Michelle Contreras-Catalán
Professor Lee Spector
COSC 247 - Machine Learning

Analyzing Amherst College's Curriculum Diversity Through NLP Clustering

Overview of the project

This project intends to analyze what are the contents currently being taught at Amherst, more specifically, I will try to address whether we are offering a diverse curriculum.

I will be using NLP K-means clustering to cluster the descriptions of all classes offered during Fall 2023 at Amherst College. The number of clusters was determined by using the elbow method, and I assessed the performance of the model by computing the silhouette score. Although I was not able to get a "good" silhouette, it is worth noting that in this context it is not so bad to have a low silhouette score because some classes are often cross-listed among different departments, and it aligns with the interdisciplinary offerings of the curriculum.

After determining an acceptable number of clusters, I decided to move on and perform K-means for a k of 100. Then, I created two visualizations, one static and one interactive, to see how all one hundred clusters behaved.

Since there is a lot of information that could be pulled from this model, I mostly focused on analyzing the diversity in the curriculum. It is worth mentioning that defining what a diverse curriculum means can be difficult, so, for this project, I will try to analyze how clusters behave by department. Given that I am not familiar with all departments, I will analyze those departments that I am the most familiar with, i.e. Computer Science and Mathematics. As a disclaimer, this is subject to my own interpretation, and it will be helpful to consult with students from other departments in future work.

Analysis by department

Here I will provide some analysis of how clusters within each department behaved. These are some of the questions that will be addressed: Is there one cluster in the department that seems to be "popular"? Are there any outliers in these clusters? If so, why could this be happening?

Michelle Contreras-Catalán
Professor Lee Spector
COSC 247 - Machine Learning

Are there any classes within the departments that are not clustered with other courses in the subject? If so, where are they being clustered?

Mathematics

Classes in the math department are being clustered into eight different groups, with cluster 24 being the most popular one:

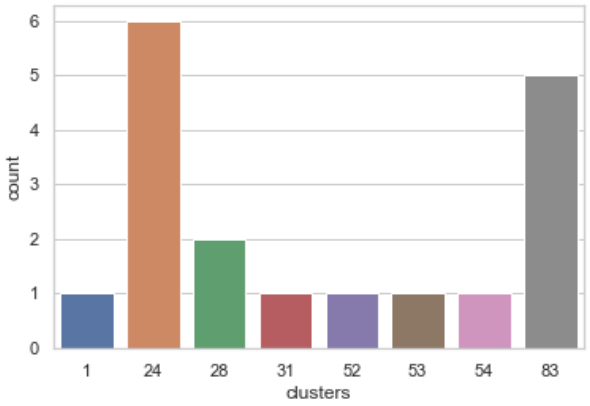

Figure 1

Cluster 24
MATH 111 - Intro to the Calculus
MATH 121 - Intermediate Calculus
MATH 211 - Multivariable Calculus
MATH 220 - Mathematical Reasoning
MATH 271 - Linear Algebra
MATH 272 - Linear Algebra W Applica
SPAN 490 - Special Topics
ECON 111 - An Intro to Economics

Figure 2

Interestingly, cluster 24 aligns pretty well with some of the <200 core courses for the math major (6 classes meet this criterion). The fact that ECON 111 is also included in the cluster makes sense, as it does involve some basic mathematics (particularly related to Calculus I-II). There is one outlier which is SPAN 490.

Cluster 83
ECON 361 - Advanced Econometrics
MATH 260 - Differential Equations
MATH 280 - Graph Theory
MATH 315 - Comp Algebraic Geometry
MATH 460 - Analytic Number Theory
MATH 256 - Combinatorial Geometry

Figure 3

Cluster 83 is the second most popular in Mathematics, consisting of the ones listed in Figure 3. These courses are mostly electives that could be considered more “applied”. It also includes advanced econometrics, which also fits within the category of applied mathematics.

Last but not least, let us take a look at the rest of the clusters. Cluster 28 groups together two of the other math core classes: MATH 350 - Groups, Rings and Fields, and MATH 355 - Intro to Analysis, which are both considered to be the more advanced components of the core. Some of the other clusters are Math special topics and honors, which are clustered

with other STEM courses of the same type. Some clusters that are also interesting to point out are Cluster 54 (Fig 5) and Cluster 53 (Fig 4), as they include new electives, that did not seem to fit anywhere else, which for the terms of this project, seems to diversify the curriculum.

