TUFTS UNIVERSITY

# Scraping and Analyzing RateMyProfessors.com Reviews for Gender Bias

Şan Akdağ

December 18, 2019

## Abstract

I wrote a web scraping script to compile all of the information available on RateMyProfessors.com about Tufts professors. I discuss the methods by which I collated, sorted, and analyzed these reviews in order to assess any underlying gender bias. I organized the reviews of male and female professors into two 150k-word documents and created embedding models for them where each word is a vector in a 300-dimensional space. The distance between two words in this space indicates how likely they are to appear together in the document. As I will elaborate upon in the results section, the findings indicate that Tufts students use the same words to rate the class regardless of the gender, but use different words to describe the professors themselves based on distances between word pairs in the embeddings.

# CONTENTS

# 1 INTRODUCTION

I wrote a web scraping script to compile all of the information available on RateMyProfessors.com about Tufts professors. I discuss the methods by which I collated, sorted, and analyzed these reviews in order to assess any underlying gender bias. I organized the reviews of male and female professors into two 150k-word documents and created embedding models for them where each word is a vector in a 300-dimensional space. The distance between two words in this space indicates how likely they are to appear together in the document. As I will elaborate upon in the results section, the findings indicate that Tufts students use the same words to rate the class regardless of the gender, but use different words to describe the professors themselves based on distances between word pairs in the embeddings.

# 2 BACKGROUND/RELATED WORK

This project involves methods for collecting, organizing, and analyzing text data. The first part of this section defines many of these concepts. The first is web scraping. Web scraping is the process of removing text data from HTML states. In the days of the early Internet this was much easier as all of the information on the site was loaded at once, and was all located in the HTML. In newer versions of the internet, asynchrony and other new features allowed website to load data on the fly, making it harder to automatically scrape data. The RMP ste loaded reviews 20 at a time and that made it difficult to scrape. I will go into further detail about this in the data collection section. In this class we learned about what we can do with corpora, or bodies of text. Once I assembled the reviews into two corpora, I created two word embedding models of those corpora from which I could try to draw some conclusions. The model I used, GloVe, maps words into a (in this case) 300-dimensional space in which distance between words indicative of semantic similarity [1].

This project was inspired by a similar project by Ben Schmdt, a professor of Digital Humanities at NYU, from 2015 called "Gendered Language in Teacher Reviews". This project allowed one to query the corpus of all RMP reviews for word frequency then broke up that data by professor gender and department. While this project did not use any high-level model to represent words in this corpus, the author was able to find meaningful results from the data and was able to create a useful web interface for the results. I did not have the capability to scrape and analyze every review for every professor on RMP so I limited my project to reviews of Tufts professors, with the hopes of being able to zoom in on the smaller data set and do an embedding. See the existing project at benschmidt.org/profGender [2].

# 3 Implementation

## 3.1 Data Collection

The first phase of this project was data collection. The data, gathered from RMP, includes name and review data for every Tufts professor listed on RMP. As mentioned, web scraping is the process by which information and content is stripped from sites. I began the web scraping process by determining what information to scrape. Each professor page contains the name, department, and individual ratings and reviews for every professor. This information is encoded in the HTML of the webpage. The HTML file is what the browser uses to render a webpage and can be viewed using the developer tools in Google Chrome (see Figure 1). Each professor is given a "TID" by the site, which is used as the token in the URL to retrieve the reviews for that professor. Once all of the professor IDs were saved to a file, I could start scraping each of the individual professor's pages. Using a tool called Selenium, I automated a Google Chrome window to load each professor's page one by one and scrape the results. On each individual professor's page, only the 20-most recent reviews are loaded when the page is visited; accessing all of the reviews required clicking a "Load More" button which would execute a JavaScript command and send a request to the server which would then send back the next 20 ratings for the browser to render. This loads more content without refreshing the page. The Selenium tool is able to locate and click the elements on the page that load the remaining reviews, which are then scraped the traditional way by parsing the HTML. The program organizes reviews into individual files and those files into folders for each professor.

