Assignment 3: Search engine

Chunzhi Xu 15551944
Zhanbin Luo 76257680
Lucas Wen 27607748

20 test queries:

1. cristina lopes
2. machine learning
3. ACM
4. Master of Software Engineering
5. informatics
6. neural network
7. Computer Science
8. department
9. research opportunity
10. aux
11. computer science and engineering
12. computer game science
13. Donald Bren School
14. data science
15. University of California
16. human computer interaction
17. irvine
18. Software Evaluation
19. University of California, San Diego
20. a

For Query 10 -- **aux**, we choose **aux** since Windows operating system prohibits us to create **aux.json**. **aux.json** is reserved filename in windows. To avoid doing that, we choose the first two initial characters to make an index file. Also, we create an index file with the word itself for words have more than five characters.(ex: student --> student.json)

For Query 1,2,3,6, we at first doing poorly on ranking the result, so we justify the value of an important word (title, subtile, bold, …) to make the result more acceptable. We try to increase important words frequency to have better score on the ranking system.

For Query 4,7,11,12,13,14,15,16,19, the efficient for these queries are poor and unstable at first since there are too many document having these words. We tested the time taken by each function one by one to find out which function takes the longest time. By doing so, we found that the efficient is poor since the index file is quite large and it is time consuming to load the file to python dict. Thus, we modify the indexer to make it create an index file with the word itself for words have more than five characters. This helps us make the query time reduce to less than 100ms for these queries.

For Query 20, we first end up having to answer since we removed all the stop words in query. Thus we and the tf-idf score since we know that the idf score can handle stop words (common words).

Read readme.md to see more information about functions and conclusions.
GitHub repository address:
https://github.com/CurtisXuCAD/Information-Retrieval---Simple-Search-Engine
(I will make it public after the due date)