Matching Sustainable Development Goals to CMU Course Offerings

Chloe Yan, Jiayue Guo, Lavanya Chawla, Peter Wu
External Advisor: Alexandra Hiniker
Faculty Advisor: Zach Branson

Introduction

- Our dataset consists of course descriptions from Carnegie Mellon University (CMU) from the Spring 2020 semester.
- We filtered the data based on the following four-step procedure:

Step 1	Step 2	Step 3	Step 4
• Remove classes with empty course descriptions or where course descriptions is same as course title • Remove URLs and special characters from course descriptions	Remove courses with uninformative courses descriptions, for example: • tbd/tba • to be added by the department • to be added at a later time	Remove cross-listed courses which have the same course description or one course description is a substring of the other	Remove commonly used words from course descriptions, such as "a", "the", "of", "class" and "student"

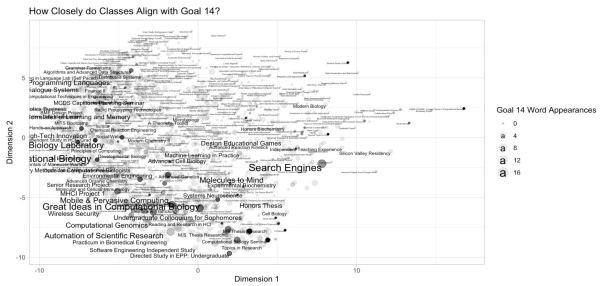
- There are 17 Sustainable Development Goals provided by the United Nations such as Achieving Gender Equality and Ending Poverty.
- Our goal is to explore ways to determine the similarity between CMU classes to each of the 17 goals.

Tf-idf Vectorization

- We utilized a metric called tf-idf (term frequency-inverse document frequency) to pinpoint words specific to each goal.
- Tf-idf in our project is a statistical measure that represents how relevant a word is to a specific goal (and not other goals).
- We calculated the tf-idf for each word in the 17 goals to find the top 25 words specific to each goal.

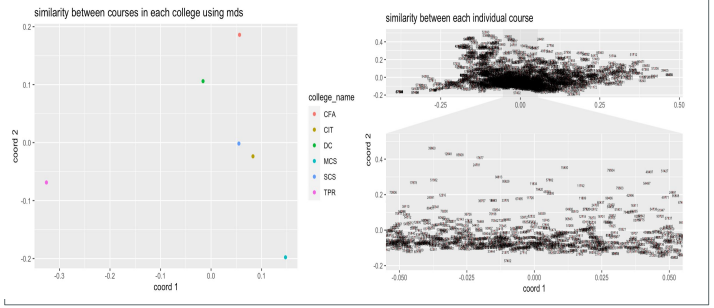
	goal_num	word	num_word_goal	tf	idf	tf_idf
1	6	water	19	0.091346154	1.4469190	0.13217048
2	14	marine	14	0.041055718	2.8332133	0.11631961
3	15	biodiversity	18	0.047872340	2.1400662	0.10244998

- We mapped each class to a 17-dimensional vector representing the number of times a class description contains a top-25 word (that we computed through tf-idf) for that goal.
- As an example, our tf-idf is able to match classes related to biology with Goal 14 on marine resources for sustainability.



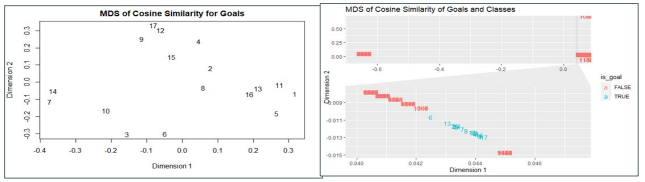
Measuring Similarity between Colleges & Classes

- We first concatenated all the course descriptions into one string for each college, then we measured the cosine similarity between each college, and use multidimensional scaling to visualize the distance of college, in terms of their course descriptions.
- TPR tends to be far apart from other colleges. SCS and CIT tend to be close together, which makes sense since they both have a lot to do with engineering.
- We again use cosine similarity to compare similarity between each individual course , and use multidimensional scaling to visualize the distance between courses.
- Courses form a huge cluster that centered at coord1 = 0, with some small clusters scattered around it.
- MDS correctly represent the similarity between courses: courses in the same department tend to be close to each other. For example, at the bottom left corner, we have most courses from drama school.

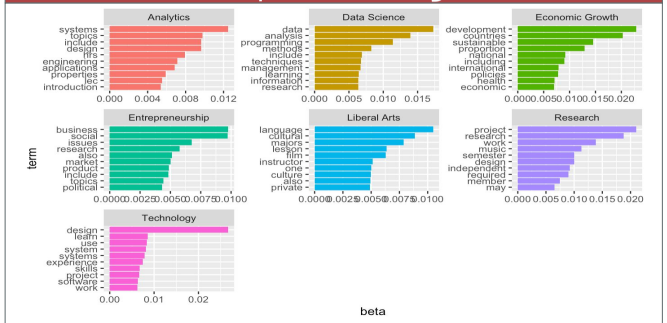


Measuring Similarity between Goals & Classes

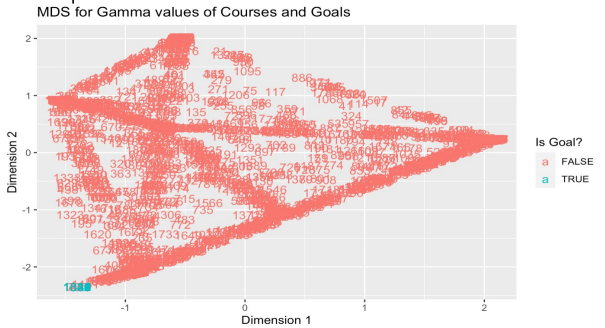
- We measured the similarity between goals using cosine similarity. The observed clusters can be divided into action-oriented (7 and 14; 13 and 16) and similar policy frameworks (9, 12 and 17; 1 and 11).
- We also performed cosine similarity on both goals and similarity and found three main clusters.
- The cluster with all the goals (identified in the zoom) can be used to identify classes similar to goals in terms of cosine similarity.
- We found that certain classes in Music and Drama departments were the furthest from the goals.



Topic Modeling



- We created a 7-topic model based on the courses and goals. A document has 7 different topic proportions (gamma values).
- The beta value of a word is the probability of the word occurring in that topic.



- There are 6 clusters, all the goals are in the bottom cluster.
- The maximum gamma value for the courses/goals in the same cluster correspond to the same topic.
- Starting from the top corner, going clockwise, ending with the middle cluster, the topic with the highest gamma value per cluster is Analytics, Research, Liberal Arts, Entrepreneurship and Economic Growth, Data Science, Technology.

Conclusion

- We used a variety of methods to analyze similarity between CMU courses and the 17 goals.
- We are able to utilize tf-idf to identify words specific to goals and compare classes' relationships to the goals.
- We use cosine similarity to compare similarity between colleges.
- We can compare goals and classes based on their cosine similarity.
- We can compare similarity of courses and goals based on their gamma values found in topic modeling.