Dimensionality Reduction for Meta-Data from Online Reviews

Eliana Espinosa, Celeste Pereira, Reinaldo Sanchez-Arias, Ph.D

St. Thomas University, School of Science, Miami Gardens, FL

Introduction

Miami Dade College (MDC) has over thousands of professors and around 90% has received plenty of reviews in the site RateMyProfessors.com (RMP). Assisted by the programming tool RStudio, collecting data and making comparisons for each instructor that has taught a mathematics and/or statistics course at MDC we can identify the correlation between why students tend to leave comments or assign specific tags to professors. Two main mathematical methods were used in this project. Principal Component Analysis (PCA), a powerful mathematical method that identifies the directions of most variation in a dataset, and k-means, and algorithm for clustering analysis that finds structure and groups based on numerical features. Dimensionality reduction is a plus to the k-means clustering method since it reduces the number of random variables to those that better describe the data. PCA and clustering have a slight difference, for one PCA tend to find a low-dimensional representation of the observations that can classify the variation in the data; as for the clustering, it finds homogenous subgroups among the observation. This exploratory study aims to show how to use meta-data associated to online reviews in finding patterns and structure in the type of comments left on the review site RMP.

Data Collection

Our case study for data analysis of online reviews involved web scraping of comments left on RateMyProfessors.com (RMP), a review site that allows college and university students to assign ratings to professors and campuses of American, Canadian, and United Kingdom institutions. Users have added more than 19 million ratings, 1.7 million professors and over 7,500 schools.

We focused our attention on the reviews left to instructors from MDC that have taught mathematics and statistics courses. We scraped the data for a total of 559 instructors, and collected the information for the *tags* included in the review. As of date, RMP allows the user leaving a comment to add up to 3 different tags when reviewing a professor. There are a total of 20 different choices:

Respected	*Hilarious*	*Caring*	*Can't skip class*
Pop quizzes	*Clear grading*	*Extra credit*	*Lots of homework*
Lots of readings	*Good feedback*	*Accessible*	*Graded few things*
Group projects	*Lecture heavy*	*Many papers*	*Participation matters*
Amazing lectures	*Tough grader*	*Inspirational*	*Test heavy*

The data collected also includes average answers to the *"would take again?"*, *"difficulty"* and *"quality""* ratings left by reviewers.

The data used was cleaned and anonymized to not include the name of the professors. We confirmed that the review left was indeed for a mathematics or statistics course in order to avoid cases in which a user leaves a review for a faculty listed in an incorrect department.

Data Science with RStudio

Programming was essential in the development of this project. R [1] is an open source language widely used in the data science community, with focus on statistical data analysis, data visualization and machine learning methods. During this project, tools from the `tidyverse` package [2] were used for data wrangling and data visualization with the help of RStudio, an open source integrated development environment (IDE) for R.

In particular we used the `ggplot2` package for *data visualization*, `dplyr` and `stringr` for *data transformation* and summaries, and `rvest` for *web scraping*. Additionally we used tools from the `tidytext` package [3] for text processing and encoding.

R Studio ggplot2 dplyr stringr

Fig 1: RStudio logo. RStudio makes R easier to use. It includes a code editor, debugging and visualization tools. Logos for the ggplot2, dplyr, and stringr packages.

Principal Component Analysis and Clustering

Exploratory Data Analysis

We performed basic exploratory data analysis (EDA) to better understand the distribution the tags used by different users, as well as to study the other variables we were able to extract from the RMP website.

Times tags have been used

Tag used	Times used	Tag used	Times used
Caring	1273	Inspirational	468
Cant skip class	1203	Participation matters	398
Lots homework	1117	Accessible	353
Tough grader	1064	Lectures heavy	316
Respected	964	Test heavy	307
Clear grading	895	Graded few things	153
Good feedback	823	Pop quizzes	92
Extra credit	744	Lots readings	89
Amazing lectures	610	Group projects	38
Hilarious	526	Many papers	38

Fig 2: Frequency of tags used in RMP as part of a review of a mathematics/statistics course.

