Curriculum changes to the College of Arts and Sciences

Removing electives based on current and past enrollment

Reducing credit hours in redesign of BFA curriculum

Logan Eichbaum  
Department of Arts  
College of Arts and SciencesUniversity of Diploma Printing  
[lke5200@psu.edu](mailto:lke5200@psu.edu)

\*Template received from IEEE.org

*Abstract*— The College of Arts and Sciences in the University of Diploma Printing has been having internal issues. For the past five years, there has been stagnant enrollment, changes in student demographics and interests, and shifts in disciplinary approaches. In this paper I will describe the necessary steps we will take to reinvigorate the curriculum and undergraduate student population. To help redesign the curriculum, I will reduce the number of credit hours to 33, define concentration areas for elective courses, and reduce the number of offered electives. To enact these changes, I will be looking at data consisting of the current courses offered, as well as enrollments for the past five years to find trends in popular classes, and classes that are taken together. I will then use my findings to determine which classes to remove from our course list, and which classes should be included in concentration areas for elective courses. The analyses done in this paper will be performed with Excel and KNIME – I will use association rules and a variety of clustering algorithms from past enrollments to create concentration areas for future undergraduate students.

# Business understanding

The current issue is that the University is facing problems in the College of Arts and Sciences. Enrollment has been stagnant for the past five years. Students have become disinterested in topics offered in our current curriculum. The Department of Arts was asked to redesign the BFA curriculum.  The College would like to reduce credit hours from 39 to 33, define concentration areas for elective courses, and reduce the number of electives offered. It seems as if they are trying to encourage more disciplined areas of study, while budgeting more towards areas that students are interested in. In this way, we can change the courses offered and help bring in more students. Using data from the past five years of course offerings, I plan to analyze which courses are not being utilized and find association rules – as well as clusters - for courses that are taken together to create concentration areas.

To be successful in this experiment, we need to group electives into concentrations of three or four without having too many groupings to avoid freedom in selecting other electives. If the number of credit hours is reduced from 39 to 33, with 21 credit hours reserved for core courses, that only leaves potentially four electives for students to enroll in. Therefore, I will look to achieve only three or four concentration groups, with most of them having three electives such that the students retain the opportunity to explore an elective outside their area of study.

Key factors in deciding concentration areas include popularity of courses as well as co-enrollment with other courses. These variables will be solved during our analysis using KNIME.

Now that I have laid out the desired goal of the curriculum shift, I will need to clean the data into a format that is ready for mining. This will involve removing duplicates, standardizing naming conventions, eliminating courses not included in the elective offerings, and making a decision on any incomplete data. These operations may be done using Excel functions. Only after this cleaning can I begin to perform the aforementioned algorithms on the data.

# Data understanding

To collect the initial data, I used the list of current courses offered, as well as a list of all student enrollments for the past five years. This data was retrieved from records obtained by the chair of the committee. The data consisted of over 4,500 enrollments with three columns: name of the students, the terms they enrolled in, as well as the courses they enrolled in. A second dataset listing all the course offerings for the ART major, including foundation courses, core curriculum, and electives was given as well.

With the past enrollment dataset having over 4,900 rows, an initial glance was taken to ascertain the quality of the data. I initially discovered a few issues with the data – the very first listed student, Bill Mumy, has many duplicate entries for courses he had enrolled in. An example is shown here:



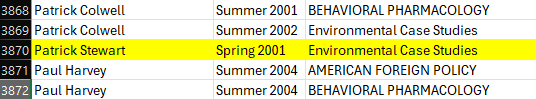
This is the case for many of the students in the dataset. There are also missing values in the dataset – when I sort by course, there are some students with missing course values in the data:



In addition to these potential problems, I saw peculiarites in certain names in the spreadsheet, such as the entry shown below:



Keeping in mind one of the business goals of this project, we want to find association rules for students who take different courses and create concentration areas for courses of interest. To do this, we need sufficient data on each student. Therefore, I might consider ignoring certain rows from the dataset that have just one course listed for enrollment when I sort by name, like the one shown below:



The data also has inconsistencies in its naming conventions. Most entries include first and last names, but some entries also include titles, have added characters in the beginning of the entry, include extra delimiters, or are duplicates. In the courses column the names of some courses are misspelled or abbreviated, leading to duplicate enrollments for some students. Some examples are included below:







The dataset we were given displays enrollments into every course offered at the University of Diploma Printing. This includes courses like Computational Linear Algebra, Analytical Mechanics, and much more. Since we are only interested in courses offered through the Department of Arts for the Bachelor of Science in Art Education, many of these courses will not be necessary in our analysis in creating concentration areas of the degree.

