Assignment 3: M3 Report

By Skylar Hoffman(21464998), Keno Ha(85953583), Jordan Williams (53041731), Jingwen Li(43521573)

Query List

Fast Queries (< 300 ms)

1. **Computer Science**
2. **Classes**
3. **Professor**
4. **Machine Learning**
5. **artificial intelligence**
6. **Courses**
7. **Informatics Webinar**
8. **Courses Courses Courses Courses Courses**
9. **No**
10. **fast**

Slow Queries (>300 ms)

1. **Machine Learning and Data Mining**
2. **A b c d e f g**
3. **to be or not to be**
4. **Machine Learning and Data Mining Machine Learning and Data Mining**
5. **machine learning computer science information artificial intelligence**
6. **computer science class on artificial intelligence machine learning data mining and information sciences taught in winter 2020**
7. **ics uci edu**
8. **information and computer science department courses professor martins**
9. **a b c d e f g h i j k l m n o p q r s t u v w q y z**
10. **i like computer science very much**

To make our queries perform better we did:

1. **We added a heuristic to remove documents that are coming from tokens with low idf scores. This improved the performance of longer queries with many tokens and documents.**

2. **Opened files at beginning of program so would not have to reopen for each query**
3. **We created a seek index and document vector length index for retrieve the inverted index for tokens and information for computing idf scores**
4. **We stored the sum of squares of the doc in memory to improve speed. It allows us to do preprocessing of the scores before query time.**
5. **We reduced the number of document vectors before calculating cosine similarity to improve performance so we did not have to compute the similarity for non-important documents.**