## 3.2 Data Collation

Once the raw data was collected I had to decide what information was relevant for the embedding. I eliminated everything except the raw text of the reviews and any tags, which students can add to their reviews. Some tags include "AMAZING", "INSPIRATIONAL", and "PARTICIPATION MATTERS". While this was enough information to assemble one corpus for all professor reviews, RMP did not supply gender information, so more work was needed to separate the reviews into two corpora. To determine if there was a gender bias, I had to first determine the gender of the professor. One approach would be to look up every professor, find their page on Tufts' website, and determine gender by the information available there (picture, name, or explicit gender status). However, this seemed too difficult and prone to error. Instead of guessing and running the risk of misgendering professors I used the information available to me, the reviews. Rather than personally make that inference, I relied on the student's use of gendered pronouns in their reviews. The solution was a Python script that went through all of the reviews for an individual professor, and inferred the gender of the professor based on whichever set of gender pronouns (he/him/his, she/her/hers) were used more frequently. This took the onus off of me to correctly identify a professor's gender, and put it on the students who reviewed them. With that information I was able to assemble two 150,000 word corpora from the review data.

### 3.3 Model, Visualization, and Analysis

Once the two corpora were assembled I created two GloVe embeddings for the two corpora. The embedding models represent words as vectors in a 300-dimensional space where distance between words in the space represent semantic similarity, or likelihood of words appearing near each other in the text. The first thing I did was to find all of the words that the two corpora have in common. This turned out to be a little over 2000 words that the two corpora had in common from which I could compare the corpora. I wrote a script that calculated the distance in the space between all word pairs. This led me to the first figure in the results section, which shows pairs of words and the distance between them in the two corpora. The next thing I tried to do was to compare the distance between every word pair in the set of common words in the two corpora and see if there was a correlation. For each pair of words in the set of common vocab I calculated the distance between the two words in the male review corpus and the female review corpus. If there was a correlation, i.e. a correlation score close to 1, between the distances of word pairs in the two corpora, that would show that words pairs tend to appear with similar frequency and that reviews in the two corpora are very similar. If there was a negative correlation, or a score closer to -1, it would mean that words that appear close together in one corpus are far apart in the other and vice versa.

# 4 Results

The first part of the results comes from the distance between selected word pairs. The scripts I wrote called vocab.py fnd words that are common between the two corpora and analyze.py finds the ones with the greatest and least distance between pairs between the two corpora. It seemed that words that describe the class and difficulty appear together with the same frequency in the two corpora. The word pairs with the greatest distance are “CLASS”/”BEWARE” which appear together with much higher frequency in the female review corpus and “very”/”jokes” which appear together with much higher frequency in the male corpus. Other standouts include “mean”/”person” which occurs more in the female corpus and “kinda”/”cute” which appears more in the male corpus.