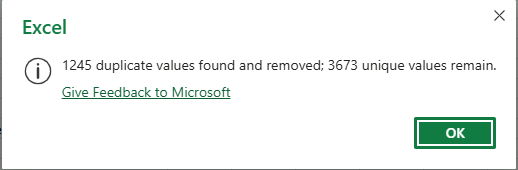
A large portion of this dataset must be preproccessed and cleaned before we can analyze it. I would first reach out to the chair committee to see if I can retrieve any of the missing information on the dataset. Then I need to decide relevant courses to include in my analysis, how I want to standardize my naming conventions, and where I need to remove duplicates. Then, I can start to mine the data for associations between classes that students take.

# Preprocessing

The first issue to take care of to properly prepare the data is to clean the data I am working with. Since we have a large amount of information, and I cannot contact the chair committee to see if I can obtain any of the missing information currently in the database, I will remove the few rows that were missing some information. I retrieve these rows by sorting them by blank cells and deleting the resulting rows.



One of the most prominent issues I noticed when doing an initial scan of the data is that there are many duplicate records. While mining, I do not want to count a duplicate entry multiple times as it could affect which courses I am deciding to remove from the listings. Therefore, I removed every duplicate value and only kept unique ones.



Upon further inspection of the course list, there were two different Art Electives named France & the European Union. Since there are no numbers indicating these courses are part of a series, during this preprocessing state I will keep in mind that I will be removing one of these listings from the course offerings.

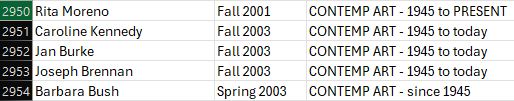
Additionally, many courses that are not in the art program are included in the course enrollment data. For the purpose of this project, we only need to include courses that are part of the core curriculum for the Bachelor of Science in Art Education, as well as electives for the same major. Therefore, I will be removing many rows that have irrelevant data such as the ones below by using a join on columns containing the same name as the courses offered in the program.



While only including course offerings that are available to Art Education majors, I noticed a variation in the naming of specific courses. For example, American Health Policy has a misspelling on several of the entries. ARTS 569: CELL BIOL. AND BIOCHEM has multiple different variations of names in the data. ARTS 497 – CONTEMP ART – 1945 to present also has the same issue with variations in naming structure. There are several other course listings that have inconsistencies in names, which I standardize in the final data set to be mined. Below are examples of the above data that needs cleaning:







In the Data Preparation step I briefly covered having two different naming records for Olympia Snowe. I removed duplicate records for this student and changed the name for these entries. I also found each student that has naming errors and corrected these inconsistencies in the data. An example of incorrectly listed data is shown below.



One of the last things to mention briefly about this data is that there are several elective offerings that have either no listings or have very low enrollment. For example, American Social Policy and Early Balkan History/Society have no enrollments in this dataset, and Evidence Based Crime and Justice Policy only has one enrollment. Before even mining, these will be likely candidates for removal from the curriculum.

Utilizing these cleaning techniques I have reduced the number of rows in the initial data from over 4900 to 2288. With the newly created data containing standardized names, unique entries, and filtered courses, I will now begin to mine the data to find association rules between electives and core curriculum, which will help me decide which electives to ultimately remove but also which electives I should use to create concentration areas for core classes. I will consider other mining techniques as well to help me discover the best results.

