and visualized as word clouds. Table 2 contains the most frequent trigrams and Figure 1 depicts the word cloud for Sociology course descriptions. The concepts that appear most often in sociology course descriptions over time are: examine societal perspectives, law enforcement courts, and explore one’s ethics.

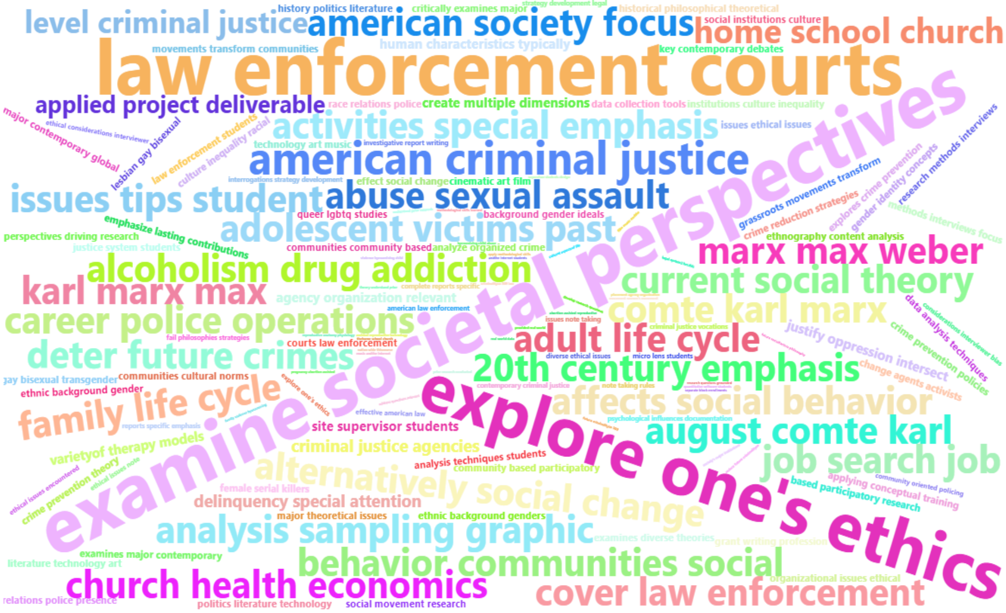
**Table 2.**

*Frequent Tri-grams Identified in Sociology Course Descriptions*

| Tri-gram | Frequency |
|---|---|
| Examine, societal, perspectives | 22 |
| Law, enforcement, courts | 22 |
| Explore, one’s, ethics | 20 |
| American, criminal, justice | 13 |

**Figure 1.**

*Word Cloud of Tri-grams from Sociology Course Descriptions*

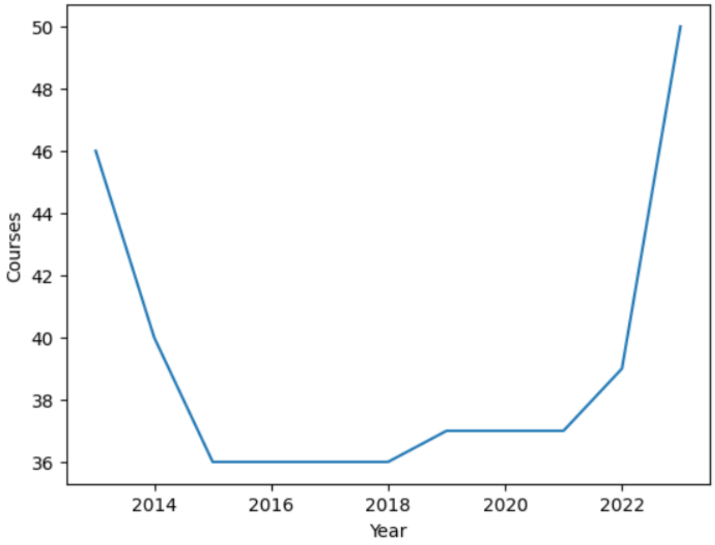


## Word Clouds and Topic Modeling in Python

Topic modeling combines the frequency of a word’s appearance in a corpus of text with the number of documents in which it appears within the corpus to predict “topics,” or combinations of words. Topic modeling was applied to Computer Science (CS) course descriptions using Python. CS course descriptions were combined in a single data frame and some general descriptive information generated including the number CS course descriptions included in each year’s catalog. This provided clues about changes from year to year. In CS, the number of courses dropped in 2015, remained stable until 2019, rose slowly from 2019 to

2022, and then increased sharply in 2023 as is illustrated in Figure 2, suggesting major curriculum shifts in 2015 and 2023.

