Mapping Sustainable Development Goals to CMU Course Offerings

By: Shannon Sun, Angela Chen, Logan Saito
Faculty Advisor: Zach Branson
External Advisor: Alexandra Hiniker

Introduction

- Our dataset consists of seven semesters (Fall 2019 to Fall 2021) of course descriptions from Carnegie Mellon University (CMU).
- We analyzed this data and the United Nations' 17 Sustainable Development Goals (SDGs).
- Our goal was to match courses to SDGs they address.

Data Processing

- We processed our data by removing invalid/duplicate descriptions, stopwords, whitespace, punctuation, numbers, and URLs, and stemming words.
- We then put the data in a document term matrix (DTM), a large matrix where each row represents a course and each column represents the frequency of a word in the course description or goal.
- We explored the use of different distance metrics, primarily cosine similarity (below on the left) and euclidean distance (below on the right), and applied them to our DTM to generate distance matrices.

$$\cos(\theta) = \frac{\mathbf{A}\cdot\mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}, \quad d(x,\, y) = \sqrt{\sum_{i=1}^{n}(y_i - x_i)^2}$$

- Cosine similarity is a good metric for course-goal matching, but isn't easily interpretable on a plot due to plot axes being arbitrary in quantity.
- Euclidean distance is used instead to find dissimilarity between colleges and departments.

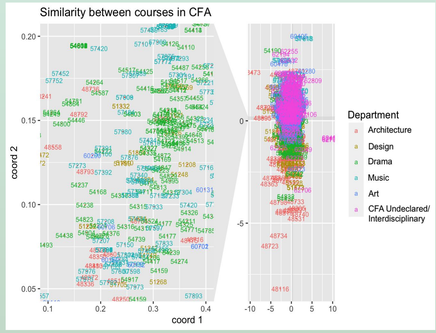
Keywords

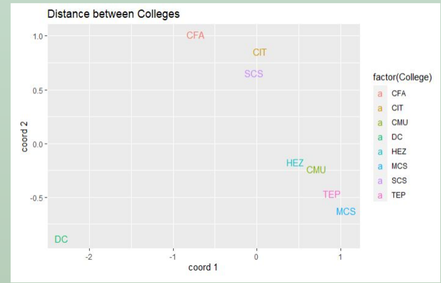
Goal 1	Goal 2	Goal 3	Goal 4	Goal 5	Goal 6	Goal 7	Goal 8	Goal 9	Goal 10	Goal 11	Goal 12	Goal 13	Goal 14	Goal 15	Goal 16	Goal 17
poverti	agricultur	mortal	educ	girl	water	energi	labour	industri	migrat	citi	wast	climat	marin	biodiv	violenc	south
disast	food	disea	learn	women	sanit	renew	employ	research	cost	disast	consumpt	adapt	ocean	forest	bribe	debt
dimen	plant	health	skill	sexual	wastewat	clean	growth	infrastructur	per cent	municip	materi	chang	fish	degrad	author	partnership
poor	preval	medicin	secondari	empow	freshwat	modern	materi	valu	remitt	settlement	food	disast	fisheri	speci	experienc	statist
social	farm	vaccin	vocat	reproduct	drink	fuel	migrant	credit	destin	urban	compani	secretariat	coastal	halt	public	coher

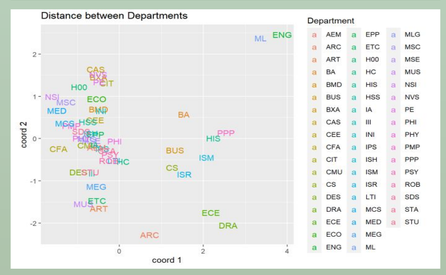
- We use tf-idf (term frequency-inverse document frequency) weighting to define the top 25 unique, keywords for each goal (see above).
- We used these keywords to compare courses with goals and see how well each college address the SDGs
- For instance, Heinz seemed to do the best overall at addressing each goal (see left plot in EDA for SDGs).

