CS Rankings Estimation

Question: Using the subreddits of accredited universities as a representation of the university, is there an accurate way to ascertain a school's relative CS ranking?

Note: We will use CSrankings.org as an "actual" ranking for comparison.

Hypothesis: Of three measures that we plan to use as a tool for comparison-- frequency counts, Vector Space, and PageRank-- we propose that PageRank will be the most accurate indication of relative computer science program rankings, as the algorithm is the most objective. We can see that this is the case, as PageRank only considers the names of the schools, while the vector space model and the frequency count are both subject to bias in creating keywords.

Methodology: We used Python and the PRAW wrapper to extract the 500 newest posts from the subreddits of the listed top-20 computer science programs. We analyzed the resulting data using three methods:

1: The Vector Space Model

Building on the starter code from HW4, we compared the documents of each subreddit's posts with the document containing relevant CS keywords using cosine similarity. That is, after populating a document with common cs-related keywords-- CS, CSE, CIS, EECS, Comp Sci, CompSci, CSCI, Computer Science, Computer Sciences, Computerscience, C Science, Cscience-- we used the vector space model to get a measure of similarity between this document and the document representing every university.

2: A basic frequency comparison of CS keywords.

This method called for simply counting the number of occurrences of the cs-related words in each document and using this data as a measure of relative ranking.

3: A PageRanking comparison.

This method relied on references to other schools' or their respective subreddits. We then ran PageRank by considering those references "outlinks".

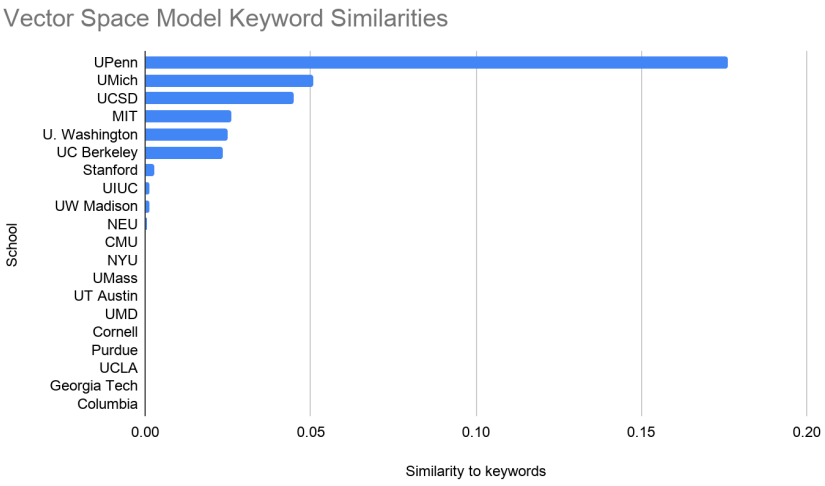
Results: We found that the most accurate method was the frequency comparison. Though this disproved our hypothesis, we were excited to note the effectiveness of frequency comparison.

We first note the “accurate” relative order of cs programs:

1. CMU
2. UIUC
3. MIT
4. Stanford
5. UCSD
6. Berkeley
7. Cornell
8. UMich
9. U. Washington
10. UMD
11. GA Tech
12. NEU
13. Columbia
14. UW Madison
15. UPenn
16. UT Austin
17. Purdue
18. UMass
19. NYU
20. UCLA

We now take a closer look at each individual method:

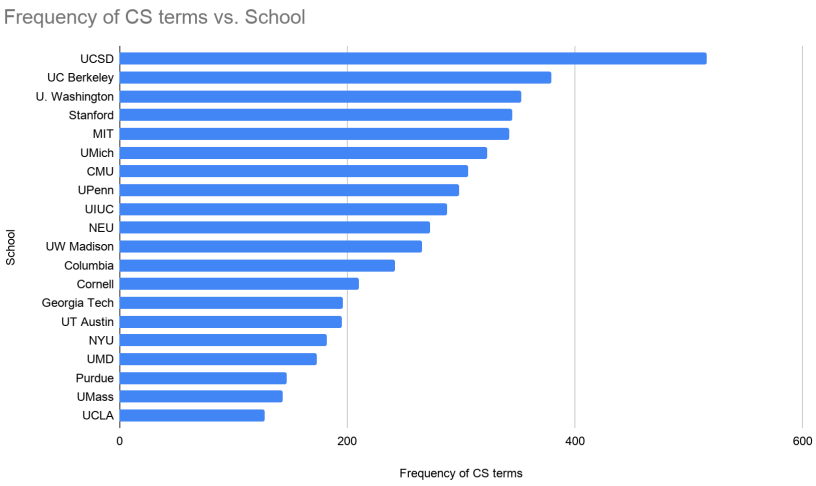
1: <u>The Vector Space Model</u>



This method had an average error of being off by 4.9 rankings.

