Data Mining Honors Project Final Report

## Motivation and objectives

The project compares UT Arlington professors based on their outcomes in the Student Feedback Surveys (SFS) conducted at the end of each semester and their ratings on the Professor Rating Site RateMyProfessors.com, winner of two People's Voice Webby Awards for 2015. Based on the data in both datasets, UT Arlington professors are classified into five groups in for each dataset (a total of ten groupings) and the results from both datasets are evaluated and compared to give a full picture of the efficacy of UT Arlington Professors.

## Data Mining/Analysis Tasks Tackled

The chief data mining task was clustering performed using a k means clustering algorithm, but other tasks were performed as follows:

* 1. **Downloading appropriate webpages from the websites**
  2. Download the Rate My Professor webpages for UT Arlington professors: The Rate My Professor data can be retrieved from the website, ratemyprofessors.com. Each professor is rated on a different webpage, therefore, the rating information for each professor has to be retrieved individually. It is a tedious process that is accomplished by saving each individual webpage. Some UT Arlington professors are not on the Rate My Professors website and others are present but have zero ratings. Those with zero ratings are ignored.
  3. Download the UTA SFS webpages: The SFS dataset can be downloaded from the UTA website at <http://www.uta.edu/ier/Surveys/sfs/results-yes.php>. The page gives a brief description of the data and contains links to the data at the bottom of the page, arranged by semester. The data is stored in tables on a webpage. To get it, you must save the webpage onto your computer. One link is wrong. The link to download the data for summer 2011 professors with last names starting from L to Z links to the wrong page. Repeated attempts to fix this were made. Calls were placed to the UTA department responsible for maintaining access to the data, but the only response was a high pitched noise on the phone. In any case, it is unlikely that the problem would have been fixed in time to be of use to the project. Consequently, only the information that could be obtained was used.
  4. **Mining the web pages downloaded for rating information**
     + - 1. Converting the tables from the UTA SFS webpage into an internal data structure:

**The data were transferred to Microsoft Excel Files**. The first attempts to make use of the UTS SFS dataset were done by use of a java library called Jsoup. The jar file for this library can be obtained at Jsoup.org. However, a problem was encountered. All the webpages obtained from the UTA website shared similarities, but they did not have the exact same structure. This made it impossible to use one simple method to parse all the webpages. Rather than write a separate method (or use if statements) to parse each webpage separately, a more elegant solution was sought.

The problem being that the webpages were so different, the solution was to make the structure of the webpages similar enough that they could be parsed together. Since all the data was in the form of tables, it was decided that the Microsoft Excel program should be used. All the webpages were copied into Microsoft Excel and edited into one format as follows:

* + 1. **All but one header rows were deleted**. The webpages were widely inconsistent in that some pages had up to four header rows and others had only one or two. This makes parsing the data complicated because the program has to determine how many header rows there are.
    2. **All whitespace in the data was deleted**. While this might seem drastic, it was an attempt to fix another inconsistency in the data. Some pages had whitespace in the names of the professors while others did not. This made it difficult to accurately compare two names from different webpages to determine if they are the same name. It can be undertaken in the code to detect and delete whitespace in the name field, but the choice was made to simply delete the whitespace in the excel files and make the code simpler. This choice to delete the whitespace did not affect any information but the headers (which were ignored anyway) and the names (which were the target of the strategy).
    3. **Extra columns were deleted.** Another inconsistency in the data was that before Fall 2013, all the SFS data had six columns corresponding to the six questions asked in the survey. From Fall 2013, however, the sixth column containing the statement, “This instructor is one whom I would recommend to other students” is missing. Therefore, this column was deleted from the other data, or simply ignored.

Some webpages also had extra empty columns before the data. Those were deleted so that the first column contained the names of the professors.