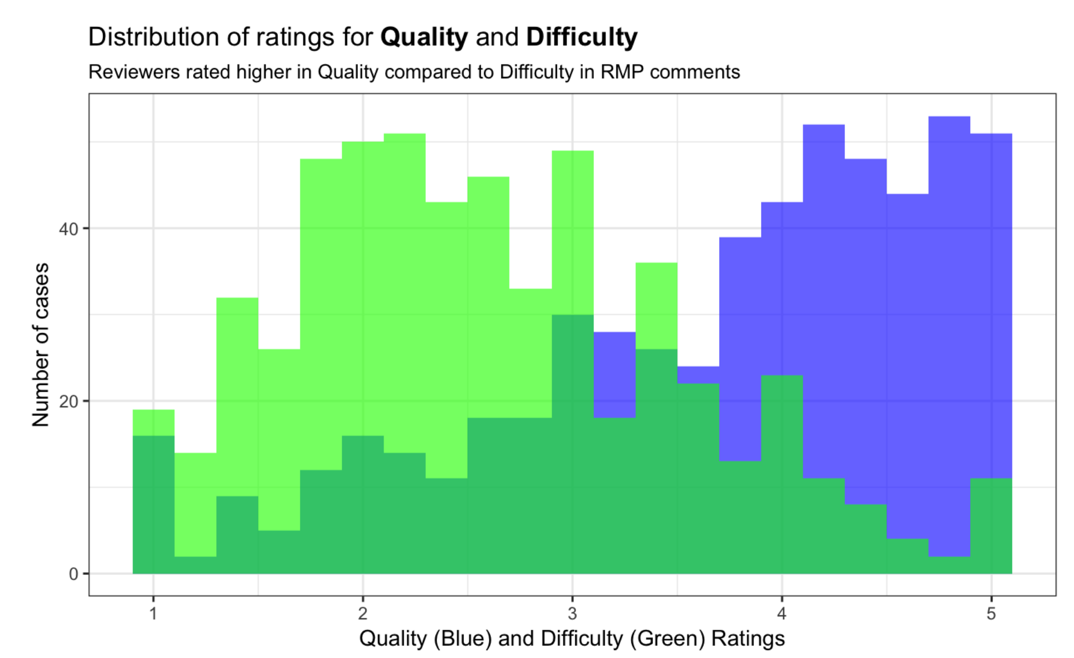

Fig 3: Histogram showing the distribution of ratings for "Quality" and "Difficulty" of mathematics and statistics courses.

Principal Component Analysis (PCA)

The method of Principal Component Analysis (PCA) is used to reduce the dimensions of a data frame finding the directions of the most variations [5]. Below we show the relative importance of the different tags in the first 6 principal components (the first six components capture 88.98% of the total variation in the data we considered).