Cluster 53
BIOL 191 - Molecules, Genes & Cells
ECON 111E -- Intro to Economics/Envir
ECON 435 - Adv Open-Econ Macroecon
MATH 325 - Calculus of Variations
PSYC 122 - Statistics for Behav Sci

Figure 4

Cluster 54
BIOL 280 - Animal Behavior
BIOL 281 - Animal Behavior w/Lab
MATH 345 - Functns Complex Variable

Figure 5

Computer Science

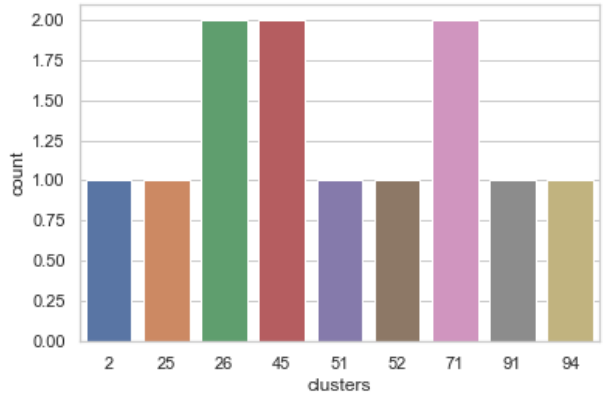

Figure 7

Cluster 71
COSC 211 - Data Structures
COSC 311 - Algorithms
GEOL 401 - Plate Tectonics
PHYS 343 - Dynamics
PHYS 460 - General Relativity

Figure8

When comparing the distribution of clusters of Computer Science versus Mathematics, there is no prominent cluster in CS. We can note, however, that one of the clusters that have two classes grouped is Cluster 71, which consists of the classes listed in Figure 8. It is interesting to see that COSC 211 and COSC 311 are together since these are usually considered as only one class at most colleges. One other point that is interesting about this cluster is that two physics classes are included, which could be seen as incoherent, but comparing this with the number of people who double major in physics and computer science makes sense.

In terms of classes that will be offered for the first time next year, one of them is “Computing Hardware”, which is fairly different from other Computer Science classes. As the title announces, “Computing Hardware” is focused on

Cluster 91
ARHA 187 - Contemp. Native Am. Art
COSC 265 - Computing Hardware

Figure 9

hardware aspects, whereas most of our current offerings focused on software aspects. This class was clustered in group 91, which only includes one other class: ARHA 187 - Contemp. Native Am. Art. Both classes are quite different from each other, but they share that, among their own department they also do not have anywhere else to fit in. Therefore, Cluster 91 can be considered an outlier, and thus these classes can be seen as those that are diversifying the curriculum.

Predicting the cluster of a new course description

Given that sci-kit Learn provides a *predict* function, I will provide the function descriptions of classes not listed for next semester to see where they would fit in the cluster. In doing this, these are some questions that will be kept in mind: Is this class being clustered with other classes from its department? If not, where is it being clustered?

I have chosen three classes that I consider to be different within their department, as they discuss topics that are usually not brought up often in the curriculum offering. Below are the descriptions of the three classes (offered in previous years).

- **Being Human in STEM:** This is an interactive course that combines academic inquiry and community engagement to investigate identity, inequality and representation within Science Technology Engineering and Mathematics (STEM) fields--at Amherst and beyond. We begin the course by grounding our understanding of the STEM experience at Amherst in national and global contexts. We will survey the interdisciplinary literature on the ways in which identity - race, gender, class, ability, sexuality- and geographic context shape STEM persistence and belonging. We will bring this literature into conversation with our own Amherst experiences. These challenging conversations require vulnerability, openness and the ability to tolerate discomfort. We will work from day one to build a brave space whose foundation is trust, accountability and growth. Students will design group projects that apply themes from the literature and our seminar discussions to develop resources and engage the STEM community, whether at the college, local, or national level. Coursework includes critical reading and discussion, reflective writing, and collaborative work culminating in community engagement proposals which students will share with the campus and the broader public.

Michelle Contreras-Catalán
Professor Lee Spector
COSC 247 - Machine Learning

- **Thinking Like a Computer Scientist:** Analytical thinking is inherent in every aspect of computer science. We need to be able to answer questions such as: how do I know that my program works correctly? How efficient is my approach to solving a problem? How does human-readable code get translated into something that can run on physical hardware? What problems are even solvable by computers? In order to study such questions, computer scientists must be able to communicate with one another using a common language, express ideas formally and precisely, and reason logically about these ideas. This course will introduce mathematics as the primary analytical tool used by computer scientists. Topics may include but are not limited to set notation, symbolic logic, proof techniques such as induction and contradiction, and applications of these topics in computer science. Much more important than any individual topic, however, is the experience that students will gain with formal reasoning.
- **Voting and Elections: A Mathematical Perspective:** The outcomes of many elections, whether to elect the next United States president or to rank college football teams, can displease many of the voters. How can perfectly fair elections produce results that nobody likes? We will analyze different voting systems, including majority rule, plurality rule, Borda count, and approval voting, and assess a voter's power to influence the election under each system, for example, by calculating the Banzhaf power index. We will prove Arrow's Theorem and discuss its implications. After exploring the pitfalls of various voting systems through both theoretical analysis and case studies, we will try to answer some pressing questions: Which voting system best reflects the will of the voters? Which is least susceptible to manipulation? What properties should we seek in a voting system, and how can we best attain them?