* + 1. **Columns in the wrong order were re-arranged**. Starting in Fall 2013, the questions asked in the Student Feedback Survey were reworded. The order of the questions was also changed. The files containing data from Fall 2013 onward were rearranged. None of the data was changed but the positions of the columns were modified to match those of the earlier files.
    2. **The data was stored as a tab separated text file**. Initially, the modified excel files were stored as CSV (comma separated values) files, but this gave rise to confusion in reading the files. The names of the professors are written in the form LastName,FirstName. When saving the data in a CSV file, therefore, the extra comma is interpreted as a column separator. Saving the files at text solved the problem.

1. Convert the ratemyprofessor webpage information into an internal data structure
2. Converting the collected data into classes
3. Plotting the results of the clustering on a graph and producing tables showing the results.

## 3. Datasets Collected and Used

* + - 1. UT Arlington Student Feedback Survey

The UT Arlington Student Feedback Survey enjoys feedback from thousands of classes over five years and fourteen semesters. It calculates the average rating of each professor over their different classes and through the different semesters. This enables it to account for differences in how a professor performs in different classes and for different subjects.

The Student Feedback Survey asks students to score their teachers on a scale of 1 to 5 on six different outcomes as follows (The last outcome was not considered in this project):

     1. The instructor provided clearly defined expectations (CDE).

     2. The instructor used teaching methods that facilitated my learning (GTM).

     3. The instructor encouraged me to take a role in my own learning, to ask questions, and to participate (SP).

     4. The instructor was well prepared to teach (WP).

     5. The instructor was available outside of class either electronically or in person (Available).

* + - 1. Rate My Professor.com data for UT Arlington Professors

The Rate My Professor dataset rivals the UT Arlington SFS dataset in the number of ratings per professor but has a smaller number of UT Arlington professors. Still, it rates professors on different criteria than the UT Arlington Student Feedback Survey and therefore has its merits.

The RateMyProfessors dataset collects information about teachers in five categories, but only the first four were used in this analysis. The categories are:

1. Overall Quality

2. Helpfulness

3. Clarity

4. Easiness

5. Average Grade

It is different from the UT Arlington SFS dataset in that it does not group the ratings by class, semester, or year. Consequently, it does not reflect changes in professors over time or consider differences in how the same professor performs in different classes.

## 4. Design of Methods

All methods are part of one of four classes:

1. The Cluster class, which is the main class, containing the implementation of the clustering algorithm and controlling the flow of the program.
2. The Professor class, which stores details about the Professors collected from the dataset.
3. The Rating class, which stores information about a Professor’s rating from the RateMyProfessor dataset.
4. The Point class, which stores information about a Professor’s rating from the Student Feedback Survey dataset.

## 5. Implementation of methods

There are five major methods, all belonging to the Cluster class. They are parseRMP(), parseSFS(), clusterRMP(), clusterSFS(), and compare(). clusterRMP() and clusterSFS() implement a k means clustering algorithm to cluster the data into five clusters and evaluate the results. parseSFS() and parseRMP() are supporting methods which parse the data from the SFS and RMP datasets respectively. The compare() method compares the ratings of professors in both clusters. Specifically, it checks to see if professors who are rated highly or poorly in either dataset are also rated highly or poorly in the the other dataset.

All other methods in the four classes merely support the functions of these methods.

## 6. Results and Evaluation

**6.1. The Student Feedback Survey Results**

The professors were divided into five clusters as show in the following table:

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Clearly Defined Expectations (CDE) | Good Teaching Methods (GTM) | Student Participation (SP) | Well Prepared (WP) | Available | Number of Professors | Average Score | Percentage of Professors |
| Cluster 1 | 4.6512 | 4.6214 | 4.6714 | 4.7108 | 4.6684 | 798 | 4.6646 | 24.72 |
| Cluster 2 | 4.3677 | 4.2722 | 4.3802 | 4.4603 | 4.4123 | 1122 | 4.3785 | 34.76 |
| Cluster 3 | 4.0352 | 3.8745 | 4.1027 | 4.1515 | 4.1598 | 783 | 4.0647 | 24.26 |
| Cluster 4 | 3.6252 | 3.3972 | 3.7544 | 3.7775 | 3.8629 | 411 | 3.6834 | 12.73 |
| Cluster 5 | 2.7765 | 2.5721 | 3.1925 | 2.9156 | 3.4290 | 114 | 2.9771 | 3.53 |