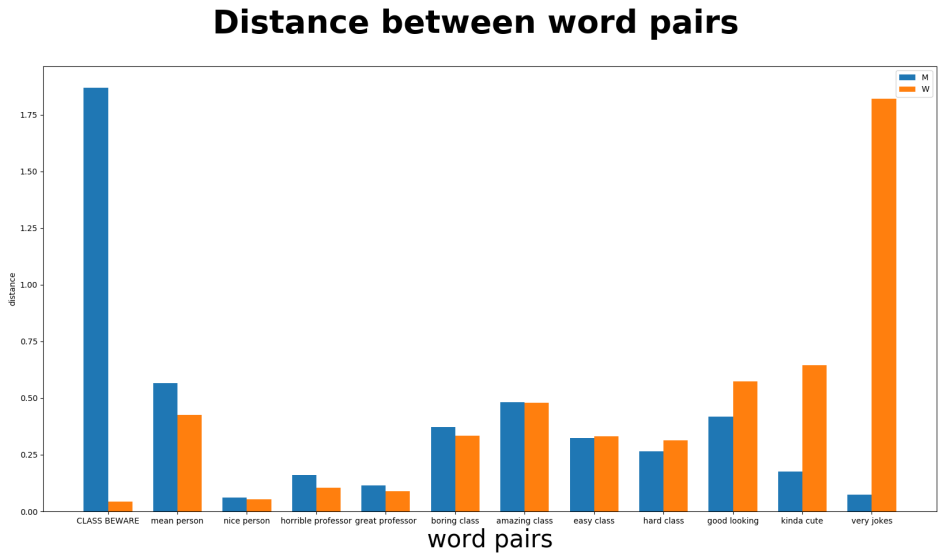


Figure 4.1: Distances Between Word Pairs Across the 2 Corpora

From here, I could tell that there were clear differences in the shape of the model that represented these corpora. I was able to represent both of these corpora visually using the tensorboard projector from the tensorflow project. This program takes the 300-dimensional vectors from the embedding model as input and visualizes them as a sphere in a 3D space. The projector revealed that both corpora have two main clusters at the pole. Visually they appeared similar, but there were differences in the distances between common word pairs so I ran a linear regression on the distances between common word pairs and found a statistically insignificant coefficient of determination of 0.537. While the correlation wasn’t negative (i.e. a big distance between words in one corpus implies a small distance in the other), this result was very far from the 95% confidence interval that certifies a statistically significant result. The projector visualizations of the two corpora can be seen at the links in the appendix.