# Mining

Filtering Data

Before even beginning the mining process, a method I can use to determine which electives to remove from the course offerings is to simply use the “COUNTIF” function in excel to determine the number of times a course was enrolled for the past few years in these course offerings. I will use a threshold of 8 for these electives to be generous, since a small discussion-based, tightly knit class could still be generated with just a few amounts of students. According to this, the following electives can be removed from the course offerings:

* AMERICAN SOCIAL POLICY – 0 enrollments
* ANALYZING THE POL WORLD – 6 enrollments
* ART AND RELIGION – 7 enrollments
* COMM and THE PRESIDENCY – 7 enrollments
* CONTEMPORARY SOCIO THEORY – 1 enrollment
* EARLY BALCAN HIST/SOC - 0 enrollments
* ELEMENTARY GERMAN 1 – 1 enrollment
* EVIDENCED BASED CRIME AND JUSTICE POLICY – 1 enrollment
* FRENCH THOUGHT SINCE 1945 – 7 enrollments
* FRENCH THOUGHT TILL 1945 – 0 enrollments

Already with this observation, we can remove 10 electives from the 33 elective offerings to generate 23 electives and 9 core courses for data mining listed below:

|  |
| --- |
| * DEVIL'S PACT LIT/FILM |
| * 20TH CENTURY RUSSIAN LITERATURE: FICTION AND REALITY |
| * A WORLD AT WAR |
| * AESTHETICS |
| * AFRICAN-AMERICAN LIT: AFRICAN-AMER LIT:CHANGE |
| * AMERICAN HEALTH POLICY |
| * AMERICAN SOUTH 1861-PRES |
| * ANALYTICAL MECHANICS * ART: ANCIENT TO 1945 – 6 enrollments |
| * AUGUSTAN CULTRL REVOL |
| * BECOMING HUMAN |
| * BEHAVIORAL PHARMACOLOGY |
| * BRITISH POETRY 1660-1914 |
| * Business German A Micro Perspective |
| * CELL BIOL & BIOCHEM |
| * COMMUNICATIONS INTERNSHIP |
| * COMPARATIVE POLITICS |
| * COMPUT LINEAR ALGEBRA |
| * CONTEMP ART - 1945 to PRESENT |
| * CONTEMPORARY AFRICAN ART |
| * CONTEMPORARY POL.THOUGHT |
| * DEVIL'S PACT LIT/FILM |
| * EARLY MESOPOTAM HIST/SOC |
| * ELEMENTARY ARABIC II |
| * ENVIRONMENTAL STUDIES RESEARCH SEMINAR JUNIOR LEVEL |
| * ENVIRONMENTAL SYSTEMS II |
| * EUROPE IN A WIDER WORLD |
| * EXPERIMENTAL WRITING SEM |
| * FOOD/FEAST ARCH OF TABLE |
| * FRANCE & THE EUROP.UNION |
| * FRESHWATER ECOLOGY |
| * AUGUSTAN CULTRL REVOL |

If we include the Foundation courses and the Core courses in the mining process, these will skew the results with association rules and outliers since each student already must take these classes. As a result, I will remove the 2 Foundational Courses and 7 Core courses from the data to be mined for more accurate results.

1. Association Rules

For this project we are interested in how popular certain electives are, and which electives are most taken with core courses that can constitute concentration areas. Essentially, we are looking at how often courses appear on their own and which courses they appear most frequently with. As a result, I will use an association rule miner in KNIME to determine relationships between electives. I will use values such as support and confidence to determine which electives are taken most often together as well as discover which electives are not being utilized and can be removed from the course offerings.

Before mining, however, it is useful to find total counts of electives taken to understand which courses are being utilized most. Unsurprisingly, the core courses have very high enrollment counts, but notably American Health Policy had a high number of enrollments with 143 compared to other offered courses. When deciding which courses to keep in the curriculum, I will likely be selecting American Health Policy as one.

1. DBSCAN Clustering Algorithm

Upon consideration of known clustering algorithms, we can start by using the DBSCAN algorithm due to its ability to create clusters in arbitrary shapes. DBSCAN – or Density-Based Spatial Clustering of Applications with Noise, is an agglomerative algorithm that uses local density to determine groupings for data points [1]. I can use it to assign electives to their own neighborhoods that will allow me more options in selecting concentration areas.

However, in order to use this algorithm, as well as future algorithms, I need to convert the data to a format that can have mining performed on it – currently, the data is just listed as strings. Therefore, I will use string distances to create numerical distances between the courses that the algorithm can take as inputs.

