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”

Introduction

- There are 17 Sustainable Development Goals provided by the United Nations such as Achieving Gender Equality and Ending Poverty.

- Our goal is to explore methods to determine the similarity between CMU classes to each of the 17 goals.

tf-idf Vectorization (CMU classes → mathematical objects)

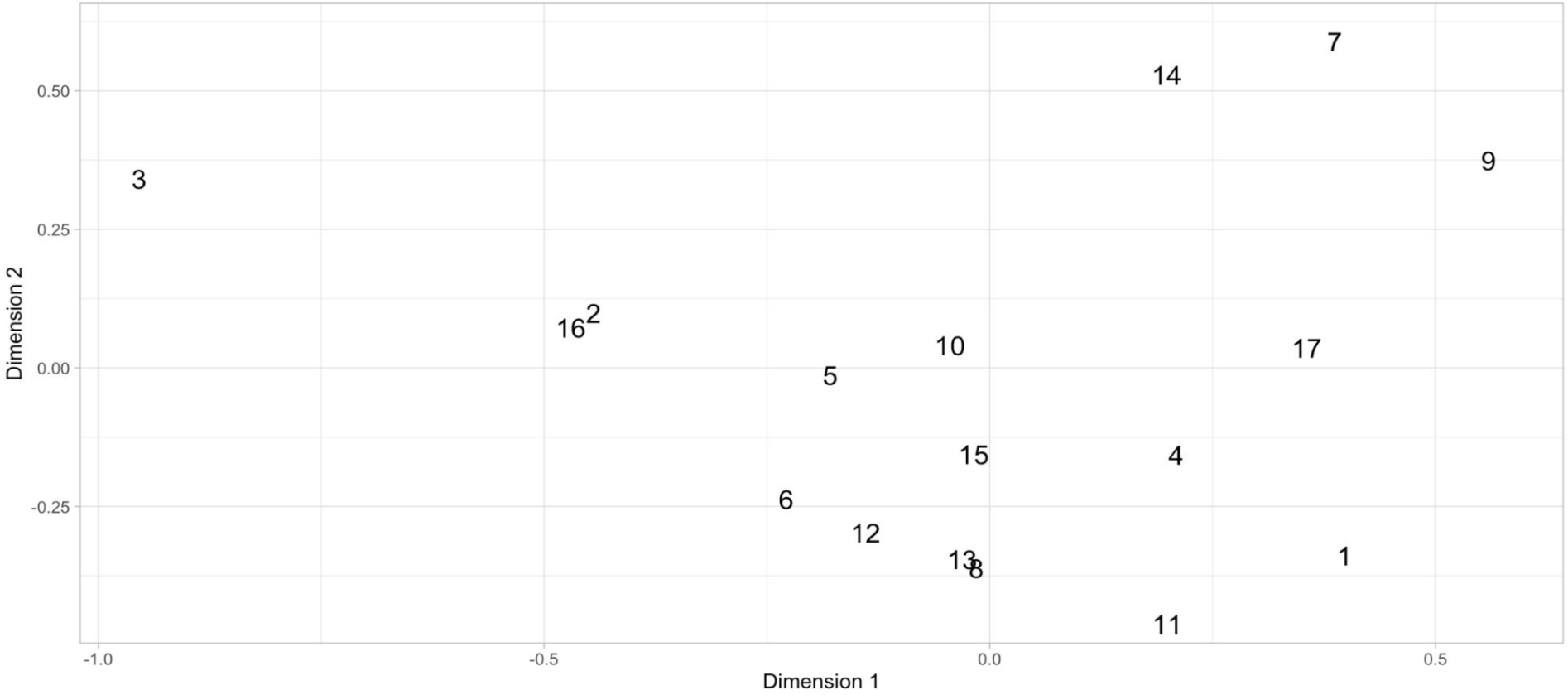
- tf-idf is a statistical measure that represents **how relevant a word is to a specific goal** (and not to the other goals).
- We calculated the tf-idf for each word in the 17 goals to find the top 25 words specific to each goal.

