EVALUATING TRENDS IN DATA SCIENCE CURRICULUM THROUGH TOPIC MODELING AND TEXT MINING

By

GABRIEL EMILIO MANTINI

BACHELOR OF SCIENCE, DATA SCIENCE

Spring 2022

A Project submitted in partial fulfillment of the requirements for the degree of

Master of Science in Computer Science, Data Science Track

Florida Polytechnic University

Lakeland, Florida

EVALUATING TRENDS IN DATA SCIENCE CURRICULUM THROUGH TOPIC MODELING AND TEXT MINING

By

GABRIEL EMILIO MANTINI

BACHELOR OF SCIENCE, DATA SCIENCE

Spring 2022

A Project submitted in partial fulfillment of the requirements for the degree of

Master of Science in Computer Science, Data Science Track

Florida Polytechnic University

Lakeland, Florida

|  |  |  |
| --- | --- | --- |
| Approved by | Signatures | Date |
| Dr. Shahram Taj  Committee Chair |  | |
| Dr. Reinaldo Sanchez-Arias  Committee Member |  | |
| Dr. Lizhen Shi  Committee Member |  | |
| Dr. Tom Dvorske  Vice Provost of Academic  Affairs – Graduate Division |  | |

Contents

[Abstract 4](#_Toc101871792)

[Introduction 4](#_Toc101871793)

[Methodology 5](#_Toc101871794)

[Results 9](#_Toc101871795)

[All Course Lectures/Notes 9](#_Toc101871796)

[Single Course Lecture/Notes 11](#_Toc101871797)

[Introduction to Machine Learning 11](#_Toc101871798)

[Introduction to Computational Thinking and Data Science 14](#_Toc101871799)

[Economic Applications of Game Theory 16](#_Toc101871800)

[Course Descriptions 18](#_Toc101871801)

[Florida Poly 18](#_Toc101871802)

[MIT 21](#_Toc101871804)

[RStudio and Packages 23](#_Toc101871806)

[Conclusion 24](#_Toc101871807)

[References 25](#_Toc101871808)

[Appendix 27](#_Toc101871809)

## List of Tables

[Table 1 - Plan of Study for MIT B.S. 6-14: Computer Science, Economics, and Data Science 6](#_Toc101886922)

[Table 2 – Top 5 words in 11 topics of MIT Machine Learning Lecture Notes. 13](#_Toc101886923)

[Table 3 -First 9 Topics Generated of MIT Data Science Course Descriptions 22](#_Toc101886924)

## List of Figures

[Figure 1- Plan of Study for MIT B.S. 6-14: Computer Science, Economics, and Data Science. 7](#_Toc101886904)

[Figure 2 - Plan of Study for FPU B.S. Data Science (2022-2023 Catalog) [25] 8](#_Toc101886905)

[Figure 3 – Initial Word Cloud of Combined Course Lecture Transcripts/Notes 9](#_Toc101886906)

[Figure 4 – Bar Chart for most frequent words of Combined Course Lecture Transcripts/Notes 10](#_Toc101886907)

[Figure 5 – Word Cloud of Machine Learning Lecture Notes 11](#_Toc101886908)

[Figure 6 – Model tuning analysis of Machine Learning Lecture Notes Topics 12](#_Toc101886909)

[Figure 7 - Dendrogram of MIT Introduction to Machine Learning Topics 14](#_Toc101886910)

[Figure 8 – Dendrogram of MIT Introduction to Computational Thinking and Data Science 14](#_Toc101886911)

[Figure 9 – Top terms of MIT Computational Thinking and Data Science 15](#_Toc101886912)

[Figure 10 – Dendrogram of MIT Computational Thinking and Data Science Topics 15](#_Toc101886913)

[Figure 11 – Word Cloud of frequent words in Economic Applications of Game Theory course lectures 16](#_Toc101886914)

[Figure 12 – Bar Chart for most frequent words of Economic Applications of Game Theory course 17](#_Toc101886915)

[Figure 13 – Word Cloud of frequent words in FPU Data Science Course Descriptions 18](#_Toc101886916)

[Figure 14 – Topic Model Selection of FPU Data Science Course Descriptions 19](#_Toc101886917)

[Figure 15 – Topic Modeling of FPU Data Science Course Description 19](#_Toc101886918)

[Figure 16 – Dendrogram of FPU Data Science Course Descriptions 20](#_Toc101886919)

[Figure 17 – Word Cloud of frequent words in MIT Data Science Course Descriptions 21](#_Toc101886920)

[Figure 18 - Dendrogram of MIT Data Science Course Description 23](#_Toc101886921)

# Abstract

Several studies have been done on capturing student engagement, performing sentiment analysis, and predicting student performance through the use of online education data, but little has been done in the way of comparative analysis of multiple curriculums data. Here we report the difference in Florida Polytechnic University’s B.S. Data Science with Massachusetts Institute of Technology’s equivalent degree through the use of topic modeling and data visualization of text mining processes. Latent Dirichlet allocation is performed to generalize each course into a blend of a number of topics that capture each word’s presence from course descriptions, lecture videos, and lecture transcripts. Moreover, model tuning is performed to identify the best number of topics selected by the algorithm to bucket words in to and give the highest amount of coherence within topics. Word clouds, bar plots, and dendrograms are used to further visualize these comparisons. Different levels of granularity are explored: both curriculum to curriculum analysis, and single course to course.

# Introduction

Open, online course material has seen a steady rise in popularity as technology has become increasingly accessible. A pioneer of the open higher education movement and community is Massachusetts Institute of Technology (MIT), who initially launched MIT OpenCourseWare in April 2001[1]. MIT OpenCourseWare includes courses across an array of disciplines taught by the school, and historically included provided homework problems, tests, lecture notes, and readings lists. Recent years have expanded on the offerings, such as in 2018 when complete audio/video lectures began to be available from those that had occurred in a traditional classroom setting from prior semesters [2].

With the rise of data science in industry, there has been a natural response of growth of data science programs and courses in traditional university offerings, as well as a growth of DSBA course offerings from online alternatives. Glassdoor’s annual ranking shows ‘Data Scientist’ as 3rd in Best Jobs in America for 2020, which would support that continued growth [16]. On a conceptual and learning level, questions are open for how data science curriculum should be structured, and which topics should be prioritized. MIT OpenCourseWare showcase the development of online learning with potential for larger learning communities and peer tutoring, through sheer volume that isn’t otherwise possible [2].