1. K-medoids algorithm

To get a better grasp of clusters we can use to identify election concentrations, I want to use K-medoids to segment the data. The idea would be to randomly create centroids and use the string distances created earlier in the DBSCAN algorithm in order to assign different courses to available centroids. I can use these results and compare them to the DBSCAN algorithm to find insights into what groups of courses students seem to enroll in.

The K-medoids algorithm has the potential to yield different results since I can set the number of clusters that the algorithm will yield before calculation. From removing electives with little to no enrollment, I have 23 electives to work with. To create groups of at least three electives, I will have to use 7 as my number of clusters. This algorithm then randomly selects 7 points in the data set as initial medoids, and forms clusters by assigning points to those initial medoids while re-computing them [1].

1. K-means

Since we have already performed calculations using string distances, a simple yet effective algorithm to consider next is K-means. Similarly to the K-medoids algorithm, we can group the data into 7 predefined clusters [1] and compare the results to Association Rules, DBSCAN, and K-medoids.

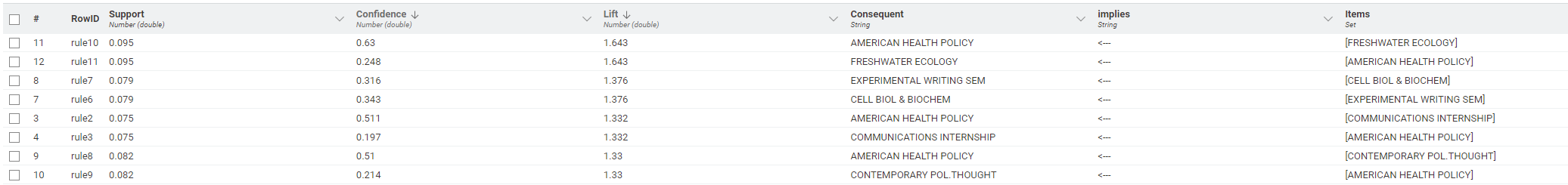
However, while preparing this algorithm, I found that simply using the string distances in conjunction with the filtered data would not work since k-means requires a single input with data points and their numerical values. Therefore, I decided to use a value counter and a normalizer node to create numerical weights to each cluster. Afterwards, I joined the weighted counts with the list of courses to assign weights to each course. I then joined this with the complete filtered elective data and ran it through the K-means algorithm, filtering out student names and count after execution. Then, I removed duplicate rows and created a scatter plot of the resulting 7 clusters.

1. Hierarchical Clustering

The last algorithm I will use to mine the data is Hierarchical Clustering. The output of this will be a dendrogram that can help me determine groupings of three-four electives that I can use in conjunction with the other algorithms. For the agglomerative method I am using, each data point is assigned to its own cluster, and a distance matrix is computed for the clusters. Then, the most similar pairs of clusters are merged until there is just one giant cluster [2].

One advantage Hierarchical Clustering has over the other clustering algorithms is its flexibility and ability to display data in a structured manner. To configure this algorithm, I used the string distance data from prior configurations along with data that I had grouped by course name. This allowed me to only cluster data by the course name.

1. Evaluation
2. Association Rules



The resulting association rules contain both core courses and elective courses. Since we are mainly interested in which electives we are keeping, as well as concentration areas for the electives, it would be more impactful to look at the association rules of just the elective courses. As a result, I filtered out core courses from this mined data and created a new output of association rules based only on the electives.



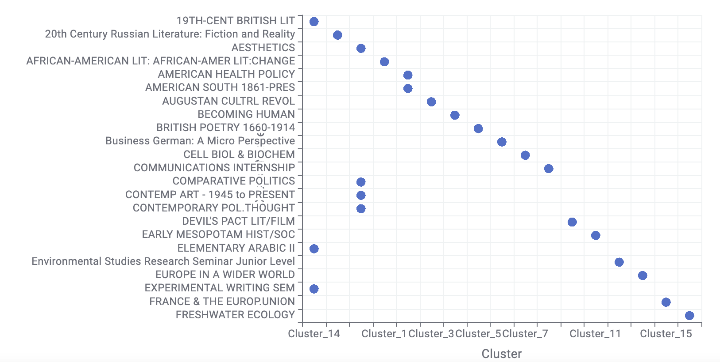
Some key association rules with high lifts are listed below, and will be used for future consideration on election concentrations:

