Restructuring Elective Curriculum for the Departments of Arts at the University of  
Diploma Printing in Romanigstan

Nick Smailer   
*Student, SWENG 545  
Pennsylvania State University*

World Campus

nrs5434@psu.edu

*Abstract*—This document discusses the combined application of statistical analysis and frequent itemset mining to assist the Arts department in their goal of increasing enrollment by reducing electives offered and determining elective groupings. Statistical analysis was used to determine viable thresholds for course elimination or “group leader” status. Using the resulting dataset, courses were assigned to a “group leader” based on rules generated using the Apriori algorithm.

Keywords—data mining, summary statistics, frequent itemsets, quantile, Apriori algorithm, confidence, lift

# Introduction & Problem Definition

To begin, it is crucial to understand the issue facing the University of Diploma Printing in Romanigstan. In recent years, their Arts department has been plagued with stagnant enrollment. As a result, they want to restructure the curriculum of their Arts department to bring in more students, and in effect, more revenue. To do this, the university has requested a reduction in the offered electives to coincide with a reduction in course hours required for the Arts program. In addition to course reduction, the university has asked for the electives to be grouped together. By lowering the total number of offered electives and grouping them into related subsets, the university hopes to revamp their Arts program and begin an upward trend in enrollment. Now that the problem has been explained and expectations defined, the approach taken to achieve these goals will be summarized, as well as the structure of this document.

The first step towards the solution was data exploration and cleaning. This is an integral step, as it provides key insights about course enrollment across a three year time frame. After cleaning the data, introductory analysis was preformed to gain an understanding of the data distribution and structure. Next, a more in depth statistical analysis was pursued in order to determine the courses to be removed. Finally, courses were grouped together using the Apriori algorithm to generate frequent itemsets. The final result was an elimination of 13 courses to bring the total amount of electives offered to 20 courses based on historical enrollment; and the creation of course groupings that provide structure with flexibility for students. Now that the project goals and outcomes have been summarized, the topics mentioned will be explored in depth.

# Data Exploration and Early Findings

## Introductory Findings

Upon inspecting the data, it became apparent that there were quite a few discrepancies throughout the dataset. Before diving into these issues, for the readers understanding, the contents of the dataset will be explained. The enrollment dataset consists of three fields- Student name, Semester, Course name. Each course enrollment has a unique record, so for example, if a student took ARTS 493 and ARTS 494, these would be two different entries. In addition to the enrollment data, an accurate list of course names was provided to check the data against.

Having the ability to compare the two, right away the course names were inspected. The enrollment dataset had 170 unique course names, while the factual set only contained 42 courses, so there was clearly many errors within the enrollment dataset. After discovering this huge disparity, it became necessary to inspect the student names for inconsistencies as well. Multiple spelling issues were discovered, but the main issue this illuminated was repeated records. Now that a few pressing issues had been identified, the next step was to address them and create a viable and trustworthy dataset.

## Data Cleaning

To start the data cleaning process, right away duplicates were eliminated to ensure each record is unique. Null values were also removed at this time. This dramatically reduced the data size from about 5,000 records to 3,661. The next step was to correct the issues with the course names. This was quite the undertaking considering there were over 100 incorrect names included throughout the dataset. To achieve this, the courses were examined one at a time and compared to those contained in the factual dataset. Additionally, the course number field was added to identify courses in a way aside from just their name. The courses that had no parallel in the list of course names were assigned with a course number of 1, effectively grouping them all together. Inconsistencies with student names were not adjusted considering they are not important to the project goals. Having eliminated the duplicate and null records and adjusted the course names to match the real course names, the data cleaning stage was considered to be complete.