Alongside this, it is of interest to analyze the emerging trends in OpenCourseWare course offerings and topics for the Data Science field. The growing market and demand for data scientists leads to desiring skill sets and knowledge from courses [16], and it is in the interest of a fledgling university in Florida Poly to see how an industry leader in MIT is prioritizing their curriculum. This begs the question: How similar are Florida Poly’s Data Science courses and topics to those that the MIT offer through its equivalent degree available in OpenCourseWare?

# Methodology

The selection of MIT OpenCourseWare as the primary target for analysis came after a need to effectively nuance the project and provide a novel report that is reasonably scalable. While MOOCs are compelling in their size and growth in popularity [3], the barrier to entry proved slightly higher for data collection purposes. Despite many effective studies and a national conference focused for them [4], the scope was narrowed to MIT’s program due to an intrigue with their interdisciplinary adaptation of the undergraduate data science degree [21], and the prestige of the university which served as a model benchmark for material.

In selecting MIT, an MOOC-oriented route could pivot to MITx, which are certified MOOCs delivered through the edX platform or through MITx online [38]. It was discovered that courses naturally had to be enrolled in to for viewing material, with access to course materials eventually expiring upon completion and the time frame to take this course not always being immediately. As a result, analyzing MIT OpenCourseWare material which effectively serve as a snapshot of previously taught lectured proved easier and more repeatable for obtaining lecture notes and transcripts in a timely fashion.

After the selection of MIT OpenCourseWare as the focal point of the study, the choosing of Data Science and Business Analytics was a natural next step, to condense material to our area of expertise and have domain knowledge. Further consideration narrowed this down more to Data Science, with seeing that this was a far more commonly labelled program and to additionally narrow the scope. Exploring MIT’s current degreed revealed a Data Science equivalent B.S. degree with 6-14: Computer Science, Economics, and Data Science [21], which sparked interest in doing a like for like comparison with our institution, Florida Poly, and their equivalent B.S. degree, Data Science [25]. The format of the material available consisting of course descriptions, lecture notes, and lecture transcripts, led to text mining and topic modeling being chosen as the primary mode of analysis in this study.

Initial interest to the topic was found through a dataset comprised of word embeddings and document topic distribution vectors generated from transcripts of 12032 video lectures from 200 courses that were collected from the Coursera learning platform in a 2019 study from Kastrati et. al. [14], “Embeddings and topic vectors for MOOC lectures dataset”. This dataset provided inspiration for the analysis of MOOCs, as well as giving the idea to focus on a specific area of study, with the courses being labeled with fine-grained and general level categories. Our research pivoted away from this data as it proved challenging to map the word embeddings with the related topic vectors on a category level (for focusing on just the data science courses).

Data collection began once the focus had been narrowed to analyzing the MIT OpenCourseWare material from the MIT equivalent Data Science degree courses. The catalog includes required courses which are core classes in the curriculum, in addition to a wide array of courses where you would pick one from several options. To simplify, a mock degree plan was devised that focused on including as many of the required courses as OpenCourseWare material existed for and selecting courses in the electives/” pick one of” categories that actually existed and had lecture transcripts/notes in MIT OpenCourseWare.

The catalog was taken from the up-to-date catalog.mit.edu for the 6-14: Computer Science, Economics, and Data Science degree [21]. Course descriptions were gathered from the courses here, and the course or its equivalent was then searched for in the MIT OpenCourseWare site by Name and Course Code [22].  For those that had a suitable equivalent in the site, lecture transcripts were downloaded if available, and lecture notes were analyzed otherwise. The table of MIT OpenCourseWare course material that was selected, along with their Florida Poly equivalent, is seen here:

Table 1 - Plan of Study for MIT B.S. 6-14: Computer Science, Economics, and Data Science

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| MIT Course | MIT Course Code | Lecture Transcripts | Lecture Notes | Florida Poly Equivalent | Florida Poly Course Code |
| Economic Applications of Game Theory | 14.12 | N | Y | Game Theory and Strategic Decisions | ECO4400 |
| Intermediate Macroeconomics | 14.05 | N | Y | Principles of Macroeconomics | ECO2013 |
| Introduction to Machine Learning | 6.036 | N | Y | Machine Learning | CAP4612 |
| Optimization Methods in Business Analytics | 15.053 | N | Y | Optimization Theory | MAP4202 |
| Design and Analysis of Algorithms | 6.046 | Y | N | Algorithm Design & Analysis | COP4531 |
| Introduction to Algorithms | 6.006 | Y | N | Data Structures & Algorithms | COP3530 |
| Introduction to Computational Thinking and Data Science | 6.0002 | Y | N | Introduction to Data Science | COP2073 |
| Introduction to Computer Science in Python | 6.0001 | Y | N | Introduction to Programming Using Python | COP2034C |
| Linear Algebra | 18.06 | Y | N | Linear Algebra | MAS3105 |
| Principles of Microeconomics | 14.01 | Y | N | Principles of Microeconomics | ECO2023 |

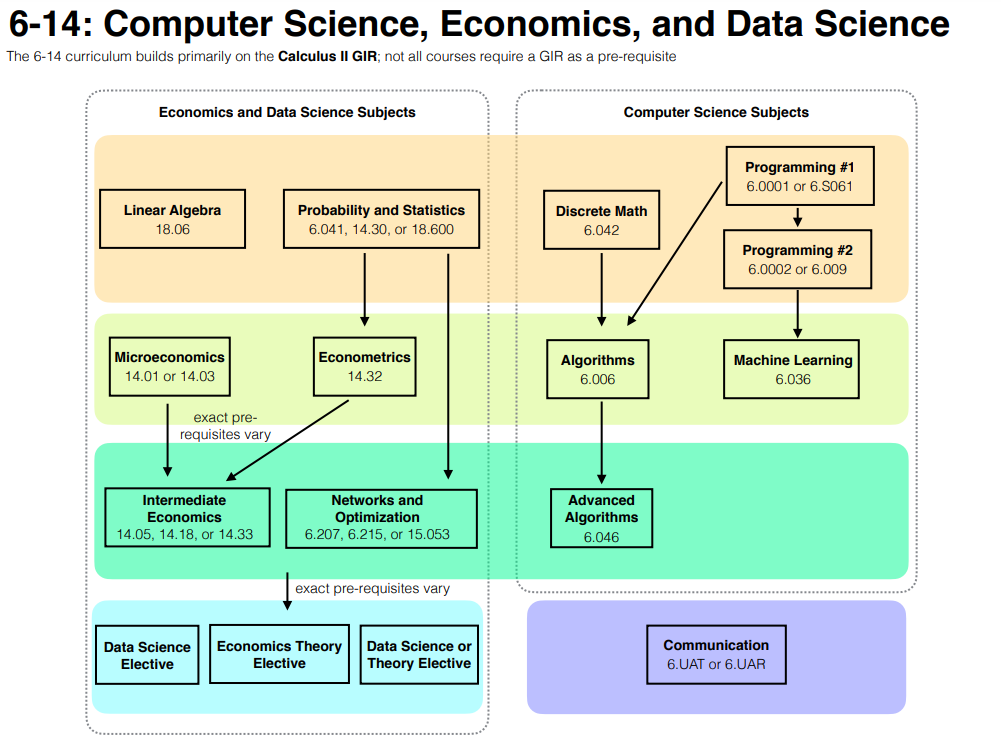


Figure 1- Plan of Study for MIT B.S. 6-14: Computer Science, Economics, and Data Science.