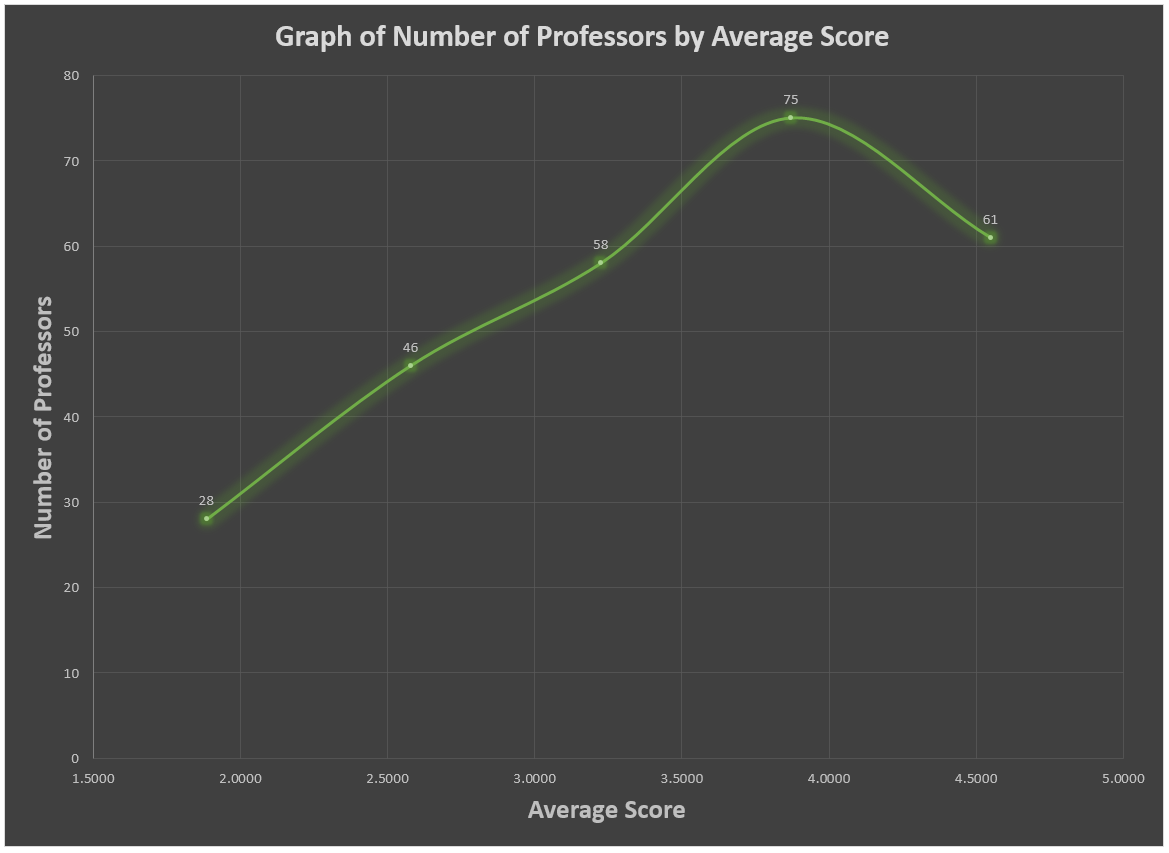
* The most professors belong to cluster 2, with an average score of 4.3785 out of five (87%). This is equivalent to a B+ letter grade.
* Together, the two largest and highest scoring clusters contain more than half (59%) of UTA Professors
* The smallest and lowest scoring cluster has an average score of 2.9771 out of 5 and contains only 3.53% of UTA Professors
* Therefore, according to the SFS dataset, UTA professors score very well on the polled criteria.