## A table with numbers and symbols Description automatically generatedEarly Takeaways

During data cleaning, three courses were identified to have 0 listed enrollments: ARTS 496, ARTS 543, ARTS 579. These were eliminated before any further analysis, because with 0 enrollments, clearly there is no need to retain them or include them in analysis. Next, the dataset was filtered to only include the remaining electives. After that, a count of total enrollment for the remaining electives was generated to identify both high and low enrollment courses. This showed the highest enrollment came from ARTS 493 which had 143 enrollments, and the lowest came from ARTS 585 which only had 1 enrollment. Illuminating both the high and low enrollment electives allowed for early consideration of remaining courses that should be eliminated and courses that absolutely must be retained.

# Statistics and Resulting Decisions

## Summmary Statistics

After exploring the data and gaining an understanding of the values within, analysis began. To begin the analysis, summary statistics were generated. The total number of enrollments over the past three years was 855- not 855 students, but 855 elective enrollments. Average elective enrollment was 29.483, and the standard deviation of the dataset was 32.463. This shows that the data is quite spread out, meaning that elective enrollment has large fluctuations. This was already seen previously when the minimum and maximum enrollments were obtained. Next, the 25%, 50%, and 75% quantile values were 10, 16, and 50 respectively. The 25% quantile value seemed quite low, so this was inspected and determined to be accurate. Having generated and examined summary statistics, it was time to make some decisions based on them.

A close up of a number

Description automatically generatedA screenshot of a computer

Description automatically generated

## Outcomes

After examining the summary statistics and comparing some of the results to the figures obtained during exploratory analysis, a few important decisions were made. These choices were made largely based on the quantile values and what they say about certain courses. The first major decision was which courses to remove.

One of the main goals of this project is to reduce the number of offered electives based on a history of low enrollment. Initially, there were over 30 elective courses. As mentioned above, some had very low interest with one course having only a single enrollment. Considering how low the 25% quantile value was, this value was deemed as a logical threshold for course elimination. As a result, all courses at or below the 25% quantile value of 10 were eliminated to reduce the number of electives offered. The courses that were eliminated are as follows: ARTS 485, ARTS 495, ARTS 543, ARTS 553, ARTS 561, ARTS 567, ARTS 573, ARTS 575, ARTS 585. The next major decision made was which course will be “group leaders”.

Group leaders in the context of this project means courses that the course pairings or groupings will be based on. Each group leader will have their own unique group of courses assigned to it. The group leaders are the most popular courses, and the courses in their respective groups will be assigned based on frequent itemsets- this will be discussed in depth later on. While the removed courses were decided using the 25% quantile value, the group leaders were decided using the 75% quantile value. The 75% quantile value was 50, and there were only four courses above that threshold. The courses in order of enrollment from greatest to least are as follows: ARTS 493, ARTS 587, ARTS 569, and ARTS 555 respectively. The lowest enrollment out of those four was 64, well above the 75% quantile value mentioned previously. In addition to the statistical reasoning, another benefit of having four groups instead of another number is the flexibility. Having four sets of five allows students to select a group they are interested in without being bound to every single course. This allows students to “customize” the program in a sense to fit their interests which I believe will cultivate interest, as well. Having determined which courses to remove and which will be group leaders, it was time to begin determining course groupings.

# Model Generation

## Initial Attempts

Before beginning the discussion on this section, it is worth noting that all model generation and refinement was done using the analytical platform Knime. To begin model generation, the records were rolled up by student. Instead of a unique record for each student’s specific course enrollment, each student had a unique record containing all the courses they had enrolled in. From there, the Apriori algorithm was used to generate frequent itemsets based on the new dataset. The Apriori algorithm determines association rules based on how many times items appear together (itemset) compared to how many times they appear individually. The Apriori algorithm was chosen over options like FP-tree mainly out of user preference, but this is also a popular way to generate association rules. A common application of this technique is purchase history to determine items that are frequently bought together, but it applies itself well to this project.

The initial parameters used were, minimum set size of three, support of 4, and confidence of 20%. This provided a picture of the most frequent item pairings, but did not prove useful for the purposes of the project. The results showed that the four courses determined as group leaders were very frequently selected together. For example, lots of the itemsets with high appearance, confidence and lift had group leaders as both antecedents and consequents. Obviously, this is not very helpful because it would be counterintuitive to group the most popular courses together. As a result, the parameters needed adjustment to allow for more useful findings.