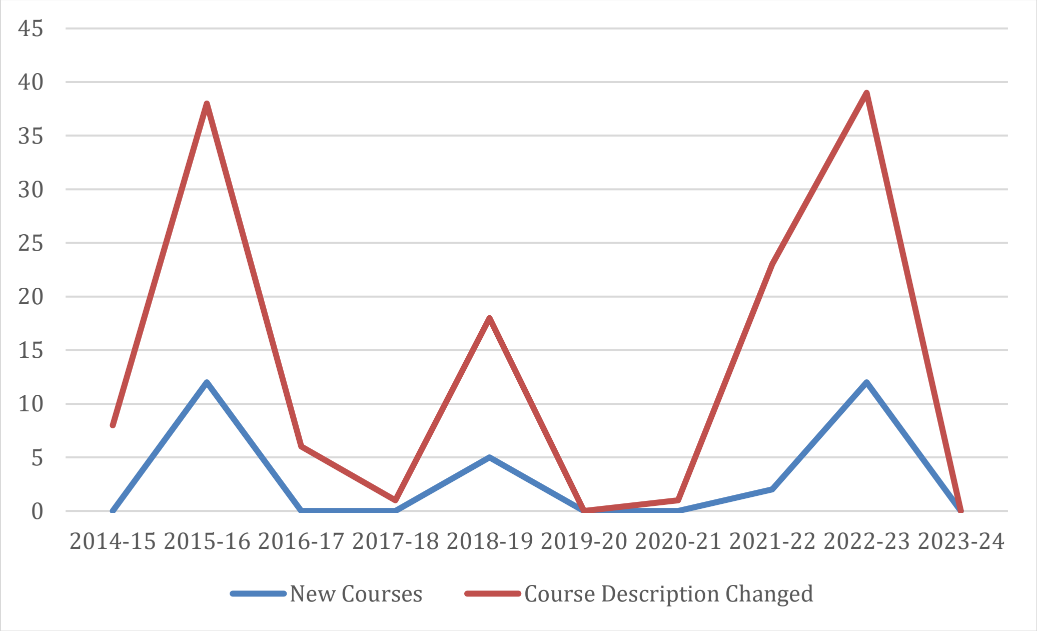
**Figure 2.**
*Number of Unique Computer Science Courses per Year.*



Next, courses that did not contain topical titles or descriptions, such as directed readings, practicums, and special topics courses were removed. Course descriptions were broken apart (i.e., tokenized) by sentence and sentences containing information about credit hours and prerequisite requirements removed. Course descriptions were grouped by year so that comparisons could be made from one year to the next. This would allow collection development librarians to identify specific topical changes at the times of major shifts in 2015 and 2023.

Figure 3 illustrates years in which new courses were added and existing course descriptions were changed. It supports the result that major curriculum changes occurred in 2015 and 2023. It also suggests that addition of new courses and changes in existing course descriptions tend to happen in the same year.

**Figure 3.**
*New Courses and Changes to Course Descriptions by Year for CS*

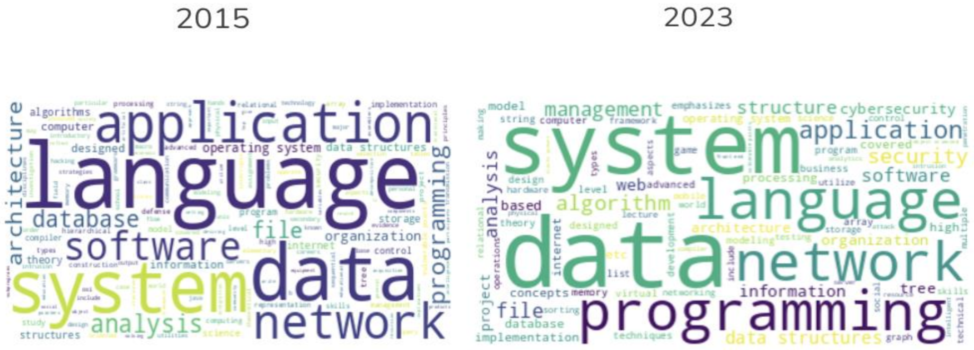


Since it was clear that there were major course/curriculum changes in CS course descriptions in 2015 and 2023, the next part of the analysis compared course descriptions from those two academic years using word clouds and topic modeling as a potential alternative to tri-grams.

The topic model used in this analysis was latent dirichlet allocation (LDA), one of the most frequently used topic modeling techniques. In topic modeling, word clouds are used to verify that the data cleansing is sufficient. It is important to know whether additional data cleansing is necessary before training the topic model (Kapadia, 2022). Figure 4 contains the word clouds from the cleaned course descriptions for 2015 and 2023. Comparing them suggests changes in topical focus in the CS curriculum that the library may use to adjust its collections budget allocations.

**Figure 4.**

*Comparison of Word Clouds from CS Course Descriptions in 2015 and 2023*



The topic models created from CS course descriptions from 2015 and 2023 illustrate similar but more specific results. Table 3 compares sample 2015 course catalog topics to sample 2023 course catalog topics.