*(General Education Courses not included)*

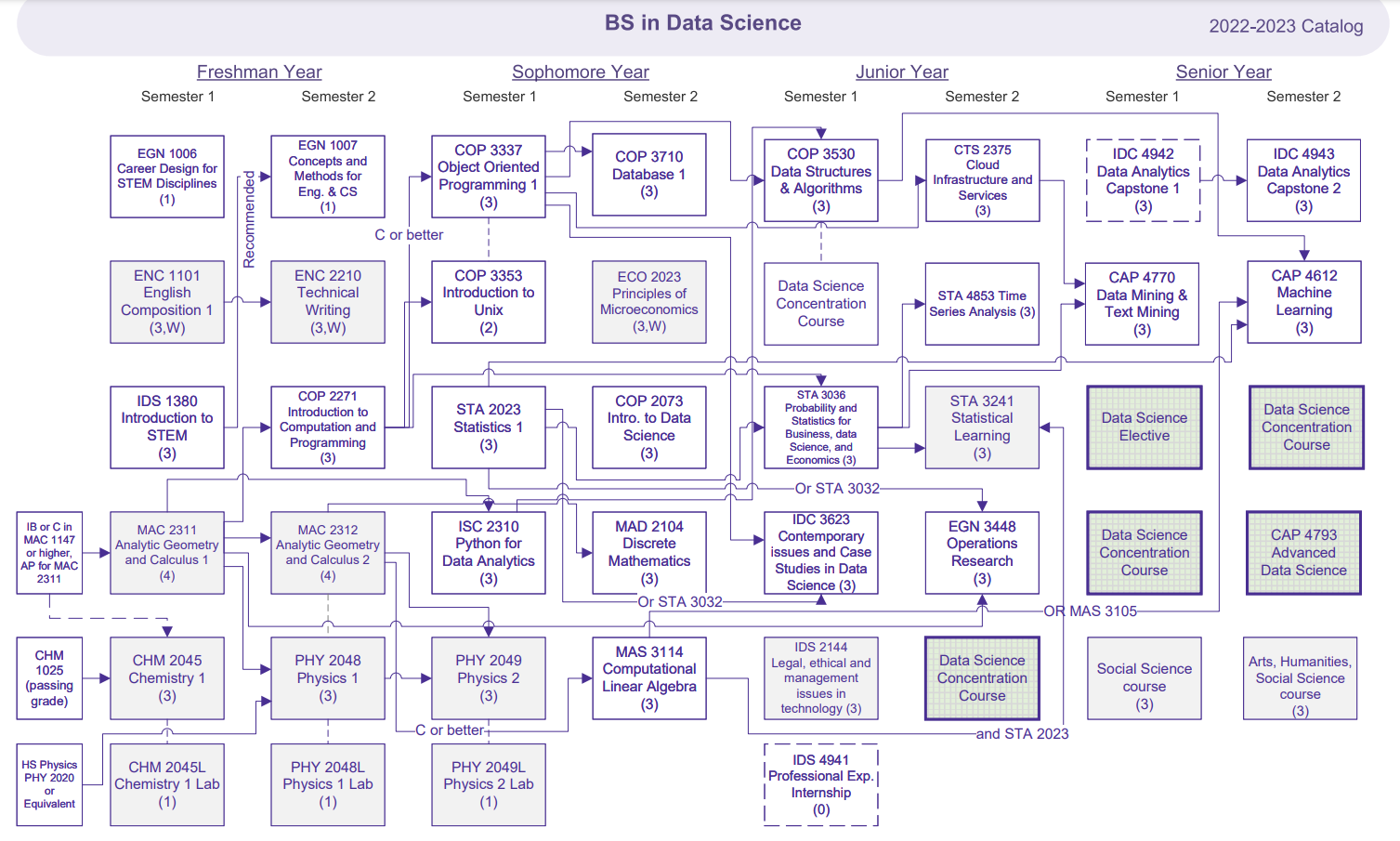


Figure 2 - Plan of Study for FPU B.S. Data Science (2022-2023 Catalog) [25]

*(General Education Courses included)*

A select few courses were excluded from the catalog for analysis in the lecture transcripts/notes. These were notably the “Project-based” courses, which were senior level courses that culminated in a research project, and oral presentation deliverable. These courses were not available through MIT OpenCourseWare.

For course description analysis, information was taken from all available courses that are possible to be taken in both the catalog for MIT (Figure 1) and FPU (Figure 2). Limiting this to just the courses that were analyzed for lecture transcripts/notes was considered but taking in all available courses was done to get a better holistic view for topic modeling. One factor that could be criticized for doing so is bias creation in considering courses and topics with many similar alternatives and being overrepresented in the topic modeling experiment. A counterargument to this is that the presence and degree of these alternatives would confirm the desire of the program to feature those courses and topics, with their prevalence being good justification for their analysis (even if all were not to be taken in one student’s enrollment).

General Education Requirements are excluded as they are very functionally similar, and MIT does not list the courses directly in the program, and the scope of the study is in predominantly the data science relevant curriculum.

