Final Project Report

COS 424 Interacting with Data (2014)

Byron Vickers, Mike McKeown, Gecia Bravo Hermsdorff

1 Introduction

We want to build a ranking of student enjoyment and satisfaction on graduate (PhD) programs at universities across the US.

We want to do this for a couple of reasons: It's pretty entertaining, for one, but it's also valuable information for any prospective graduate student trying to decide where to apply, or which offers to accept. The current tools available for doing this are pretty limited: sites such as `usnews.com` and `forbes.com` offer little if any information on student satisfaction, while less formal ranking systems such as `collegeprowler.com` (now moved to `colleges.niche.com`) have sparse data and suffer from the bias problems inherent in any crowd sourced review system.

As mentioned above, there are a number of review websites which offer partial data on student satisfaction in graduate programs. There are also several university-specific results on student satisfaction; internal surveys done by universities to gauge what they are doing well and what needs to be improved.[1]

We would like to take a completely different approach to the problem: using sentiment analysis on the acknowledgment sections of submitted theses to try and gauge how satisfied students were at the point where they finish their degrees and are about to leave the university.

There are 3.6 million english PhD theses listed on `proquest.com`. We propose to use a subset of these as our text corpus, most likely selecting a thousand or so from each of several top-ranked US universities. We will then extract the acknowledgments text from each thesis, run a (possibly made-from-scratch) sentiment analysis algorithm on it, and use these to build a profile of the university's student satisfaction.

While the rankings website data is useful as a sanity-check for our model (we expect our results to correlate strongly with the rankings, but probably not exactly), the university-commissioned student satisfaction surveys should give a good value for the ground-truth values, and we can use these to validate the predictive power of our algorithm. We can also use resources such as `http://nlp.stanford.edu:8080/sentiment/rntnDe` to check the performance of our sentiment analysis algorithm if we choose to develop one from scratch ourselves.

In attempting to rank university graduate programs based on student satisfaction, we would like to apply machine learning techniques to a text corpus consisting of dissertation acknowledgment sections from different universities. We believe there will be a correlation between features of the dissertation acknowledgment sections and student satisfaction. For example, if an acknowledgment is longer and more positive, it may indicate the student is more satisfied than terse acknowledgment sections.

We plan to explore a small collection of methods to extract university rankings based on the dissertation acknowledgments. We will first need to extract features from the acknowledgments. A couple features we have discussed already include the length of the acknowledgment section and the count of keywords in the text that may have a positive or negative connotation, however, there are probably others that we will explore as well. The features could then be used individually to obtain some sort of score for each acknowledgment (in which an average for each university would result in a ranking) or

[1] Some examples: `http://www.utexas.edu/news/2011/12/01/grad_climate_study/`, `http://www.ohio.edu/education/news-and-events/upload/Final-2013-Student-Satisfaction-Report-8-23-13.pdf`, `http://ga.berkeley.edu/files/page/surveyreport.pdf`

could be aggregated for each university (based on all acknowledgments from the university) and obtain a ranking score for each university.

As for the machine learning algorithms, there are a number of different approaches to explore. As mentioned previously, we could use sentiment analysis to obtain a score for how positive or negative an acknowledgment section is. This would allow us to rank universities based on their average sentiment score. Another possibility is to use a regression to determine a score for each acknowledgment or university. We could also use a binary classifier to determine if a given acknowledgment is better than another and use this information to determine a score for an acknowledgment or university.

There are also machine learning algorithms specifically tailored toward ranking, such as Ranking SVM, Bayes Rank, etc. While we have not done too much research into how these algorithms work and the assumptions they make yet, we plan to do so in determining which ones we will explore. In short, there are many data analysis techniques that can be applied to this problem, and we intend to explore a subset of them that make valid assumptions about our problem and match up with the type of data we have and our goal.

2 Data Collection

Write about how we collected the data – how we scraped it, what methods we used to decide exactly which PhDs to get, and so on.

Then talk about how we got the acknowledgements sections out of the PDFs. What the issues were? How often extraction failed, roughly. Bag of words and positive ratio extraction from the text.

[Remove this section before submission]

The theses used for our analysis were taken from the ProQuest Full Text Thesis and Dissertation Database[2]. This database contains 1.5 million theses, when doing searches for English-language documents with full text available. The database includes metadata specifying features such as date of submission, University, Department, subject, and so on. All of this is freely available to licensed users – notably, any user whose requests originates from within the IP block of a subscribing university (such as Princeton).

The data collection occurred in two stages: collecting the theses, and then running a parsing and extraction routine on them to pull out the text of the acknowledgements section.

Collecting theses (scraping)

ProQuest has a database API exposed at `http://fedsearch.proquest.com/`. This access point seems to accept SRU (Search/Retrieval via URL)[3] queries, but fails to respond to the standard `Explain` operation, thus making its valid query terms unaccessible. Documentation for the service appears to be non-existent.[4]

We were successful in reverse-engineering the API to the extent where we could collect 1000 URLs of fulltext English theses. This is the dataset we went on to use for an analysis of PhD satisfaction by state over the United States. However, this dataset was insufficient for any analysis of individual universities.

At this point we switched to scraping documents via the standard, human-facing ProQuest search interface. When queries are made through this interface, ProQuest attempts to authenticate the humanity of the downloader using captcha-gateways for any downloads beyond a small number. Via various subterfuges, however, we were able to circumvent these measures and download 200 theses for each of the Ivy League universities.[5] These form our second dataset.

Feature Extraction

Blah blah blah blah blah blah blah. All yours, Mike.

[2] Searchable via `http://search.proquest.com/pqdtft/`

[3] See `http://www.loc.gov/standards/sru/`

[4] Though one user reports the existence of a document available from ProQuest on request: see `bibwild.wordpress.com/a-proquest-platform-api/`. He claims it is not very explanatory.

[5] The number 200 was chosen mainly to keep data processing times reasonable. With enough time (or compute power), our method should be easily extensible to at least several thousand theses per university

3 Data Analysis

How we analysed the data, once we had it out of the PDFs and into R.

Models we used (gaussian emission, and naive bayes). Issues with implementing the models.

4 Results

Lots of pictures. Tables. Discussion of pictures. Discussion of tables. Anything else that fills space.

5 Conclusion

Summary of entire project. Issues, successes. Directions for extension in future work. Pithy concluding sentence.

6 References

List of relevant papers, some of which we should definitely be citing in the main text.

7 Appendices?