

Brian Zhao 96102936
Kevin Zheng 92169660

Before optimization:

Well-performing queries:

1. ACM
2. Cristina lopes
3. Hitesh Sajnani
4. Pattis
5. Thornton
6. Software engineering
7. Python
8. Master of software engineering
9. Mswe
10. Wics
11. Java

Badly-performing queries:

1. Uci computer science undergraduate
2. ICS 121
3. ICS 46
4. University of california irvine
5. Water
6. Taco bell
7. Information retrieval
8. ICS
9. To be or not to be
10. What when where how why
11. Computer science

The badly-performing queries did so because they gave irrelevant results. Their results tended to be long pages of text that just happened to have a lot of the query words in them. To improve query result relevance, we implemented cosine ranking and increased the weighting of title and heading words. Now, top results tend to have important query words as their title or main topic.

“What when where how why” and “to be or not to be” performed badly because they took too much time. The problem is that they are composed of all stop words, which are very frequent. To improve search speed, we sorted our postings by tf-idf score and implemented a maximum posting retrieval limit for queries with mostly stop words.

We also implemented pagerank in order to increase the scores of home pages and authoritative pages. This strategy greatly improved queries such as mswe, cristina lopes, and master of software engineering. Home pages and faculty profile pages are ranked higher because of this.