# Results

## All Course Lectures/Notes

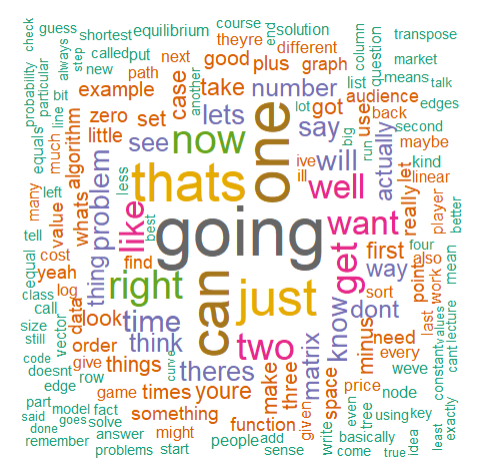
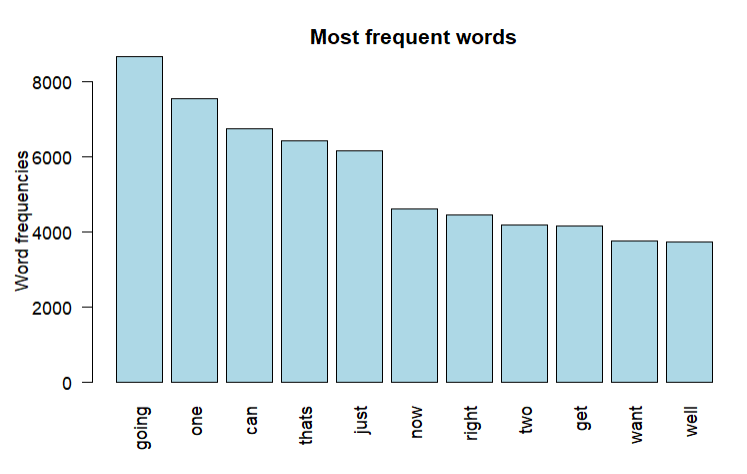


Figure 3 – Initial Word Cloud of Combined Course Lecture Transcripts/Notes

Initial effort in data visualization when trying to amalgamate all of the courses proved to generate a lot of noise, as seen above.



*Figure*

Figure 4 – Bar Chart for most frequent words of Combined Course Lecture Transcripts/Notes

Topic Modeling is a form of statistical modeling that is useful for discovering underlying themes that occur within a collection of documents [41]. Initial Topic Modeling included a lot more noise than anticipated. Although common stop words were removed, further word extraction was required to get more value out of the information. At this point, it proved necessary to narrow the scope further and reduce the load of input into the topics. For this, text mining was performed on course description/syllabus data.

## Single Course Lecture/Notes

### Introduction to Machine Learning



Figure 5 – Word Cloud of Machine Learning Lecture Notes

Word clouds are frequency representations of the number of occurrences of a word in a given set, with the larger, centered words occurring in larger frequencies, with the smaller, offset, edge words occurring at lower frequencies [24]. Word cloud of all words from lecture notes in MIT Introduction to Machine Learning courses show snippets of things we’d expect to be seeing in its content. “Gradient” and “image” are two of the largest recurring words, which are parts of the notable techniques of gradient descent and image recognition. Training, data, and set are all reoccurring and will be repeated in many of the preprocessing steps for Machine Learning.

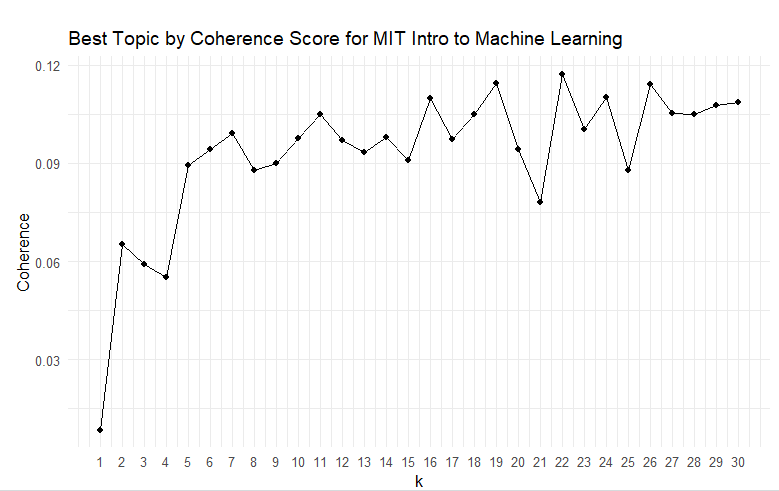


Figure 6 – Model tuning analysis of Machine Learning Lecture Notes Topics

The coherence score above from model tuning analysis for model selection serves as a guideline for the number of topics to select when running Latent Dirichlet allocation (LDA). LDA is a form of topic model that is applied to text data and works by breaking down the document matrix in to two smaller parts, a matrix consisting of topics and once consisting of the topic word [42]. The correlation of relevance between the topic words within an LDA defined topic is measured through the coherence score, to get a better idea of the number of topics is a sweet spot in having understandable correlation within topics, while not having too many topics to the point of extreme overlap [43].

Latent Dirichlet Allocation is guided by every document being a mixture of topics (i.e., Document A is 50% Topic 1 and 50% Topic 2), and every topic being a mixture of words (i.e., Engineering topic might be “Computer”, “Chemical”, “Mechanical”), with words being able to be shared between topics [39].

Model tuning using the Latent Dirichlet Allocation was conducted to identify the appropriate number of topics to manually set in topic modeling. Results yielded the highest coherence score when set to k=14 topics. Coherence scores in this case are simply a measure of how much the words in the same topic make sense when they are put together [26]. Higher quality of topics is seen with higher coherence scores, as there will be more related words together and the topic is more understandable. k=7, k=8, and k=12 are also viable by not having too much less coherence while requiring less topics to group terms.

Coherence scores in this case are simply a measure of how much the words in the same topic make sense when they are put together [26]. Higher quality of topics is seen with higher coherence scores, as there will be more related words together and the topic is more understandable. That being said, the optimal number of topics for topic selection is not purely based off of coherence score. One common strategy is to use the “elbow method”, which in this case is the point where the curve bends. The reasoning here is that at this point there are diminishing marginal returns in the coherence score while increasing the number of topics to the point where they aren’t individually clear what it is uniquely capturing [43].

The number of topics chosen to use for LDA in the Machine Learning course is k = 11, to provide a middle ground between coherence score and a more succinct number of topics

Table 2 – Top 5 words in 11 topics of MIT Machine Learning Lecture Notes.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| T\_1 | T\_2 | T\_3 | T\_4 | T\_5 | T\_6 | T\_7 | T\_8 | T\_9 | T\_10 | T\_11 |
| Learning | Descent | Algorithm | Function | Values | data | Image | Loss | Fall | Network | output |
| Policy | Gradient | Linear | Step | Vector | Model | Input | Data | Set | neural | updated |
| Action | Matrix | Training | Update | Feature | Class | Size | Function | Study | loss | time |
| Reward | Question | Classifier | Network | Set | Training | Dimensional | Learning | Question | weights | values |
| Optimal | Regression | Data | Iteration | Discredit | Regression | Values | Values | Linear | gradient | called |

Words in the table above appear in ascending order of phi-value, which is a direct correlation of a higher probability (calculated as (pr(word|topic)) of the word belonging to a topic [26]. In the figure below, a Dendrogram is visualized that uses the distance between two probability vectors (Hellinger distance) to determine if topics are closely related. The closer a branch is to another on the Dendrogram tree, the closer one topic is considered related to another. The height of the Dendrogram is an indicator of the order in which the clusters were joined, as a reflection of the distance between the topics [40].