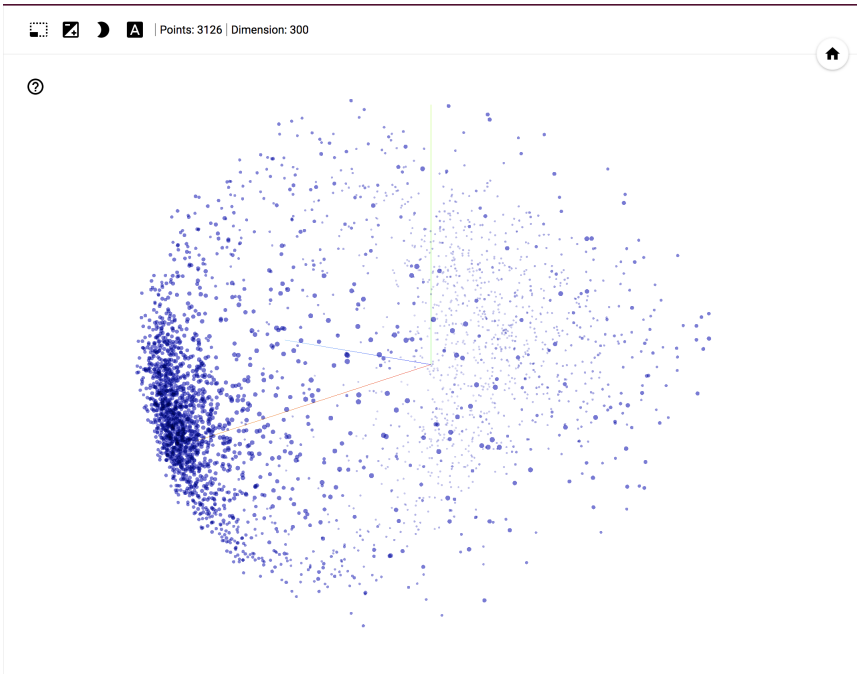


Figure 4.2: Male Corpus Embedding

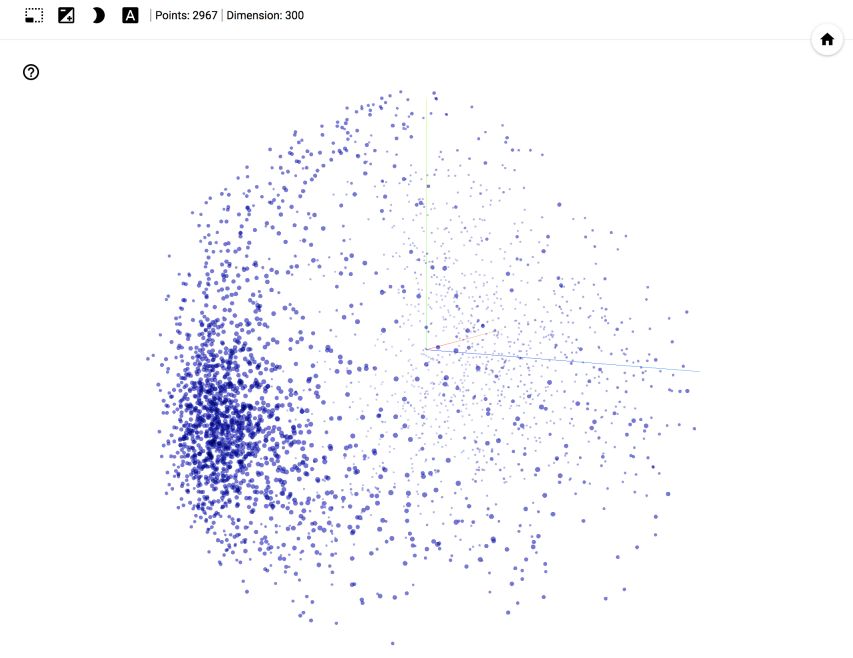


Figure 4.3: Female Corpus Embedding

# 5 Discussion

Given this limited data set and limited result, it is fair to argue that these results are insignificant. But as similar studies suggest RMP reviews reflect what might be expected of traditional measures of student learning [3][4]. There seems to be an academic consensus on the validity of these reviews and the purpose they serve when formal student evaluations are not accessible.

## 5.1 Quality Control and Future Work

There were also some issues. The first was that I did not remove any proper names from either Corpus, which definitely influences the embeddings and nearest neighbors in the vector spaces. If I were to do this again, I would not include proper names in the corpus. Another issue is the size of the data set. Although there were over one thousand professors on the site, only a handful of professors had over 50 reviews reviews. This resolved to two 150,000-word corpora but it would be interesting to do an embedding on a corpus at the scale of the whole website. The resources I had used to theorize this project both either had much larger data sets. The project that was the main inspiration for this one was over all of the reviews. An improved version of this project might try to see if these trends are consistent across colleges and universities in similar cohorts like the NESCAC, Ivy League, or public vs. private colleges of a specific region before scaling to the entire site. Another idea I had for an improvement on this project would be to assemble two extra 150,000-word corpora from random reviews of professors on the site, all from different departments and universities and compare the embeddings done on those corpora with those of Tufts professors' reviews. This would put my results into better perspective if we could see how the distance between vectors in the Tufts corpora compares to those of the random corpora.

# REFERENCES


[1] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1532–1543. [Online]. Available: http://www.aclweb.org/anthology/D14-1162

[2] B. Schmidt, "Gendered language in teacher reviews." [Online]. Available: http://www.benschmidt.org/profGender

[3] J. Otto, D. A. Sanford Jr, and D. N. Ross, "Does ratemyprofessor. com really rate my professor?" *Assessment & Evaluation in Higher Education*, vol. 33, no. 4, pp. 355–368, 2008.

[4] M. J. Brown, M. Baillie, and S. Fraser, "Rating ratemyprofessors. com: A comparison of online and official student evaluations of teaching," *College Teaching*, vol. 57, no. 2, pp. 89–92, 2009.

# 6 Appendix

Male embedding link:
http://projector.tensorflow.org/?config=https://raw.githubusercontent.com/sanakdag/CorporaFinal/master/tensorboard/m_projector_config.JSON

Female embedding link:
http://projector.tensorflow.org/?config=https://raw.githubusercontent.com/sanakdag/CorporaFinal/master/tensorboard/f_projector_config.JSON