* FRESHWATER ECOLOGY -> AMERICAN HEALTH POLICY
* CELL BIOL & BIOCHEM -> EXPERIMENTAL WRITING SEM
* COMMUNICATIONS INTERNSHIP -> AMERICAN HEALTH POLICY
* CONTEMPORARY POL. THOUGHT -> AMERICAN HEALTH POLICY

It is important to note that this data was generated with a very low support value. While the lift indicates there are actionable insights here regarding 2-itemsets, we should be wary about using these for definitive analysis. So far, I have potential consideration for concentration areas, but I would like to explore the clustering algorithms that can potentially give me better results.

1. DBSCAN

While configuring the DBSCAN node in KNIME, I used trial and error to find an epsilon value that would create clusters of only three or four classes in the data. This way, we would be able to use the results in a way that creates concentration areas for the electives. Below is an output for an epsilon value of 0.65, with a minimum of three data points needed in a cluster since we are looking for clusters of three or four:

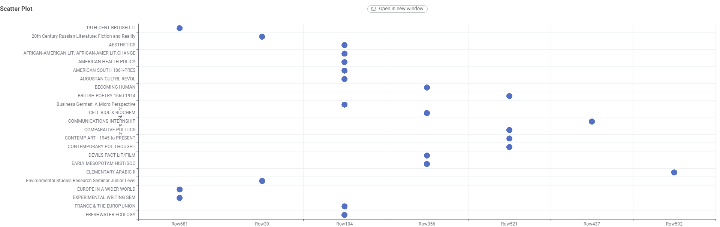


Using the DBSCAN algorithm yielded three clusters that had more than one course within them. The following groupings will be considered at the evaluation phase when deciding which results to use in practice:

* 19TH-CENT BRITISH LIT
* ELEMENTARY ARABIC II
* EXPERIMENTAL WRITING SEM
* AESTHETICS
* COMPARATIVE POLITICS
* CONTEMP ART – 1945 TO PRESENT
* CONTEMP POL. THOUGHT
* AMERICAN HEALTH POLICY
* AMERICAN SOUTH 1861-PRES

1. K-medoids

To configure the K-medoids algorithm, I set the partition count to 7, since we are trying to create clusters of 3-4 out of a list of 23 electives. I will look at the resulting output and use it for more insights on which electives to group into concentration areas alongside the output for the DBSCAN algorithm.



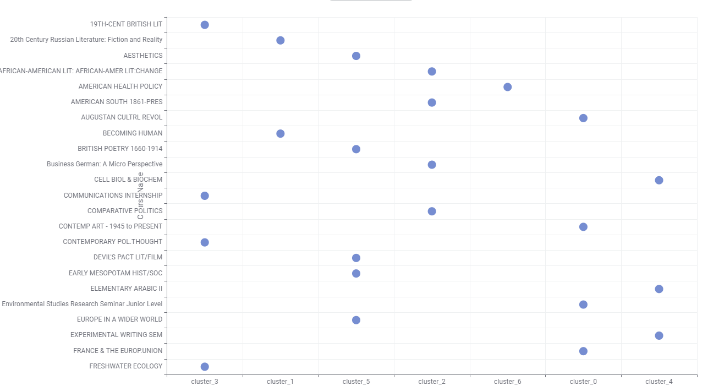
One drawback of the k-medoids algorithm is that although it created seven partitions, the clusters contained anywhere between one and eight data points. The algorithm was able to group some sets of three to four electives into clusters, listed below:

* 19th-CENTURY BRITISH LIT
* EUROPE IN A WIDER WORLD
* EXPERIMENTAL WRITING SEM
* BECOMING HUMAN
* CELL BIOL & BIOCHEM
* DEVIL’S PACT LIT/FILM
* EARLY MESOPOTAM HIST/SOC
* BRITISH POETRY 1660-1914
* COMPARATIVE POLITICS
* CONTEMP ART – 1945 TO PRESENT
* CONTEMPORARY POL. THOUGHT

One aspect we can note immediately after looking at the results of the DBSCAN and the K-medoids algorithm is that there is a consistent grouping between the electives Comparative Politics, Contemp Art – 1945 to Present, and Contemporary Pol. Thought. During our evaluation it is likely that these three courses will compose a concentration area, as well as potentially one more course.