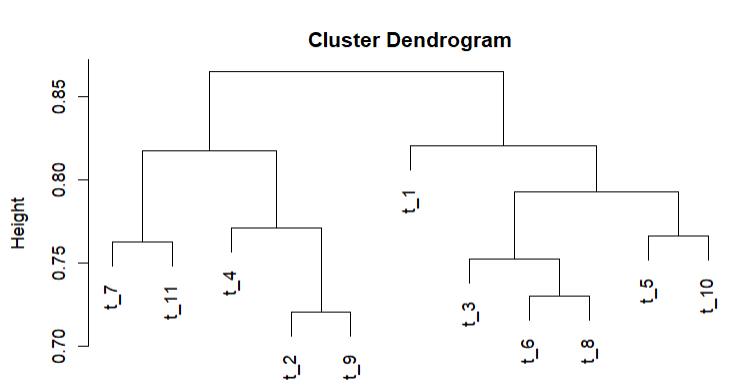


Figure 7 - Dendrogram of MIT Introduction to Machine Learning Topics

Our two figures above give us a good idea of some of our reoccurring words and their relatedness from our LDA defined topics. On the left-hand side of the dendrogram from the first split we see words in topics 7,11,4,2, and 9. Top words from here include image, input, size, output, values, matrix, gradient, descent, data, and function. On the right-hand side of the dendrogram from the first split we see policy, optimal, algorithm, training, model, learning, loss, neural, and network amongst others.

This split gives us a rough view that the topics on the left are lower level and granular and about the data, while the topics on the right are more of the analysis of the data itself and those techniques.

### Introduction to Computational Thinking and Data Science

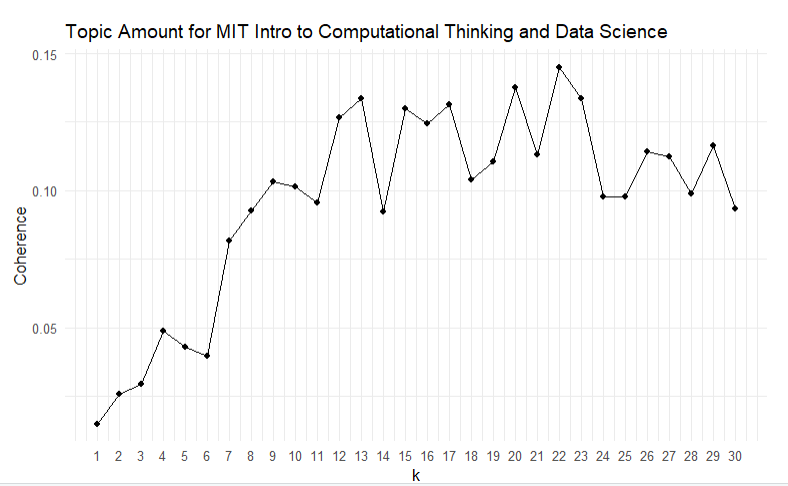
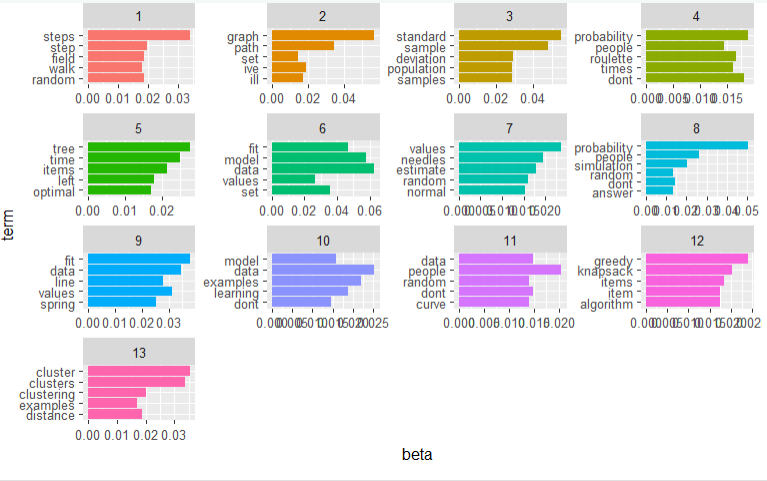


Figure 8 – Dendrogram of MIT Introduction to Computational Thinking and Data Science

Model tuning suggests little to no return after approximately k=13 topics, so LDA was performed with 13 topics.



*Figure*

Figure 9 – Top terms of MIT Computational Thinking and Data Science

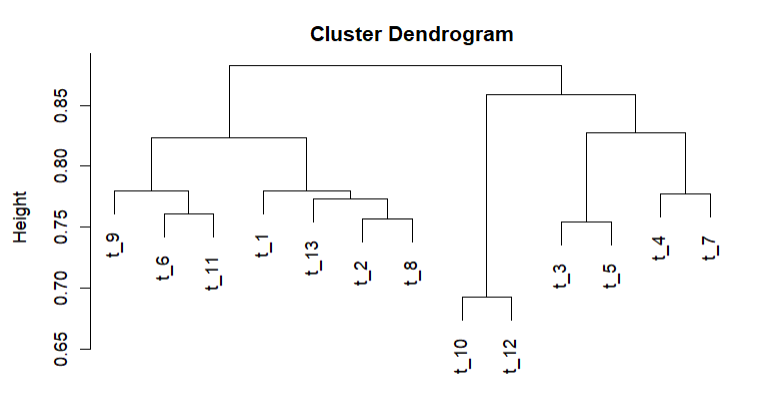


Figure 10 – Dendrogram of MIT Computational Thinking and Data Science Topics

This time, the top terms were visualized in a bar plot to give a better idea of the significance of the words relative to each other in each group. Some words that stick out significantly within their respective groups include, “steps”, “graph”, “standard”, and “probability”. Probability additionally occurs in two of the first 8 topics, showing statistics as a prevalent topic in this course.