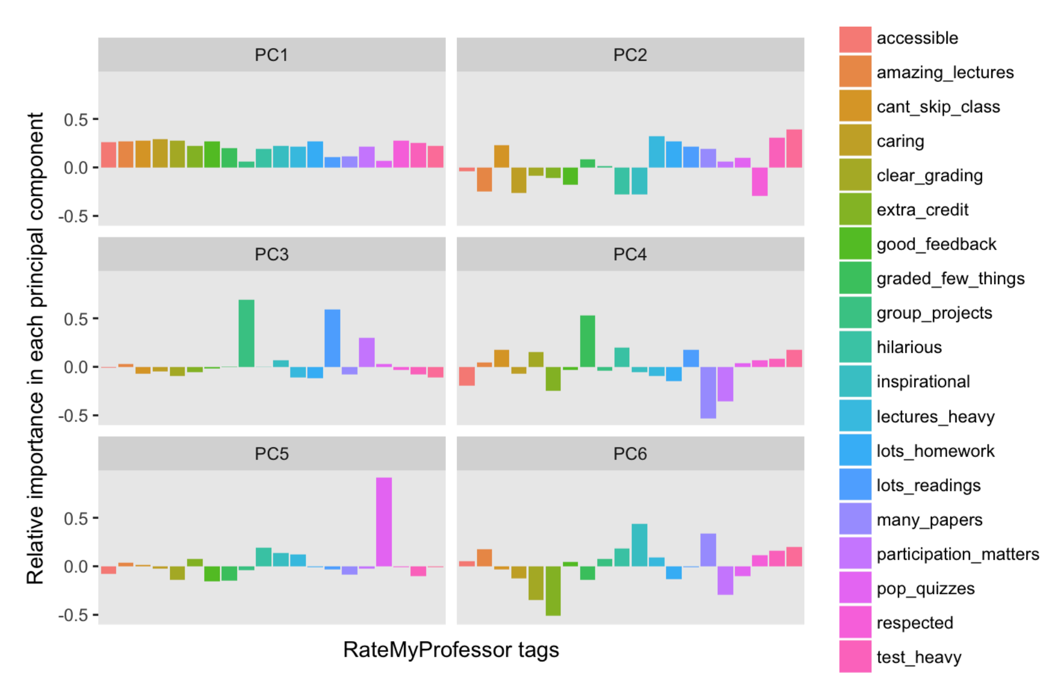

Fig 4: Contribution to the first six principal components. Notices how in components 3 and 4, very specific tags contribute the most to the direction of largest variance

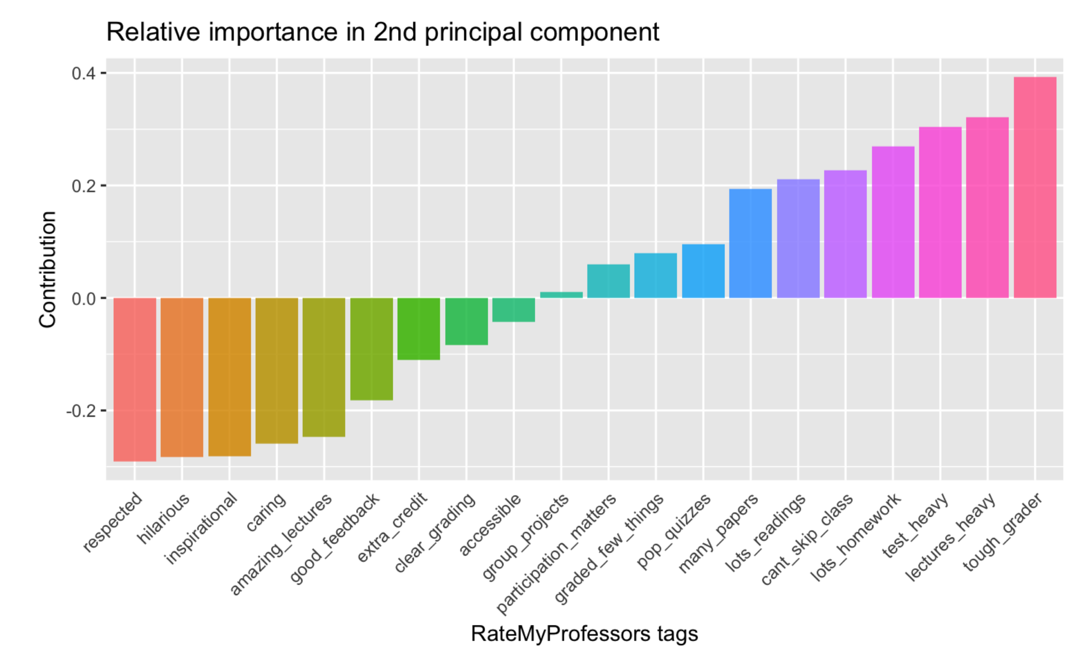

Fig 5: Contribution of each tag to the second principal component. Notice that the this principal component stretches from tags associated to the instructor to tags associated to the grading structure of the course.

PCA allows us to think and reason about high dimensional data. Part of that is projecting many dimensions down onto a more plottable two dimensions.