	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
4	14	fisheries	12	0.035190616	2.8332133	0.09970252
5	9	industrial	12	0.046153846	2.1400662	0.09877228
6	7	renewable	5	0.042372881	2.1400662	0.09068077
7	7	developing	8	0.067796610	1.2237754	0.08296783
8	2	agricultural	13	0.036723164	2.1400662	0.07859000
9	3	mortality	12	0.027027027	2.8332133	0.07657333
10	13	climate	10	0.042372881	1.7346011	0.07350004

Identifying Keywords from Goal 14

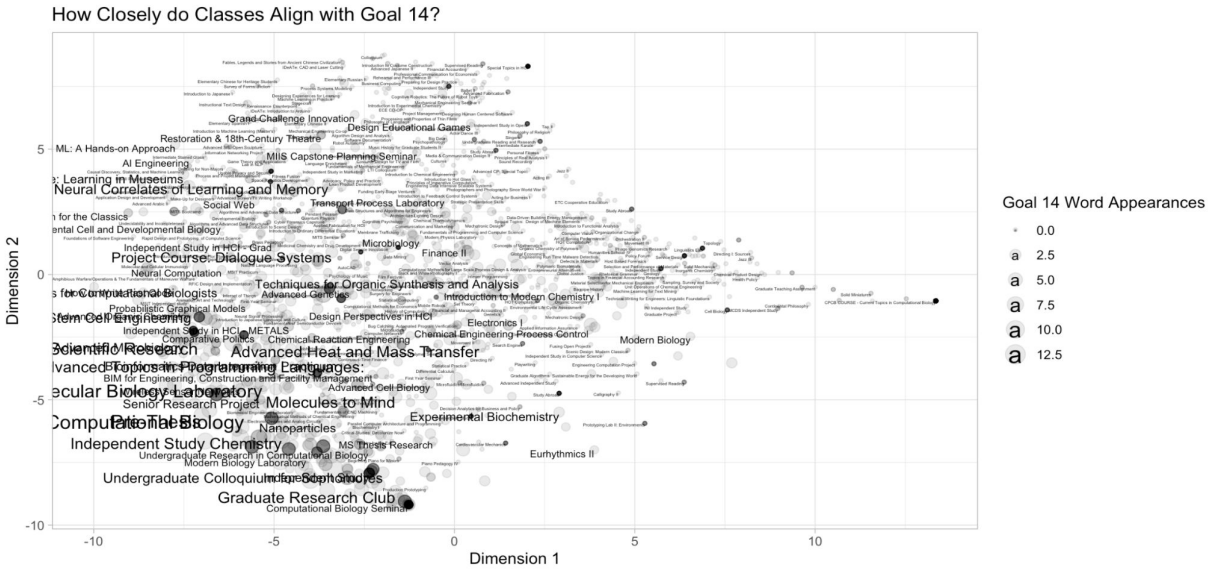
- We integrated our domain expertise of the goals to refine our keywords to only the most important words.
- **Goal 14 = “Conserve and sustainably use the oceans, seas and marine resources for sustainable development”**
- "marine" "developing" "law" "nations" "sea"
  "conservation" "contribute" "illegal" "biological" "recognizes"
  "reflected" "restoration" "stocks" "resources" "scientific"
  "impacts" "transfer" "research" "management" "convention"
  "degree" "legal" "pollution" "significant" "smallscale"
- "law" "nations" "conservation" "contribute" "illegal"
  "biological" "restoration" "stocks" "resources" "scientific"
  "impacts" "transfer" "research" "management" "pollution"

Which goals tend to appear together in the classes?