The right side of the cluster dendrogram implies bigrams (two-word segments) of standard deviation, model learning, greedy algorithm, and knapsack items. The left side of the dendrogram implies bigrams of clustering examples, probability simulation, model fit, and data sets. This shows the course being broken down into some optimization topics such as the knapsack problem and greedy algorithms on one side, with the other focuses including simulation problems and other unsupervised problems.

### Economic Applications of Game Theory



Figure 11 – Word Cloud of frequent words in Economic Applications of Game Theory course lectures

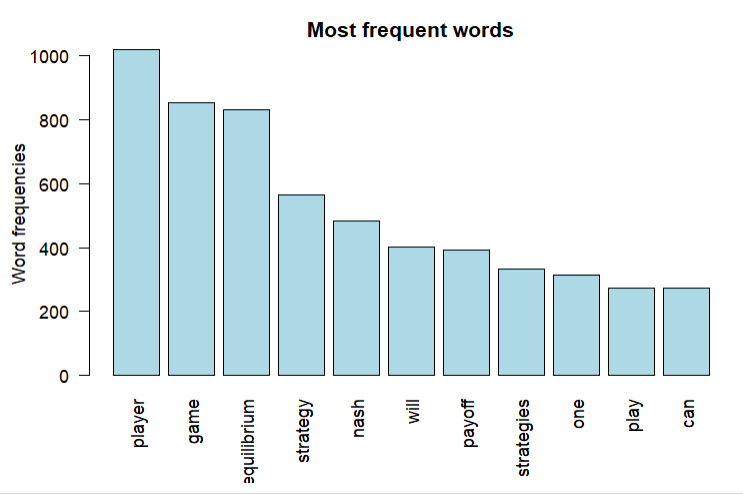


Figure 12 – Bar Chart for most frequent words of Economic Applications of Game Theory course

A quick look at a third courses lecture material before taking a more general look into course descriptions shows an Economic Applications of Game Theory course from MIT OpenCourseWare. While the prior two were required courses, Game Theory exists as an elective in the MIT 6-14 catalog [21]. This is similar to Florida Poly’s catalog for Data Science, which also has a Game Theory course as an elective and not a required course [25].

## Course Descriptions

### Florida Poly

## 

Figure 13 – Word Cloud of frequent words in FPU Data Science Course Descriptions

Course descriptions in Florida Poly Data Science courses show a concentration in data analytics and computer science principles, with the emphasis on data, programming techniques, and modeling design.

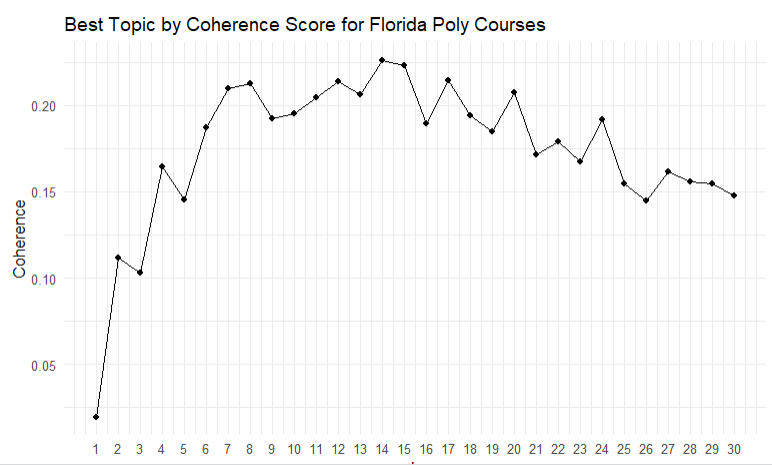


Figure 14 – Topic Model Selection of FPU Data Science Course Descriptions

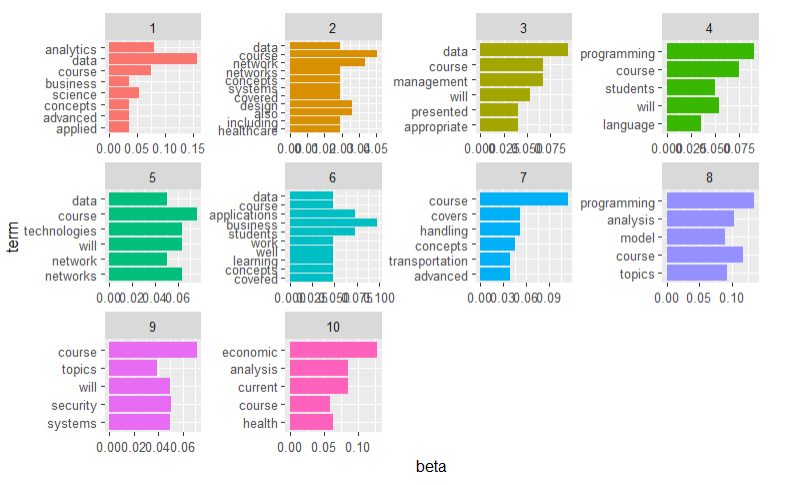


Figure 15 – Topic Modeling of FPU Data Science Course Description

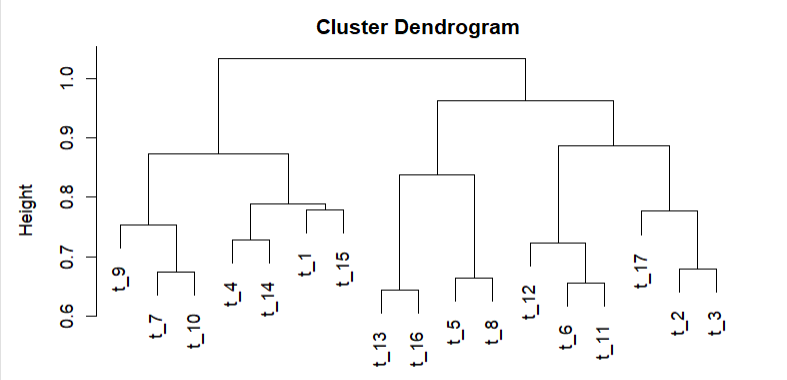


Figure 16 – Dendrogram of FPU Data Science Course Descriptions

Recurring words within the topics on the left side of the dendrogram are language, design, system, logistics, economic, models, information. Recurring words within the topics on the right side of the dendrogram are programming, software, computer, data, techniques, networks, performance, security, algorithms. This shows an emphasis on more traditional computer science topics on one branch along with the supplementation of data analytics and business/economic topics on the other.