## Improvements

While the initial findings were interesting, for the most part they were obvious and not too useful. In an effort to create more useful itemsets, the thresholds were lowered. The minimum set size was changed to 2, support was changed to 3, and confidence was changed to 3%. While these parameters are all quite low, it ensured that all of the remaining electives would be included in the groupings. These adjustments yielded a more useful result, but still needed modifications for interpretation. The result table was split into four groups, one for each group leader. The groups were created by filtering the antecedent column to only include one group leader per table. Splitting the result into smaller groups based on group leaders provided a much more readable and actionable output. Now that the parameters had been adjusted to include all courses and the results had been filtered, decisions could start being made about the course groupings.

# Course Assignment

When assigning the remaining electives to group leaders, multiple different factors were considered. They will now be explained:

* Itemset appearance- How many times were both courses selected by one student? This is crucial to take with a grain of salt however, because it hinges upon how often the group leader was selected.
* Confidence- How often were both electives selected by the same student out of ONLY the records including the group leader?
* Lift- How often do the electives appear together compared to how often they were EXPECTED to appear together?

Lift was given the highest priority out of the different factors. A rule with a lift greater than 1 is considered actionable, so this was a focus in creating the groups. In situations where the itemset appearance, confidence, and lift were extremely close, the group leader with the lower total enrollment was given the elective in question. This was decided in an effort to make the course groups more enticing. For example, ARTS 493 had the highest enrollment by a large margin. This implies that no matter what, students are going to enroll in this course. As a result, if the elective in question has a very similar itemset appearance, confidence, and lift to a group leader with a lesser enrollment, the elective is assigned to the group leader with lower enrollment. Now, the decisions regarding group assignments will be outlined, going in order from group leader with most to least enrollments.

## ARTS 493

ARTS 493 was a cut above the rest. Far and away the most popular elective, and as a result, there were quite a few other electives that frequently appeared alongside it in itemsets. With ARTS 493 as the antecedent, consequents of ARTS 491, ARTS 547, ARTS 494, ARTS 497, and ARTS 488 all had a lift greater than one, meaning they are actionable. That being said, only four of the five listed could be selected. Upon comparing the support, confidence, and lift to the other three groups, it was clear ARTS 488, ARTS 494, and ARTS 497 belonged in this group. ARTS 547 had very similar values to another group leader (ARTS 555) and due to ARTS 493 having the higher enrollment, ARTS 547 was assigned to the other group. Based on the values, ARTS 491 fit the best with this group, however, ARTS 557 was selected instead. While ARTS 557 did not align well with this group, it did not align well with any group. Since ARTS 493 had by far the highest enrollment, it was decided that this would be the best placement to “hide” ARTS 557. ARTS 557 did not have low enrollment, so it would not have made sense to remove it, it just did not fit well into any group. Ultimately, ARTS 488, ARTS 494, ARTS 497, and ARTS 557 were assigned to this group.

## ARTS 569

ARTS 569 had the second highest enrollment at 103. This group was difficult to assign, because many of the electives with a lift over 1 had higher lifts in other groups or had been assigned to ARTS 493. ARTS 581, ARTS 494, ARTS 551, and ARTS 488 had a lift greater than 1 in relation to ARTS 569. As mentioned above, ARTS 488 and ARTS 494 had already been assigned, and had a higher lift in relation to ARTS 493. Additionally, ARTS 551 had a higher lift in relation to ARTS 587, so initially, it was also going to be placed in a different group. Upon further inspection ARTS 587 had enough options with lift greater than 1 so cede ARTS 551 to ARTS 569. All things considered, this is not the strongest group based on support, confidence, and lift, but only because courses fit better elsewhere. In the end, ARTS 491, ARTS 551, ARTS 565, and ARTS 581were assigned to this group. These courses had lifts of 0.771, 1.05, 0.9, and 1.369 respectively.