1. K-means

Below is the scatterplot that resulted from the K-means algorithm.

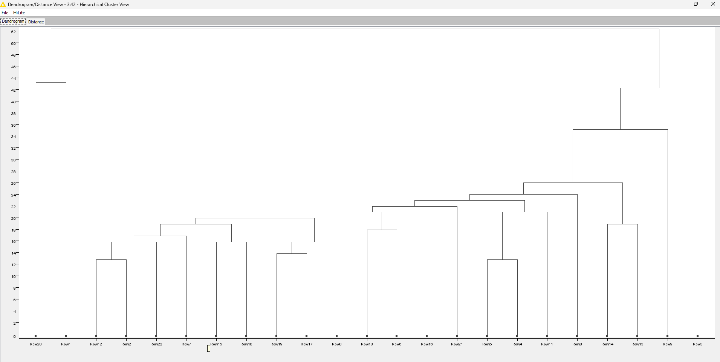


Some notable clusters with three or four electives are listed below:

* 19th-CENTURY BRITISH LIT
* COMMUNICATIONS INTERNSHIP
* CONTEMPORARY POL. THOUGHT
* FRESHWATER ECOLOGY
* AFRICAN AMERICAN LIT: AFRICAN-AMER LIT: CHANGE
* AMERICAN SOUTH 1861 – PRES
* BUSINESS GERMAN: A MICRO PERSPECTIVE
* COMPARATIVE POLITICS
* CELL BIOL & BIOCHEM
* ELEMENTARY ARABIC II
* EXPERIMENTAL WRITING SEM

Upon observation, it was apparent that the resulting clusters were very dissimilar to the ones output by K-medoids and the DBSCAN algorithms. This is likely due to the method in which I assigned numerical values to the courses. I will later analyze how significant these results are, and whether to include them in the decision for determining elective specializations.

1. Hierarchical Clustering



After observing the output, all that was left to do was to use key-value pairs to determine which RowID corresponded to which course name. Using the dendrogram from the Hierarchical clustering output, I created the following clusters of 3 or 4:

* COMPARATIVE POLITICS
* 20TH CENTURY RUSSIAN LITERATURE: FICTION AND REALITY
* FRESHWATER ECOLOGY
* AUGUSTAN CULTRL REVOL
* EXPERIMENTAL WRITING SEM
* ELEMENTARY ARABIC II
* 19TH CENTURY BRITISH LIT
* AMERICAN SOUTH 1861 – PRES
* AMERICAN HEALTH POLICY
* CELL BIOL & BIOCHEM

After using five separate mining techniques, I am ready to begin evaluating our models and deciding which courses to put in concentration areas.

Using Association Rules only generated a few actionable insights for elective concentration; however, these few rules are important for potentially completing specialization areas. Below are the classification rules I will look at when creating the concentration areas:

* FRESHWATER ECOLOGY -> AMERICAN HEALTH POLICY
* CELL BIOL & BIOCHEM -> EXPERIMENTAL WRITING SEM
* COMMUNICATIONS INTERNSHIP -> AMERICAN HEALTH POLICY
* CONTEMPORARY POL. THOUGHT -> AMERICAN HEALTH POLICY

After using string distances to calibrate the DBSCAN algorithm, it produces results that were replicated using the K-medoids algorithm, as well as with Hierarchical Clustering. Therefore, I will be using a combination of these three clustering algorithms with the developed association rules to generate groups of three or four courses.

The cluster output for K-means was very different from the others presented above. It produced results that were not consistent with the association rules and will not be included in the computation for concentration areas.

# Deployment

The initial goal of this project was to reduce the number of credit hours from 39 to 33, define concentration areas for elective courses, and reduce the number of offered electives. Naturally by reducing the number of electives offered, we reduce the number of credit hours, so we will focus on the latter two objectives for deployment.