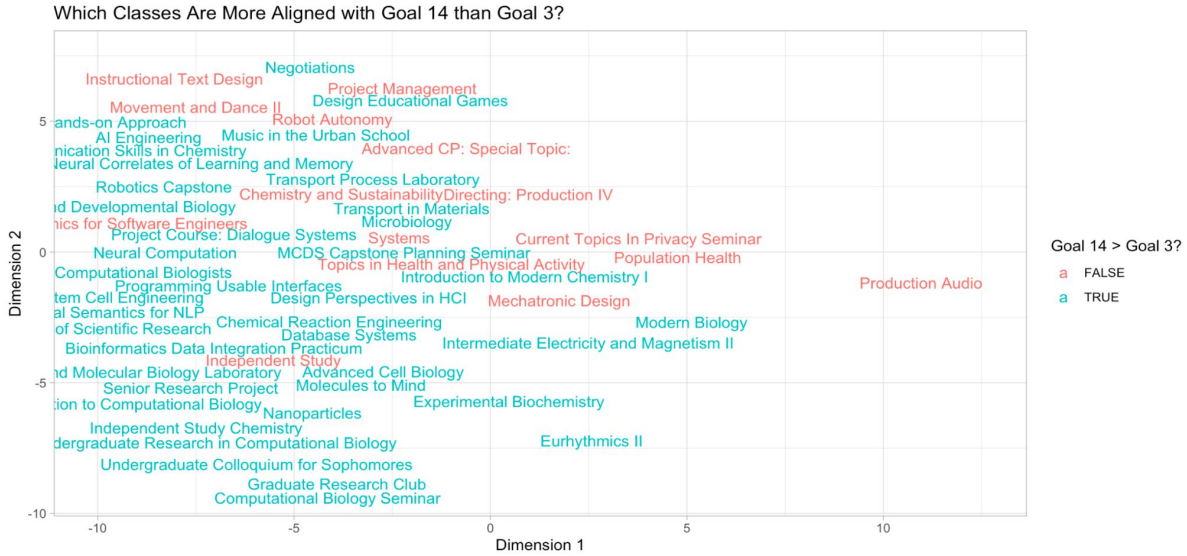
Mapping Goals to Classes

- We mapped each class to a 17-dimensional vector representing the number of times a class description contains a top keyword
- **Goal 14 = “Conserve and sustainably use the oceans, seas and marine resources for sustainable development.”**



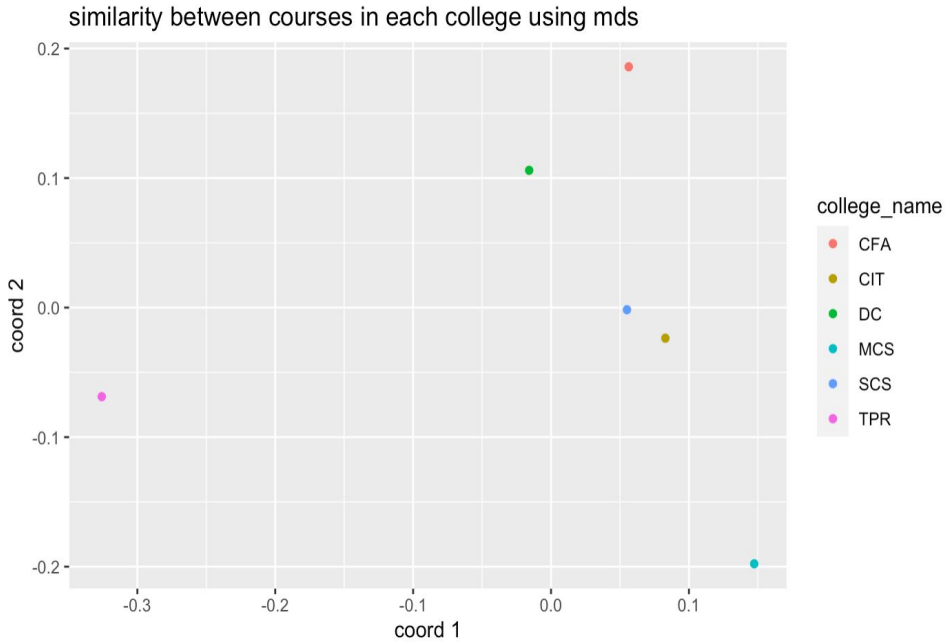
Comparing Classes' Relationship to Goals

- Goal 3 = "Ensure healthy lives and promote well-being for all at all ages."
- Goal 14 = "Conserve and sustainably use the oceans, seas and marine resources for sustainable development."