A screenshot of a computer

Description automatically generated

## ARTS 587

Selecting the course to group with ARTS 587 proved to be the easiest out of the four group leaders. ARTS 492, ARTS 547, ARTS 581, ARTS 559, ARTS 484, ARTS 583, ARTS 551, and ARTS 571 all had a lift greater than one in relation to ARTS 587. Considering the large number of courses with a strong relationship to ARTS 587, this grouping was decided last. This allowed group leaders with fewer strong relationships to retain the courses with lifts greater than 1, even if said course had a higher lift in relation to ARTS 587. This allows for more balanced groupings on the whole, and a fairer structure. The courses assigned to this group are as follows: ARTS 492, ARTS 559, ARTS 571, and ARTS 583.

## ARTS 555

To decide the courses grouped under ARTS 555, the same procedure was followed as the other groups. First, courses with a lift greater than 1 were identified. These courses were ARTS 484, ARTS 488, ARTS 547, ARTS 549, and ARTS 571. ARTS 571 was included in the ARTS 587 group because it had a higher lift in relation to ARTS 587. The same was true for ARTS 488 in relation to ARTS 493. As a result, the remaining three courses with a lift greater than 1 were assigned to this group. The final course assigned was ARTS 545. Similar to ARTS 557, this course did not have a lift greater than 1 in relation to any of the group leaders, so it was placed in this group to ensure the strength of the groupings on the whole. The course assigned to this group are as follows: ARTS 484, ARTS 545, ARTS 547, and ARTS 549.

# Conclusions & Final Thoughts

In conclusion, the predefined requirements for this project were satisfied using a variety of means. First, the data was thoroughly inspected and cleaned to ensure a consistent usable format. Next, statistics were generated which were the basis for key decisions regarding which courses to eliminate. The statistics generated were also crucial in determining which courses should serve as group leaders. Finally, after a few iterations of improvement and refinement, the Apriori algorithm was used to generate frequent itemsets and association rules. These rules provided the basis for the elective group assignments. The main considerations for group assignment were lift, confidence, and itemset appearance, but in situations where those values were extremely close, group leader enrollment was used as the deciding factor.

The final result is as follows:

1. ARTS 493: ARTS 488, ARTS 494, ARTS 497, ARTS 557
2. ARTS 569: ARTS 491, ARTS 565, ARTS 551, ARTS 581
3. ARTS 587: ARTS 492, ARTS 559, ARTS 571, ARTS 583
4. ARTS 555: ARTS 484, ARTS 545, ARTS 547, ARTS 549

Ultimately, within the four groups of five, only four rules had lifts less than 1, and one of those had a lift of 0.9. Ensuring each retained course was included and went to a group that made sense was like completing an intricate puzzle, so all things considered the result is very actionable. In addition to the hard evidence supporting the decision making process and analysis discussed in this document, there is one more “hidden” benefit. While this is a personal thought, I believe it would make the program more desirable for students. The reduction in credit hours means a reduction in the number of electives students are required to take. They are now required to take 12 credits of electives, or four courses. As a result, it may have seemed more intuitive to create five groups of four where every course has to be taken. I argue the contrary- by allowing students the flexibility to select three of the four electives under each group leader, they have freedom to pursue what they deem interesting out of the groups. I believe this customization will be an enticing feature that does not necessarily get captured or represented by analysis because it is more human. On the whole, if these modifications are implemented, the Arts department should see an uptick in enrollment, and maintain steady enrollment in the offered electives.

##### References

1. W. Bai, “Lesson 5.3 Association Rule Mining in Knime”, SWENG 545
2. W. Bai, “Lesson 6.2: Algorithm for Associative Classification”, SWENG 545
3. W. Bai, “Lesson 6.4: Example of Associative Classification in Knime”, SWENG 545

Note: In terms of the project problem of grouping courses in 3 or 4, I did this by creating sets of 4 under the “group leader”. I realize this may be a bit of a strange interpretation, but I believe my reasoning using quantile values explains why I went this route.