We noted that UPenn was a large outlier and propose a possible reason being bias. Since we browse Penn's subreddit the most, it is likely that we included more terms that can be found there, including "CIS", just because of these words being used more often in our own vocabulary. In order to counter this bias, we tried to include a more diverse group of terms like "EECS" and "CSE".

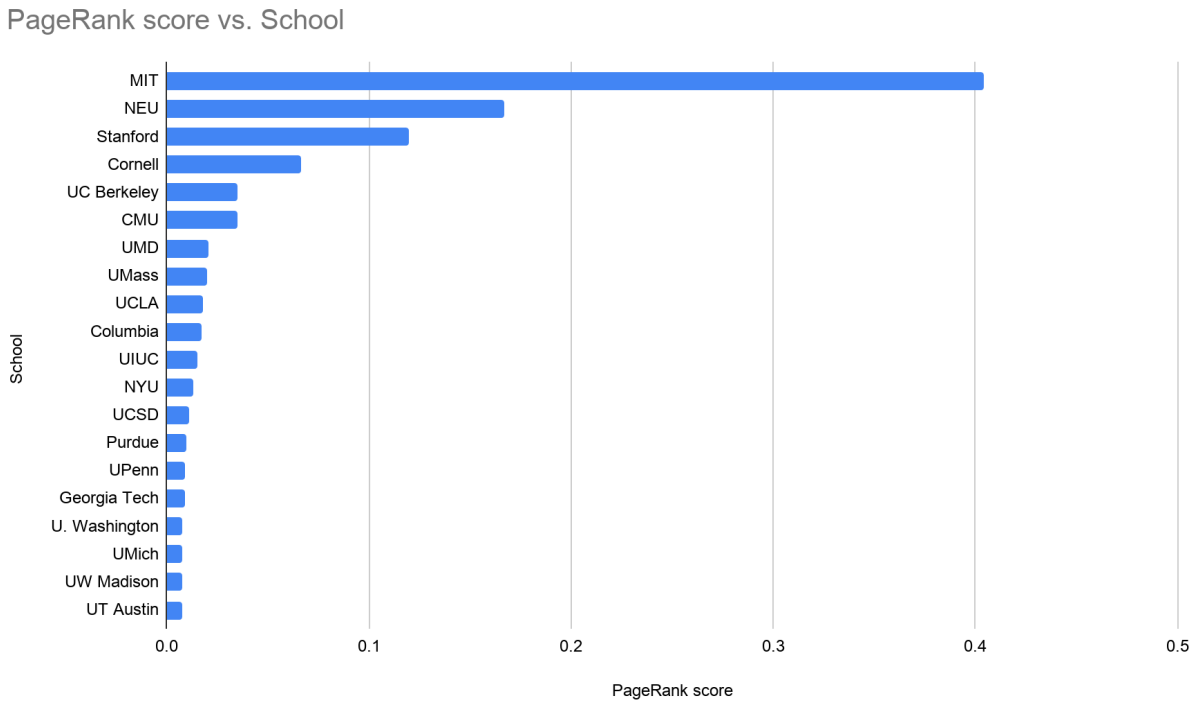
2: <u>A basic frequency comparison of CS keywords</u>.



This method had an average error of 3.3 rankings.

We note the outlier of UCSD and propose that the error could be because the method of searching for "CS" is particular would skew the data in this direction, as the school's name includes that term.

3: A PageRanking comparison.



This method had an average ranking error of 5.4 rankings.

With the values of each university using each method, we were able to calculate the error against the "true" rankings as follows:

Method	Average Error (i.e. avg ranks off by)
Vector Ranking	4.9 (total deviation of 98)
Frequency	3.3 (total deviation of 66)
PageRank	5.4 (total deviation of 108)

Note: error was calculated by finding the number of rankings each individual school was off by, getting a total, and dividing by the number of schools.

Clearly, frequency is the most accurate measure with the smallest error. We see that this could imply a correlation between the number of times a school talks about computer science and its overall ranking. It is possible that if a school has a more computer science oriented curriculum, a higher proportion of students are likely to discuss or be involved with computer science and therefore post about the topic more often. There may also be a factor of pride involved, leading to a greater number of posts containing cs-related words.

Looking forward, we wonder if the idea behind this project could be extended to consider other majors and possibly help students better understand which school might be the best fit for them. Considering that many rankings take into account tuition and location, a student solely concerned with student interests and overall atmosphere may benefit from such an analysis.

For a table of the rankings, please see the file called “Findings.pdf”. This file contains a table comparing the expected vs experimental rankings, as well as charts depicting the relative rankings and average errors. Raw data is enclosed in “Spreadsheet_Data.zip”. All code is in “RedditScraper_Code.zip”.