**Table 3.**

*Comparison of Sample Topics in the CS Curriculum in 2015 and 2023.*

| 2015 Topics | 2023 Topics |
|---|---|
| Creation of models and software applications; | Threat detection; |
| Programming, program debugging, gaming; | Managing network security; |
| System design and network analysis; | Data security software; |
| Theory, programming structures, ap architecture. | Programming languages, project oriented data structures. |

## Cross-technique Observations

Comparisons of results across researchers and academic disciplines resulted in additional observations. In anthropology course descriptions, while it was possible to extract tri-grams and create a word cloud, there were not any significant changes over time or major contrast of frequency within the tri-grams.  In contrast, in CS course descriptions there were two periods of increased change. For librarians with responsibilities for collection development and management this knowledge could inform collections budget allocation decisions.

In all of the disciplines in which course descriptions were examined and across all the techniques used in their analysis, it is clear that the results of text mining techniques require human interpretation and, ideally, subject expertise, also known as domain knowledge. Librarian liaisons often have (or develop) some subject expertise in their liaison areas and therefore should be included in this kind of data analysis and interpretation.

## CONCLUSION

In this analysis of the textual course descriptions using text mining techniques, we identified some academic disciplines within the institution where the curriculum changed little over the 10 years of the study and others where the curriculum changed in clearly identifiable periods. Such information would support library collections decision making, for example, creating a more fair distribution of collections funds. Visualizations resulting from the disciplinary analyses helped us to further identify topical changes in curricular focus, which would support materials selection. Both results suggest that the use of data science techniques in LIS may be useful to academic librarians for collection development and management.

The use of data science techniques in libraries is subject to several caveats, one of the most important being that there is no substitute for human domain knowledge, particularly in text mining. Data and results from the techniques we applied, n-grams, word-clouds, and topic modeling, make clear that results cannot be interpreted without disciplinary and institutional context. In collection development, this suggests the inclusion of disciplinary faculty in the data analysis process when librarian liaisons need additional domain expertise.

Were we to continue this project, we would explore more academic disciplines, continue to refine disciplinary stop words lists, and seek additional sources of data as has been suggested by Been, et al. (2023 and 2024), such as institutional course change request forms. In hopes that others will take an interest in our work, we have shared R markdown and Python notebooks containing the code for the analyses conducted in a publicly available Github repository (https://github.com/sarahwsutton/Data-driven-collection-development-Text-mining-college-course-catalogs). The notebooks contain additional details of our analyses and instructions for using our code with data from other institutions. This would be important to creating and refining topic models since one limitation to our use of topic modeling is the relatively small amount of data used to create them.

## REFERENCES

Association of College and Research Libraries. (2018). Standards for libraries in higher education. https://www.ala.org/acrl/standards/standardslibraries

Been, J., Thompson, M., & Weber. (2023, March 6). Of course we want to see into the future: Data mining course catalogs. Electronic Resources in Libraries, Austin, TX.

Been, J., Thompson, M., & Weber, M. (2024, March 4). Data nebulae: Shaping library collections for the future. Electronic Resources & Libraries, Austin, TX.

Johnson, P. (2018). Fundamentals of collection development and management (4th ed.). American Library Association.

Finnell, J., & Fontane, W. (2010). Reference Question Data Mining: A Systematic Approach to Library Outreach. *Reference & User Services Quarterly*, *49*(3), 278–286. http://www.jstor.org/stable/20865263

Lin, S., & Scott, D. (2023). Hands on data-science for librarians. Chapman and Hall/CRC.

IBM. (2024, April 9). What is text mining? https://www.ibm.com/topics/text-mining

Kapadia, S. (2022). Topic modeling in Python: Latent Dirichlet Allocation (LDA). Retrieved 5/13/24 from https://towardsdatascience.com/end-to-end-topic-modeling-in-python-latent-dirichlet-allocation-lda-35ce4ed6b3e0

Mathworks. (2024). What Is an N-Gram? https://www.mathworks.com/discovery/ngram.html

Rafique, A., Ameen, K., & Arshad, A. (2023). E-book data mining: Real information behavior of university academic community. *Library Hi Tech*, *41*(2), 413–431. https://doi.org/10.1108/LHT-07-2020-0176

Silge, J., & Robinson, D. (2017). Text mining with R: A tidy approach. O'Reilly Media. https://www.tidytextmining.com/#welcome-to-text-mining-with-r

Tu, Y.-F., Chang, S.-C., & Hwang, G.-J. (2021). Analysing reader behaviours in self-service library stations using a bibliomining approach. *The Electronic Library*, *39*(1), 1–16. https://doi.org/10.1108/EL-01-2020-0004