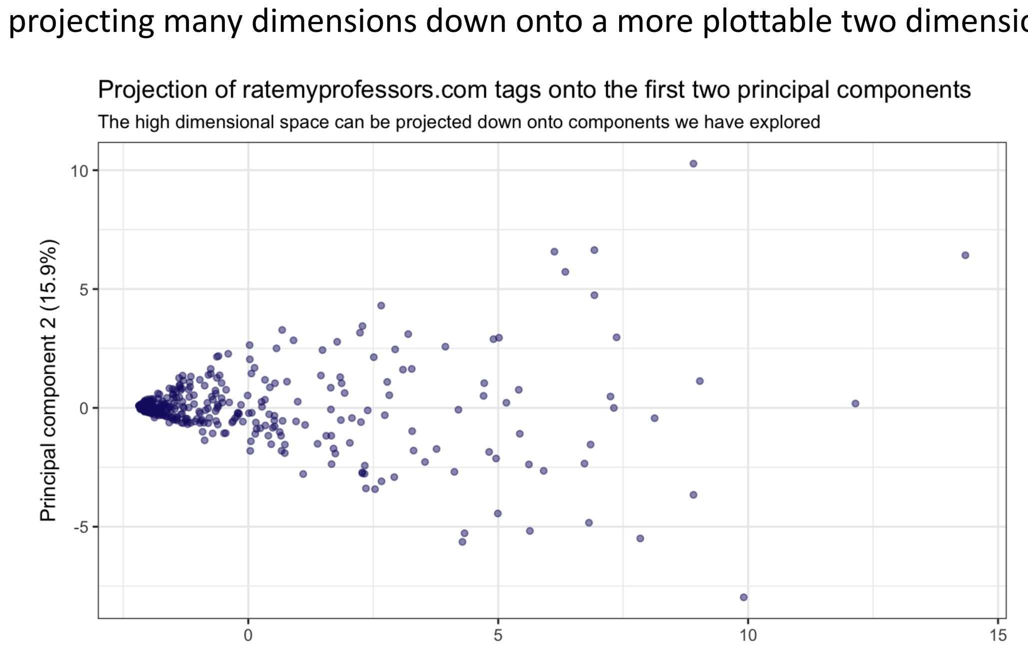

Fig 6: Projection on principal component space. Every point represents one of the 351 mathematics/statistics instructors that had at least one tag in their reviews on RMP

Clustering

K-means [6] is used to find relationships between the observation and create clusters. The method creates clusters so that the total intra-cluster variation is minimized.

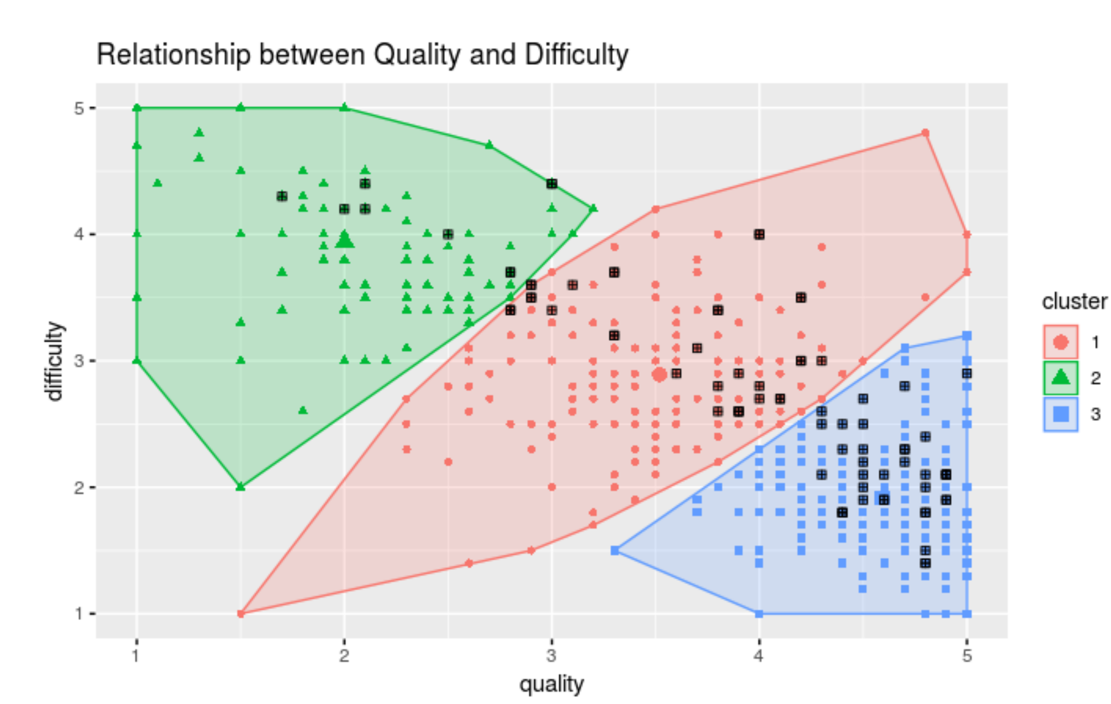

Fig 7: Grouping showing the distribution of ratings for "Quality" and "Difficulty" of mathematics and statistics courses, highlighting the points for which PC1 > 3, PC2> 2.5 and PC2 < -2.5 in Figure 6