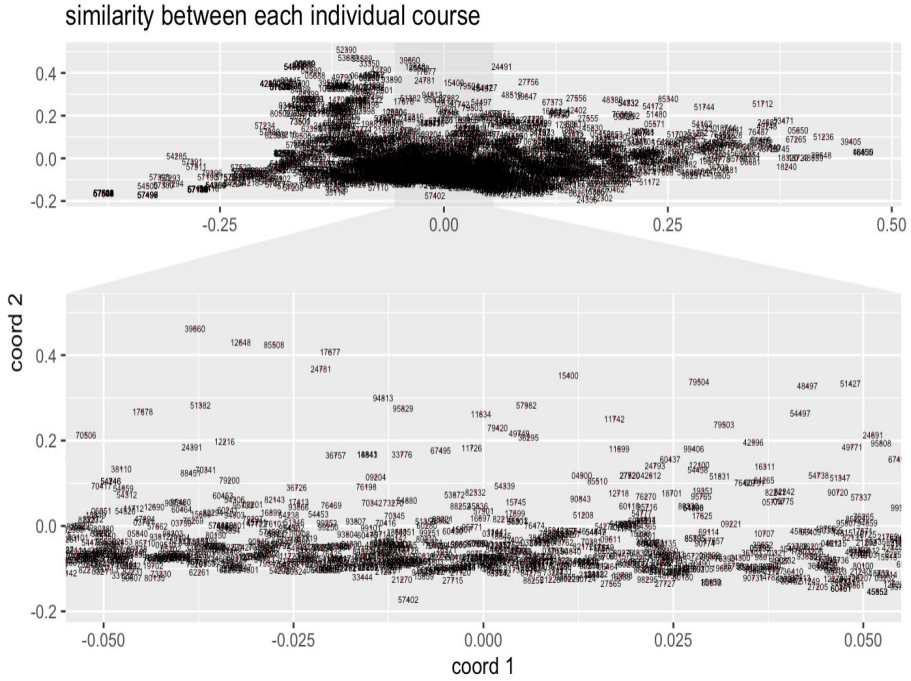
Similarity Between Colleges

- Multidimensional scaling (MDS) is a way to visualize the distance matrix, which contains the pairwise distance(similarity)of courses.
- We are measuring the similarity between two courses using cosine similarity on a document-term matrix where each row keep tracks of the frequency of each word (term) in the document.
- We concatenate all the course descriptions for a single college into one string, and use MDS to visualize the similarity between different college.



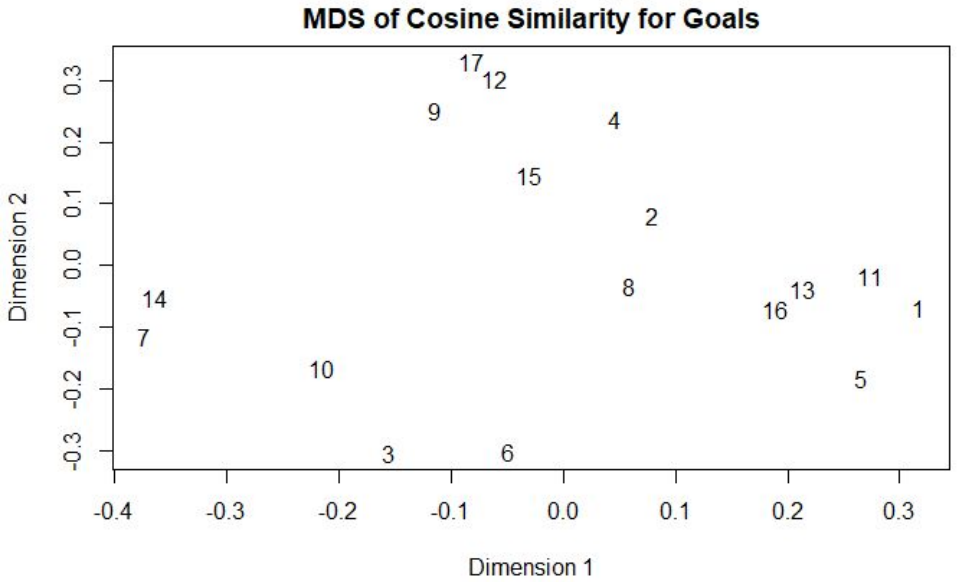
Similarity Between Pair of Courses

- We use cosine similarity and MDS on each individual course description.
- Courses cluster around coord1=0. And courses from the same department are close to each other.