### MIT

## 

Figure 17 – Word Cloud of frequent words in MIT Data Science Course Descriptions

The word cloud seen from the combination of MIT Data Science course descriptions reads much closer to a computer science/information technology degree, with more emphasis on networks, optimization, and programming. Economics is also more prevalant, which lines up with the nature of the degree being offered jointly with the strictly Economics degree. MIT’s mission statement in their catalog of this degre looking to convert these topics to “real-world challenges” is also supported with the frequency of research, social, and theory.

Table 3 -First 9 Topics Generated of MIT Data Science Course Descriptions

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| T\_1 | T\_2 | T\_3 | T\_4 | T\_5 |
| Information | Topics | Course | Management | Programming |
| Economic | Analysis | Concepts | Modeling | Will |
| Analysis | Learning | Topics | technology | Course |
| Current | Algorithm | Handling | Project | Techniques |
| Models | Data | Implementation | Logistics | Computer |

|  |  |  |  |
| --- | --- | --- | --- |
| T\_6 | T\_7 | T\_8 | T\_9 |
| Course | Engineering | Problems | language |
| Health | Software | Application | Design |
| Components | System | Science | Fundamental |
| Healthcare | Design | Solving | Knowledge |
| Tools | Policy | Computer | Computation |

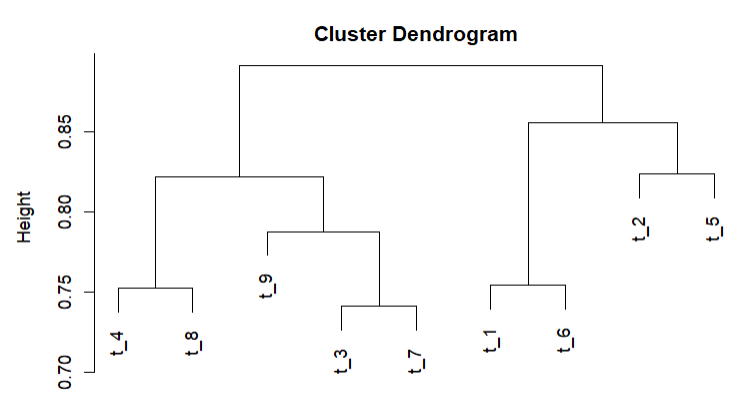


Figure 18 - Dendrogram of MIT Data Science Course Description

Further looks show the segmentation of one side of the cluster dendrogram having information, topics, algorithm, data, programming, techniques, economics as likely representing the coursework and theory side of the curriculum, with management, modeling, project, logistics, language, fundamental, design, policy showing a lot of the projects and soft skills required for deliverables.

## RStudio and Packages

All of the analysis shown was performed in RStudio Desktop Integrated Development Environment (IDE), a standard open course set of integrated tools designed for R productivity and most popular IDE for R [37]. Packages used outside of base R include tidyverse[27], textmineR[28], tidytext[29], pdftools[30], tm[31], wordcloud[32], RColorBrewer[33], topicmodels[34], and qpdf[35].

pdftools and qpdf are primarily for performing transformations and combining the raw form pdf lecture files and transcripts obtained from the MIT OpenCourseWare site. wordcloud and RColorBrewer are for word cloud visualizations and more dynamic color palates. tidyverse and tidytext are some of the most used packages, which provide functions to convert text from and to tidy formats. textmineR and topicmodels are also essential, giving functionality for analyzing diagnostics for topic models and being able to create them.

A full description of specifications in the project can be found in the Appendix below.

# Conclusion

The most notable outcome of the topic modeling performed shows a much larger emphasis in economics in the MIT degree compared to the Florida Poly degree, where as Florida Poly has a higher proportion of analytics and security subtopics. These can likely be attributed to the placement of the degree programs, where MIT 6-14 is advertised as interdisciplinary with Computer Science, Economics, and Data Science whereas Florida Poly is offered as a standalone Data Science degree, although the nature of the material itself is quite interdisciplinary as seen with our discovered topics. There is very little in the way of data security and database management covered in the MIT degree, as opposed to Florida Poly’s inclusion of them.

Future study in the analysis of MOOCs and publicly available course work is key as education continues to branch out in the open distributed digital space. Paths include comparing other top rated open courses in both data science and outside of it, as well as examining the growth and change of the curriculum over time by comparing lecture material in different years to decipher growing trends in the curriculum. A useful application of this could be to text mine job requirements of top industry positions and evaluate to what extent the courses line up with the skills in job requirements.

This study is also open for reproducibility in analyzing any of the over 900 universities internationally that are offering some form of open, online courses [23]. Comparative analysis ally be done on a curriculum level by grouping documents from all courses, as well as at a higher level by performing topic modeling at the course level to get a more granular look at what material in a specific course is being emphasized within the curriculum. This would be ideal for a university that was looking to conduct a proposal to open a new major into a department, or for a faculty member to argue for the inclusion/exclusion of a course based on its coherence to relevant material in the degree.

# References

[1] Goldberg, Carey (4 April 2001). "Auditing classes at M.I.T., on the web for free". The New York Times.

[2] “Audio/video lectures,” MIT OpenCourseWare, Massachusetts Institute of Technology. [Online]. Available: https://ocw.aprende.org/courses/audio-video-courses/. [Accessed: 26-Apr-2022].

[3] “By The Numbers: MOOCs in 2019 — Class Central,” The Report by Class Central, Dec. 03, 2019. https://www.classcentral.com/report/mooc-stats-2019/ (accessed Jan. 14, 2022).

[4] “Proceedings – Educational Data Mining 2021.” https://educationaldatamining.org/edm2021/proceedings/ (accessed Jan. 14, 2022).

[14]” Z. Kastrati, A. Kurti, and A. S. Imran, “Embeddings and topic vectors for MOOC lectures dataset,” vol. 1, Dec. 2019, doi: 10.17632/xknjp8pxbj.1.

[15] N. Nguyen and R. Caruana, “Improving Classification with Pairwise Constraints: A Margin-Based Approach,” in Machine Learning and Knowledge Discovery in Databases, vol. 5212, W. Daelemans, B. Goethals, and K. Morik, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 113–124. doi: 10.1007/978-3-540-87481-2\_8.

[16] C. Zaire, “Is Data Science Still a Rising Career in 2021,” Medium, Nov. 08, 2021. https://towardsdatascience.com/is-data-science-still-a-rising-career-in-2021-722281f7074c (accessed Jan. 14, 2022).

[21] “Computer Science, economics, and Data Science (course 6-14),” MIT Bulletin. [Online]. Available: http://catalog.mit.edu/degree-charts/computer-science-economics-data-science-course-6-14/. [Accessed: 26-Apr-2022].