6.2. The Rate My Professor Dataset Results

The professors were divided into five clusters as shown in the following table:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Cluster | Quality | Helpfulness | Clarity | Easiness | Average Score | Number of Professors | Percentage |
| 1 | 4.7410 | 4.7623 | 4.7230 | 3.9721 | 4.5496 | 61 | 22.76% |
| 2 | 4.1160 | 4.1507 | 4.0653 | 3.1560 | 3.8720 | 75 | 27.99% |
| 3 | 3.3569 | 3.3603 | 3.3586 | 2.8276 | 3.2259 | 58 | 21.64% |
| 4 | 2.7065 | 2.6022 | 2.7957 | 2.2130 | 2.5793 | 46 | 17.16% |
| 5 | 1.8714 | 1.9071 | 1.8321 | 1.9321 | 1.8857 | 28 | 10.45% |

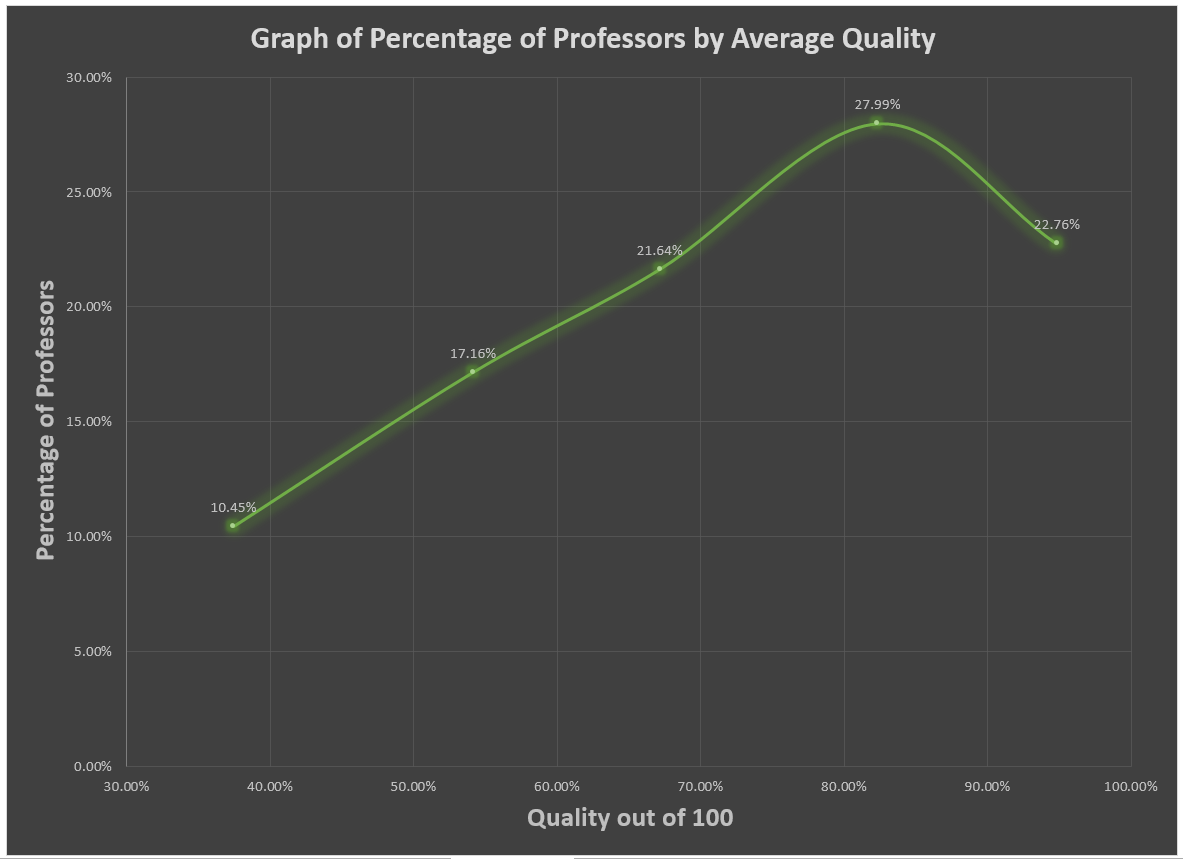
The following graphs helps visualize the data:

[](https://sites.google.com/site/tracyogunihonr/rate-my-professor/cluster%20size%20by%20average%20RMP.PNG?attredirects=0)

As can be seen from the graph:

* The highest scoring cluster with 61 professors, has an average score of 4.55 out of 5, corresponding to a percentage score of 91%.
* The next scoring cluster has the most professors (75) and an average score of 3.87 out of 5 (77%).
* In other words, the best professors have a letter grade of A- and the next best professors have a letter grade of C+. This means that only the first cluster (about 23% of UTA professors) score higher than a C.
* The lowest scoring cluster, with 28 professors has an average score of 1.88 out of 5 (37.7%). That corresponds to an F letter grade.
* The next lowest scoring cluster, with 46 professors, has an average score of 2.58 out of 5 (51.6%). That also corresponds to an F letter grade.
* Taken together, both clusters contain 27.6% of all UT Arlington professors. According to the rating, therefore, more than a quarter of UT Arlington professors score about an F.
* The middle cluster, with an average score of 3.23 out of 5 (64.52%) consists of 58 professors. The lowest three clusters taken together amount to letter grades of D or less and contain about 49% of UT Arlington Professors.

An analysis of the average quality of the professors reveals a slightly better but still unsatisfactory picture:

[](https://sites.google.com/site/tracyogunihonr/rate-my-professor/average%20quality%20percentage%20graph%20RMP.PNG?attredirects=0)

As can be seen from the graph:

* More than a quarter of UT Arlington Professors have an F letter grade.
* About 50% of the professors score a B or higher
* The remaining professors score around a D.
* The data shows a satisfactorily high number of A's and B's, but also a high number of F's.

6.3. Comparison of Both Results

The Comparison of the SFS and RMP datasets is done in two stages. The first stage compares the general state of UTA Professors as presented by both datasets. The second stage considers the degree of agreement between both datasets.

**Stage 1**

There is a wide disparity between the picture of UTA Professors presented by the Student Feedback Survey and the Rate My Professors website.

* In the SFS dataset, only 3% of UTA professors get an F letter grade, a sharp difference from 27.6% in the RMP dataset.
* According to the SFS dataset, 59% of UT Arlington professors score a B+ or higher. The RMP datasets has the number of professors scoring higher than a C at 22.76%
* When only the 'Quality' rating in the RMP dataset is considered, the number of professors scoring a B or higher rises to 50.75% of the considered professors.

The explanation for this disparity seems obvious.

* Selection Bias: While the SFS dataset strives to collect data about every professor, the RMP dataset depends on students to take initiative and rate their professors. For that reason, the dataset likely contains information mostly about the very good or very bad professors.
* Questions and Question Quality: The SFS dataset contains answers to well defined questions, while the RMP website simply asks students to rate the "Quality" and "Clarity" of a professor. Such vague questions are likely to give rise to a different kind of answer than the questions used for the SFS dataset.

**Stage 2**

To judge the degree of agreement between both dataset, one must simply check how a professor rated high (or low) in one dataset is rated in the other. To do this, only professors present in both datasets were considered.

* Of the Fifty-eight professors in the highest RMP cluster (who are also in the SFS dataset), Fifty-five are in the highest two SFS clusters, a very high degree of agreement.
* Of the Forty-three professors in the highest SFS cluster (who are also in the RMP dataset), forty are in the highest two RMP clusters.
* Of the twenty-five professors in the lowest RMP cluster (who are also in the SFS dataset), thirteen are in the lowest two SFS clusters.
* Of the two professors in the lowest SFS cluster (who are also in the RMP dataset), one is in the lowest two RMP clusters.

The most that can be concluded, therefore, is that professors who are rated highly in one dataset, are also rated highly in the other. However, professors who are rated badly in one dataset are not necessarily rated badly in the other. This makes sense because both datasets ask different questions, but the issue bears further investigation.

## Presentation/Visualization of the Outcome

The presentation and visualization of the results can be found on the website presenting the findings.

## URL of your project website

The website for the project can be found at: https://sites.google.com/site/tracyogunihonr/home