MDS between Colleges, Departments, & Courses

- We used multidimensional scaling (MDS) to visualize and display the distances between course offerings on a 2D plot.
- MDS is generated using Euclidean distance matrices using our course DTMs.
- The closer two courses are on a MDS plot, the more similar their descriptions are in terms of word usage. If courses in a given strata are more spread apart, there is high variation between course descriptions within the strata. Ex: see top right plot.
- We see that courses offered by the same department tend to be clustered together, indicating they have similar course descriptions.
- We also applied MDS to explore differences in course offerings between different colleges.s (middle right) and departments (bottom right)
- Due to the unequal numbers of courses offered across colleges, we standardize them by reporting the mean number of courses tagged as a fraction of total courses in the college.
- From our middle right plot, we can see that CIT (College of Engineering) and SCS (School of Computer Science) are close to each other, suggesting their course descriptions have similar word usage.

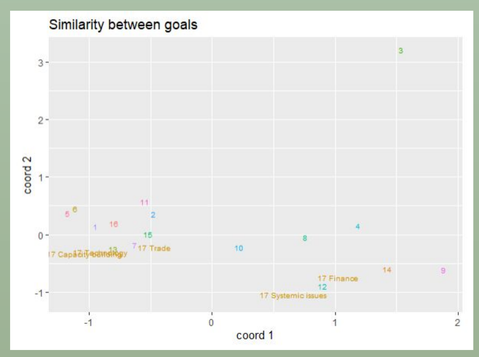






EDA for SDGs

Mean keyword count per goal per college

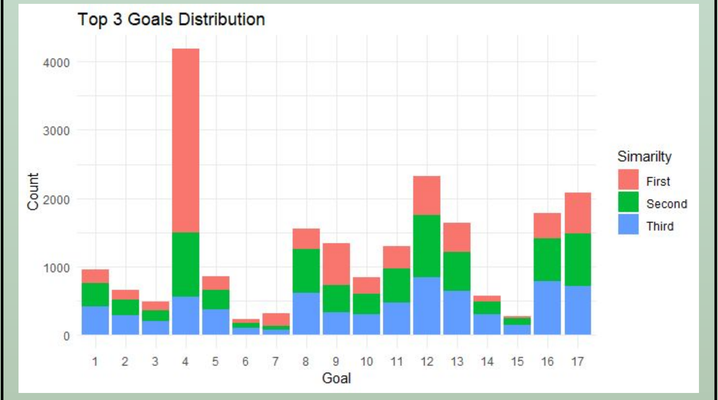


MDS plot between the SDGs

Matching Goals to Courses

Semester	Course Number	Course Title	Course Description	CS SDG 1	CS SDG 2	CS SDG 3
F19	80335	Social and Political Philosophy	Broadly speaking, political philosophers are interested in wh...	10	1	5
F19	18747	Wireless Device Architecture	Growth of the Internet of Things depends on semiconducto...	7	12	17
F19	36707	Regression Analysis	This is a course in data analysis. Topics covered include: Sim...	17	11	1

- Using cosine similarity, we were able to tag courses with the 3 SDGS they were the most similar to (as illustrated above).
- After matching all the courses with their 3 closest SDGs, we explored the overall distribution of tagged goals (illustrated below).



- Overall, CMU is doing well at addressing goal 4 (quality education), goal 12 (responsible consumption and production), and goal 17 (partnerships for the goal).
- However, CMU is doing poorly at addressing goal 6 (clean water and sanitation), goal 15 (life on land), and goal 7 (affordable and clean energy).

Conclusion

- We used a variety of methods to analyze the similarity between CMU courses and the 17 UN SDGs.
- By exploring different approaches using methods such as tf-idf weighting, similarity measures, and MDS, we were able to match courses with the 17 SDGs.
- We used Euclidean distance to measure similarity between colleges, departments, and courses.
- We were able to see how well each college (and CMU overall) does at addressing each SDG.