[22] “Search: MIT OpenCourseWare: Free Online Course Materials,” MIT OpenCourseWare. [Online]. Available: https://ocw.mit.edu/search/. [Accessed: 26-Apr-2022].

[23] Mooc.org, “Massive open online courses: An EDX site,” MOOC.org | Massive Open Online Courses | An edX Site. [Online]. Available: https://www.mooc.org/. [Accessed: 26-Apr-2022].

[24] Kabir, Ahmed Imran, Koushik Ahmed, and Ridoan Karim. "Word Cloud and Sentiment Analysis of Amazon Earphones Reviews with R Programming Language." Informatica Economica 24.4 (2020): 55-71.

[25] “Florida Polytechnic University BS in Data Science Program Concentration Electives and General Education.” [Online]. Available: http://catalog.floridapoly.edu/mime/media/25/469/BS+in+Data+Science+2022-2023+3.11.2022.pdf. [Accessed: 26-Apr-2022].

[26] F. tang, “Beginner's Guide to LDA Topic Modelling with R,” Medium, 14-Jul-2019. [Online]. Available: https://towardsdatascience.com/beginners-guide-to-lda-topic-modelling-with-r-e57a5a8e7a25. [Accessed: 26-Apr-2022].

[27] Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686

[28] Tommy Jones (2021). textmineR: Functions for Text Mining and Topic Modeling. R package version 3.0.5.

[29] Silge J, Robinson D (2016). “tidytext: Text Mining and Analysis Using Tidy Data Principles in R.” \_JOSS\_, \*1\*(3). doi:10.21105/joss.00037 (URL: https://doi.org/10.21105/joss.00037), <URL: http://dx.doi.org/10.21105/joss.00037>.

[30] Jeroen Ooms (2022). pdftools: Text Extraction, Rendering and Converting of PDF Documents. R package version 3.1.1. https://CRAN.R-project.org/package=pdftools

[31] Ingo Feinerer and Kurt Hornik (2020). tm: Text Mining Package. R package version 0.7-8. https://CRAN.R-project.org/package=tm

[32] Ian Fellows (2018). wordcloud: Word Clouds. R package version 2.6. https://CRAN.R-project.org/package=wordcloud

[33] Erich Neuwirth (2022). RColorBrewer: ColorBrewer Palettes. R package version 1.1-3. https://CRAN.R-project.org/package=RColorBrewer

[34] Grün B, Hornik K (2011). “topicmodels: An R Package for Fitting Topic Models.” \_Journal of Statistical Software\_, \*40\*(13), 1-30. doi:10.18637/jss.v040.i13 (URL:https://doi.org/10.18637/jss.v040.i13).

[35] Jeroen Ooms (2019). qpdf: Split, Combine and Compress PDF Files. R package version 1.1. https://CRAN.R-project.org/package=qpdf

[37] RStudio. [Online]. Available: https://www.rstudio.com/products/rstudio/. [Accessed: 26-Apr-2022].

[38] “MITX courses,” Open Learning. [Online]. Available: https://openlearning.mit.edu/courses-programs/mitx-courses?f%5B0%5D=course\_availability%3A62. [Accessed: 26-Apr-2022].

[39] J. S. and D. Robinson, “6 topic modeling: Text mining with R,” *6 Topic modeling | Text Mining with R*. [Online]. Available: https://www.tidytextmining.com/topicmodeling.html. [Accessed: 26-Apr-2022].

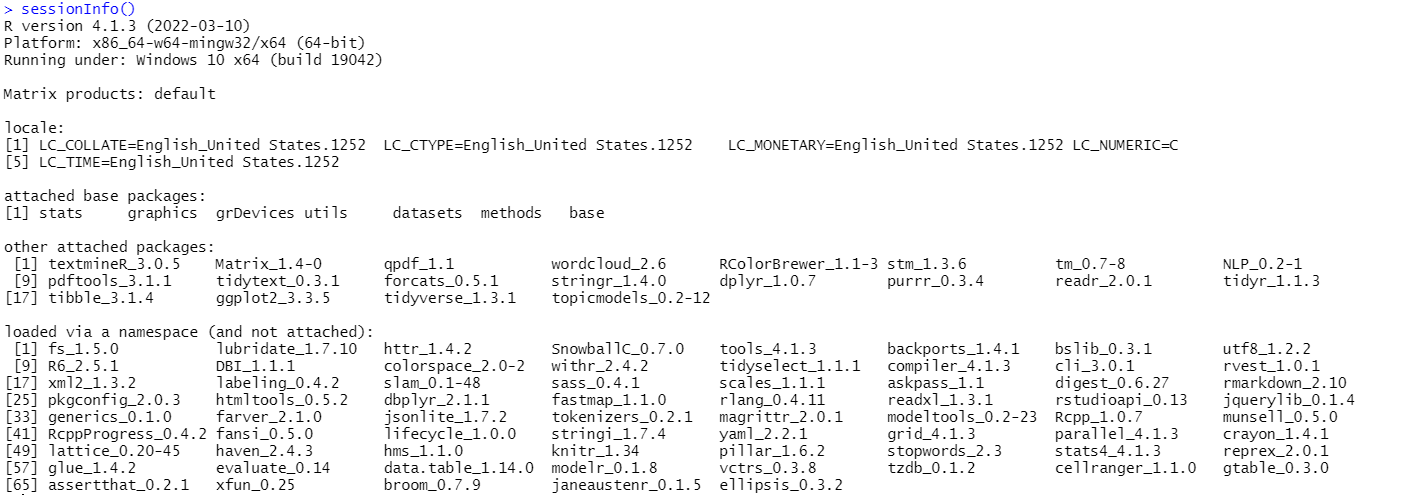
[40] Cabezas, L., Izbicki, R., & Stern, R. B. (2021). Hierarchical clustering: visualization, feature importance and model selection. *arXiv preprint arXiv:2112.01372*.

[41] Kherwa, Pooja, and Poonam Bansal. "Topic modeling: a comprehensive review." EAI Endorsed transactions on scalable information systems 7.24 (2020).

[42] Tong, Zhou, and Haiyi Zhang. "A text mining research based on LDA topic modelling." *International Conference on Computer Science, Engineering and Information Technology*. 2016.

[43] Syed, Shaheen, and Marco Spruit. "Full-text or abstract? Examining topic coherence scores using latent dirichlet allocation." *2017 IEEE International conference on data science and advanced analytics (DSAA)*. IEEE, 2017.

# Appendix



R Studio Session Info

Github repository:

https://github.com/gmantini/MIT-OpenCourseWare-Data-Science-Text-Mining