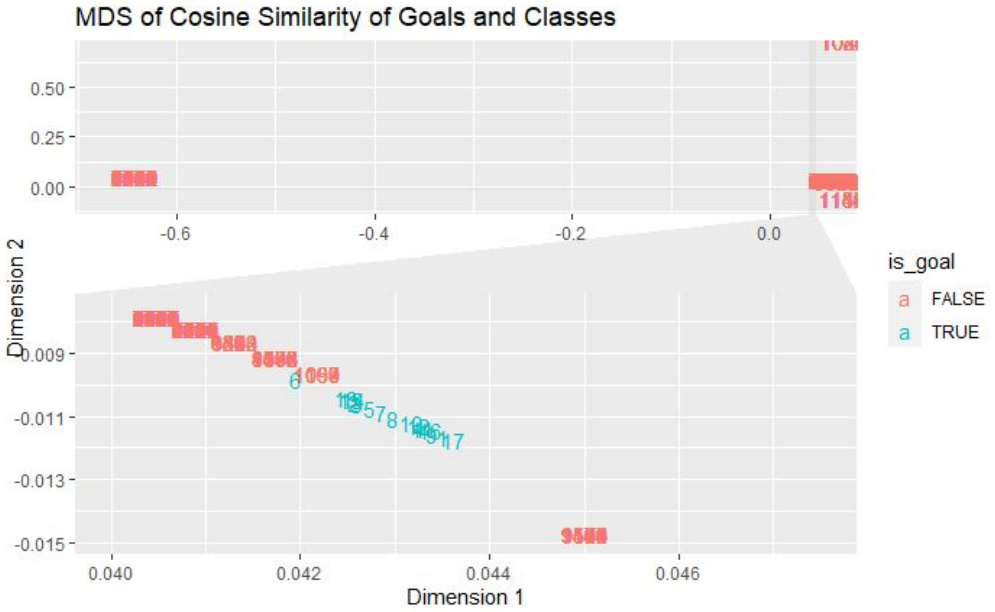
Similarity between Goals

- We performed cosine similarity and MDS on goals.
- We found clusters of goals that were similar in terms of action-oriented (14 and 7; 13 and 16) and in terms of policy (9, 12 and 17; 11 and 1)



Similarity between Goals and Classes

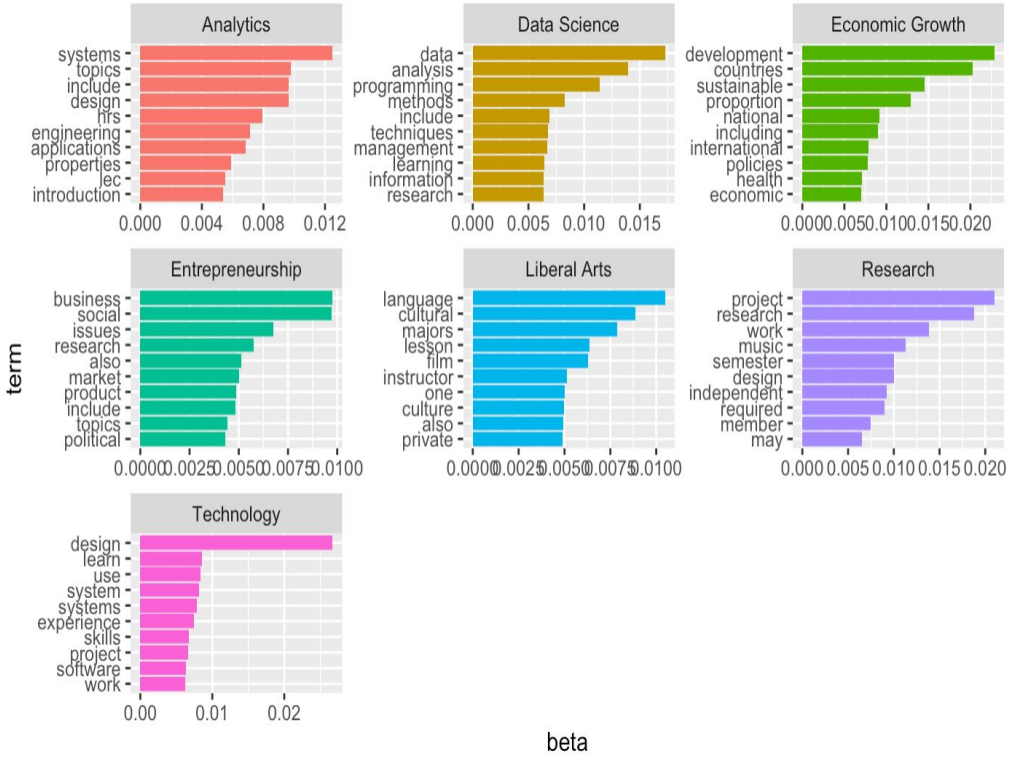
- We performed cosine similarity and MDS of goals and classes.
- We found that certain classes tended to be closer to goals than others. Certain classes in Music and Drama were the furthest from the goals.