Elbow Method

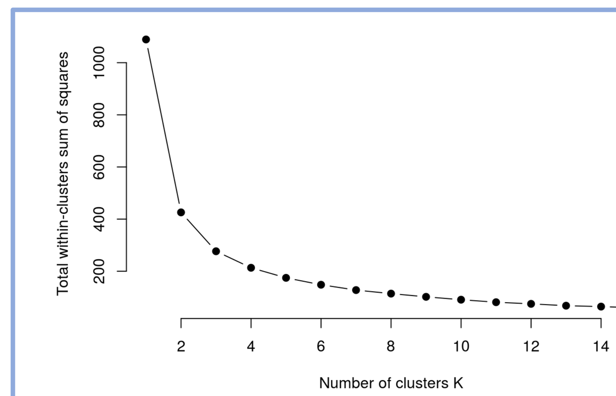

Figure 8. "Elbow method" to determine the number of cluster to consider (i.e. value of k). The total within-cluster sum of square measures the compactness of the clustering and we want it to be as small as possible.

Fig 9: R code snippet showing the implementation of the "elbow method" to find a good value of k in k-means clustering.

1. Compute clustering algorithm for different values of k. For instance, by varying k from 1 to 15 clusters
2. For each k, calculate the total within-cluster sum of square (wws)
3. Plot the curve of wws according to the number of clusters k.
4. The location of a bend (elbow) in the plot is generally considered as an indicator of the appropriate number of clusters.

````
```{r}
set.seed(123)

# function to compute total within-cluster sum of squares
wss <- function(k) {
  kmeans(read_comments[ ,c("quality", "difficulty")],
         k, nstart = 10 )$tot.withinss
}

# Compute and plot wss for k = 1 to k = 15
k.values <- 1:15

# extract wss for 2-15 clusters
wss_values <- map_dbl(k.values, wss)

plot(k.values, wss_values,
       type= "b", pch = 19, frame = FALSE,
       xlab= "Number of clusters K",
       ylab= "Total within-clusters sum of squares")
```
````

Conclusions

Based on the data collected from the "Rate my Professors" website, structure and patterns were found for MDC mathematics and statistics instructors, relating the ratings for difficulty and quality, as well as the meta-data associated to every online review left on the RMP site. Results show how professors are rated from students based on experiences and demonstrates that when professors have a *"very difficult"* rating they tend have a low quality rating, and vice versa for having low difficulty to having high quality rating. Dimensionality reduction techniques such as PCA can be used to better understand the directions of most variation in the data, and transform the high dimensional space (e.g. meta-data on online review) to a space where patterns can be found (principal component system)

References

[1] R: free software environment for statistical computing and graphics. https://www.r-project.org/

[2] `tidyverse`: an opinionated collection of R packages designed for data science. https://www.tidyverse.org/

[3] `tidytext`: Text Mining using 'dplyr', 'ggplot2', and Other Tidy Tools. https://CRAN.R-project.org/package=tidytext

[4] RMP: Rate My Professors http://www.ratemyprofessors.com/About.jsp

[5] Markus, R. (2008). What is principal component analysis? Nature biotechnology, 26(3):303-304

[6] James, G. (2017). An introduction to statistical learning: With applications in R. New York: Springer.

Acknowledgements

This work was made possible by the support of the School of Science at St. Thomas University. We also want to highlight the support of our research partners for always being present when needed and the Faculty at the School of Science of St. Thomas University. Special thanks to MDC student Kevin Osorio. *This project was supported, in part, by U.S. Department of Education grant award P03C1160161 (STEM SPACE), P031c160143 (STEM EngInE), P120A160036 (STEM ISLE), 1161177 (STEP Up), P120A140012 (SPARC). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the respective funding agency.*