Objective 1: Define concentration Areas for Elective Courses

Using our association rules and our clustering algorithm results, I will create three groups of three or four electives for specialization. The Board will decide what to name these concentrations:

Group 1:

* EXPERIMENTAL WRITING SEM
* ELEMENTARY ARABIC II
* 19TH CENTURY BRITISH LIT

Group 2:

* AMERICAN SOUTH 1861 – PRES
* AMERICAN HEALTH POLICY
* FRESHWATER ECOLOGY

Group 3:

* AESTHETICS
* COMPARATIVE POLITICS
* CONTEMP ART – 1945 TO PRESENT
* CONTEMPORARY POL. THOUGHT

Objective 2: Reduce the number of offered electives

During the first parts of the Mining phase, we were able to count the number of enrollments in electives, set a threshold for keeping electives or not, and removing electives accordingly. As a result, we were able to remove the following electives from the list of offered courses:

* AMERICAN SOCIAL POLICY
* ANALYZING THE POL WORLD
* ART AND RELIGION
* COMM and THE PRESIDENCY
* CONTEMPORARY SOCIO THEORY
* EARLY BALCAN HIST/SOC
* ELEMENTARY GERMAN 1
* EVIDENCED BASED CRIME AND JUSTICE POLICY
* FRENCH THOUGHT SINCE 1945
* FRENCH THOUGHT TILL 1945

This reduced the number of electives offered from 33 to 23, which leaves room for concentration areas as well as other options for students to enroll in.

1. Discussion of Performance

Overall, output of the association rules as well as the clustering algorithms were able to give actionable insights to the grouping of electives. Through the K-medoids algorithm, DBSCAN algorithm, and Hierarchical clustering algorithm, I was able to make similar clusters of data that proved which courses students seem to be enrolling most in, as well as which courses are being taken together. These results also make sense logically if we look at each group separately. For example, Group 1 contains elective courses that are primarily focused on writing. Group 2 contains courses about America, and Group 3 contains courses about politics and art.

1. Limitations and Encountered Difficulties

The main difficulty encountered during this project was the performance of the K-means algorithm. In order to configure the data in a manner that could be mined, I used a different method in KNIME than the string distances method I was using for DBSCAN and Hierarchical Clustering. I believe since this method was different, the produced results contrasted the other methods so much so that K-means was an obsolete method. In a future instance of this project, I would consider standardizing each distance method used before computation to gain the best results.

Another note to make during this project was in algorithms like DBSCAN and K-medoids, selection of Epsilon and radius was very important because a small shift in value could result in one cluster of all the elective courses. Before possibly changing concentration areas for the next school year, I might re-do this approach meticulously to find the perfect values for each case.

Most of the data used was gathered before 2005. For the best results, in the future I would like to see much more data for enrollments that are closer to the present.

For future iterations of this project as well, I might consider the terms of enrollment into classes as well. For example, there have been no enrollments in AESTHETICS since 2002. I will bring this to the committee’s attention to finalize concentration areas as well as possibly remove this course from the offered electives.

1. Conclusion

This project has successfully removed 10 electives that are experiencing stagnant enrollment by students. It has also used cluster analysis and association rules to find specializations that will align with student interests:

Group 1:

* EXPERIMENTAL WRITING SEM
* ELEMENTARY ARABIC II
* 19TH CENTURY BRITISH LIT

Group 2:

* AMERICAN SOUTH 1861 – PRES
* AMERICAN HEALTH POLICY
* FRESHWATER ECOLOGY

Group 3:

* AESTHETICS
* COMPARATIVE POLITICS
* CONTEMP ART – 1945 TO PRESENT
* CONTEMPORARY POL. THOUGHT

The College of Arts and Sciences were concerned about stagnant enrollment and changes in student demographics and interests, as well as shifts in disciplinary approaches. This redesign of the BFA curriculum will revamp student interest in popular topics, increase enrollment in electives, and create concentration areas that will accommodate for those shifts in disciplinary approaches.

# References

[1] Penn State. (2025) “Partitional Clustering Methods” [PowerPoint Slides] Available: <https://psu.instructure.com/courses/2383421/pages/lesson-9-notes?module_item_id=43064084>

[2] Penn State. (2025) “Agglomerative Hierarchical Clustering (bottom up)” [PowerPoint Slides] Available: <https://psu.instructure.com/courses/2383421/pages/lesson-9-notes?module_item_id=43064084>