CMU Course Landscape Scan: Information & Data Literacy

By: Madden Moore, Tong Hu, Jenny Shan
Project Advisor: Zach Branson Project Supervisors: Peter Freeman, Jamie McGovern External Advisor: Joanna Dickert

Background and Introduction

- Overall goal: To identify which courses fit the Data and Information Literacy initiatives from the CMU Core Competencies Initiative
- Data Literacy: Evaluating and using existing data, Generating and evaluating new data
- Information Literacy: Identification and navigation of the information landscape, Critical evaluation of information, Information engagement in and across communities of practice, Shaping the information ecosystem
- Dataset: CMU courses from 2019-2022 - course name, department, ID, description, etc.

Data Processing

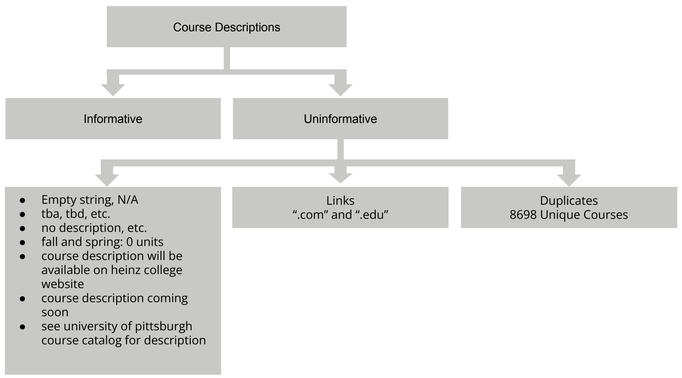
- To classify courses, it is important to analyze the description of courses.
- We eventually found some “uninformative” descriptions:

```
[1] "TBA"          [1] "No course description provided."          [1] ""
```

- To contrast, there is an informative description::

```
[1] "This course is designed to give undergraduate students experience using statistics in real
research problems. Small groups of students are matched with clients and do supervised research
for a semester.  Students gain skills in approaching a research problem, critical thinking,
statistical analysis, scientific writing, and conveying and defending their results to an
audience and to the external client who provided the data."
```

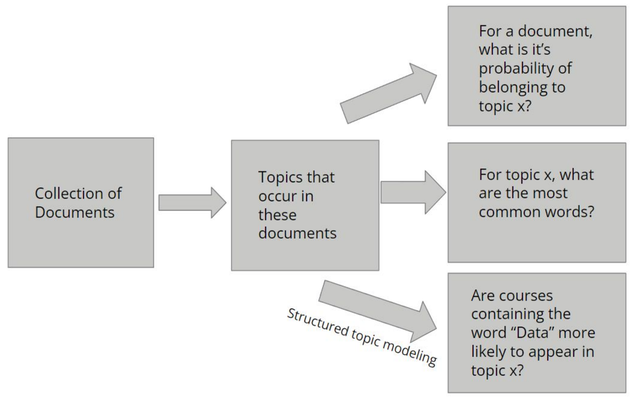
- We also found courses with merely link with uninformative content. For example, some course descriptions contain nothing but a link.
- we also notice that duplicates might cause bias in our modeling. Thus, the final step to the data preprocessing is to filter out courses with the same course ID and course descriptions.
- We found 8698 distinct courses, and after filtering out other uninformative courses, we finally found 7562 distinct courses with informative course descriptions.

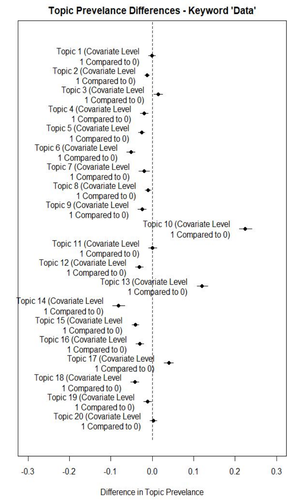


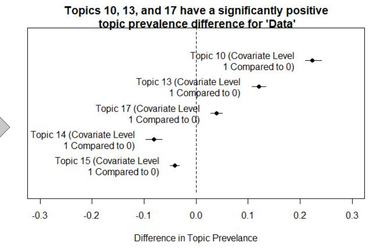
Conclusion

- We were able to map information and data literacy skills (as identified by the campus-wide working group) to CMU courses via text mining and topic modeling.
- Using keywords and phrases from the competency definitions, we identified course topics with positive and negative prevalences to information and data literacy, and spotted specific course levels and course departments that contribute the most to CMU students' information and data literacy development.
- We hope the results from the course landscape scan would help with better course design at CMU in the future.