Being Human in STEM (Offered as BIOL-150, CHEM-250) was sent to Cluster 58, which corresponds to two other classes: AMST 219 - Rebels and Reformers: Women in the Progressive Era and ECON 412 - Pluralist Applied Micro Practicum. On one hand, AMST 219 touches upon many issues of gender, civil rights, and immigration, among others. On the other hand, ECON 412 sees economics from a different angle from what the economics department offers and tries to detach the area from the math behind it. It makes a lot of sense for these classes to be clustered together, particularly, the economics class as CHEM 250 also tries to leave the math and pure science behind to focus more on issues of race, gender, and equity,

among others. This is a clear example of classes that are diversifying the curriculum and that are making the courses by departments expand beyond what is usually offered.

Cluster 7
GEOL 253 - Geospatial Inquiry
GEOL 105 - Oceanography
GEOL 271 - Mineralogy
STAT 360 - Probability
STAT 225 - Nonparametric Statistics

Figure 10

Thinking Like a Computer Scientist (Offered as COSC-121) was sent to Cluster 7, which corresponds to the courses listed in Figure 10. After carefully reading the descriptions for these classes, we can find similarities among them: For example, GEOL 253, puts great emphasis on developing a new thinking perspective on geology, which is more related to using data related to geographical locations. The other two geology classes do not seem to be quite related to the previous description, but they do align with the content itself of the other classes in the cluster. On the other hand, the two statistic classes could be related to COSC 121 as they use logical thinking applied to more computational disciplines.

Last but not least, Voting and Elections: A Mathematical Perspective (Math 150), was grouped with EUST 344/ HIST 344 - Empires in Global Hist and HIST 427 - Citizenship in Empire. The themes of these three classes are fairly similar, as they all approach topics of citizenship, issues of power, and political dynamics, among others. Both of the classes that were already in the cluster are from the History department, so in line with the hypothesis of classes that could diversify the curriculum, we could say that Math 150 does a good job of going beyond what is typically offered in the department. Moreover, students who have taken those history classes could consider taking Math 150 in the future (which does not have any prerequisites).

Conclusions and Implications

Something that is interesting is how some clusters grouped classes that at first glance do not seem quite similar, but that it is common for students to be interested in. So, we could say that one implication of this project is working as a recommendation system for students to explore other areas

Michelle Contreras-Catalán
Professor Lee Spector
COSC 247 - Machine Learning

Related work and suggestions for future work

For this project, I used K-means clustering, which is a hard clustering method. However, I believe that due to the nature of Amherst College, many of our classes could be considered interdisciplinary, and thus, it would make sense to have some room for freedom between clusters.

While looking at what others have previously done, I came up with a few papers where customer data was clustered to create recommendation systems (Mulyawan et al., 2019; Phorasim et al., 2004). It could be interesting to use students' data to analyze what classes they are taking and use this as a recommendation system in Workday, however, due to FERPA regulations this might be hard to accomplish.

Code References

The following websites and stack issues were very helpful in developing this project, with special credits to http://brandonrose.org/top100

- https://www.dataknowsall.com/textclustering.html
- http://brandonrose.org/top100
- https://towardsdatascience.com/a-friendly-introduction-to-text-clustering-fa996bcefd04
- https://www.dataquest.io/blog/tutorial-text-classification-in-python-using-spacy/
- https://medialab.github.io/iwanthue/palettes/
- https://stackoverflow.com/questions/25921762/changes-of-clustering-results-after-each-time-run-in-python-scikit-learn
- https://stackoverflow.com/questions/32244753/how-to-save-a-seaborn-plot-into-a-file
- https://stackoverflow.com/questions/57340142/user-warning-your-stop-words-may-be-inconsistent-with-your-preprocessing

Articles

Mulyawan, B., Viny Christanti, M., & Wenas, R. (2019). Recommendation product based on customer categorization with K-means Clustering Method. *IOP Conference Series: Materials Science and Engineering*, *508*, 012123. https://doi.org/10.1088/1757-899x/508/1/012123

Phorasim, Phongsavanh & Yu, Lasheng. (2017). Movies recommendation system using collaborative filtering and k-means. *International Journal of Advanced Computer Research. 7. 52-59*. 10.19101/IJACR.2017.729004.