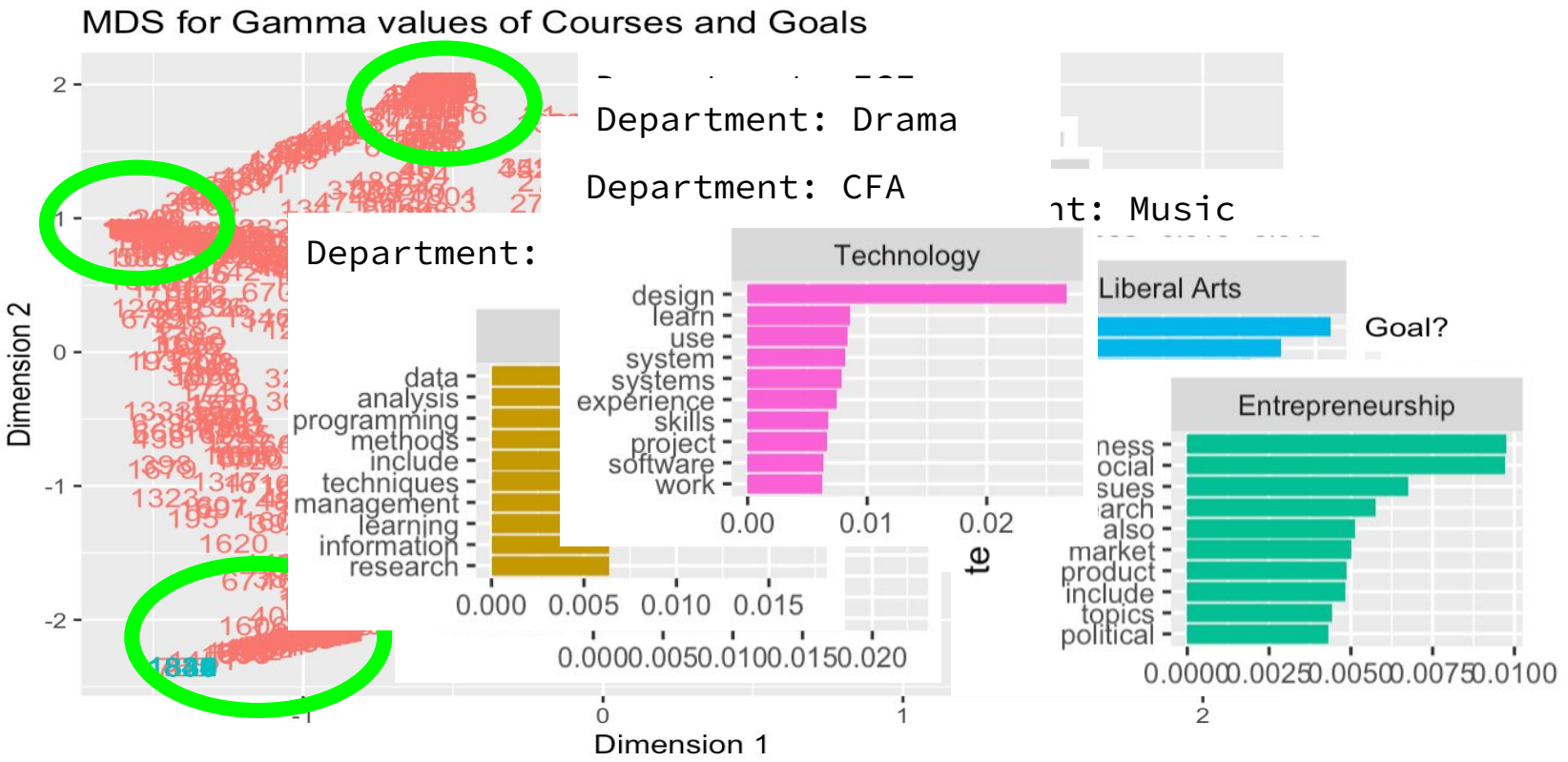
Topic Modeling

- A statistical model for discovering abstract topics that occur in a set of documents. 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.



Topic Modeling (Multidimensional Scaling)



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 can compare colleges and courses based on cosine similarity and MDS.
- 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.