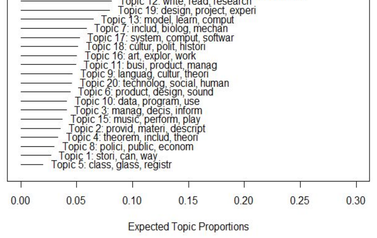
Topic Modeling Analysis and Results







For each keyword, topics with a topic prevalence difference far from 0 were selected for further analysis.



Positive Prevalence Difference Topics for Keyword “Data”

Data science topic

- Keywords: data, program, analysis, python, java, object-oriented, algorithm
- Top courses: 95-481 “Web Application Development”, 95-885 “Data Science and Big Data”, 90-819 “Intermediate Programming with Python”, 67-364 “Practical Data Science”, 17683 “Data Structures for Application Programmers”

Machine learning topic

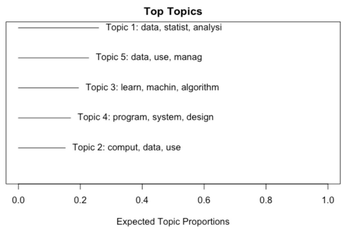
- Keywords: model, learn, compute, method, statistic, machine, regress, bayesian, model
- Top courses: 10-301 “Introduction to Machine Learning”, 18-752 “Estimation, Detection, and Learning”, 46-926 “Machine Learning I”, 46-927 “Machine Learning II”

Systems and software topic

- Keywords: system, compute, software, robot, architecture, sensor, hardware
- Top courses: 18-349 “Introduction to Embedded Systems”, 18-685 “Flexible Energy Systems”, 18-452 “Wireless Networking and Applications”, 15-213 “Introduction to Computer Systems”

Topic modeling provides collections of documents that match with the data and information literacy keywords.

To further explore the course composition, we explore the subset, the courses containing keyword “data” in the course descriptions, and obtain the following topics and keywords from a new topic model.



Topic 1 Top Words: Statistics, Analysis, Research
Highest Prob: data, statist, analysi, student, use, research, project
Topic 2 Top Words: Computer, Science, Network
Highest Prob: comput, data, use, scienc, function, network, program
Topic 3 Top Words: Learning, Machine, Algorithm
Highest Prob: learn, machin, algorithm, model, data, includ, techniqu
Topic 4 Top Words: Program, System, Design
Highest Prob: program, system, design, softwar, data, applic, comput
Topic 5 Top Words: Manage, Market, Inform
Highest Prob: data, use, manag, system, market, databas, inform

We identify one course that best exemplifies each topic. Each topic corresponds to one department that has the most courses related to data literacy skills. These departments provide courses that contribute to students’ data literacy development.

Example Courses:
Topic 1: 76107 Writing about Data
Topic 2: 15150 Principles of Functional Programming
Topic 3: 05434 Machine Learning in Practice
Topic 4: 15619 Cloud Computing
Topic 5: 45882 Digital Marketing and Social Media Strategy

Topic 1: Statistics and Data Science Department
FREX: causal, interpret, lab, laboratori, report, measur, statis

Topic 2: Computational Biology Department
FREX: genom, function, genet, neurosci, diseas, biomed, biolog

Topic 3: Department of Computer Science
FREX: learn, linear, machin, algorithm, cluster, deep, probabl

Topic 4: Institute for Software Research
FREX: cloud, privaci, web, secur, virtual, java, softwar

Topic 5: Business Department
FREX: market, gis, busi, consum, product, spatial, economi