Information Retrieval -- IS322-2024

Assignment # 3

Groups of ~4
All three must be able to describe the WHOLE program
Write a java code that

1- A web crawler that crawl wikipedia starting from the following 2 seeded
https://en.wikipedia.org/wiki/List_of_pharaohs
2- Build the inverted index for visited pages
3- get a query (set of a number of words)
4- compute the cosine similarity between each file and the query
5- rank the top k=10 files according to the value of the cosin similarity

Hints:

```
String result = "";
String[] terms = phrase.split("\\W+");
int len = words.length;
double scores[] = new double[N];  // N= collection size (10 files N =10)
//1 float Scores[N] = 0
//2 Initialize Length[N]
//3 for each query term t
for (String term : terms) {
//4 do calculate w t, q and fetch postings list for t
    term = term.toLowerCase();
    int tdf = index.get(term).doc_freq;  // number of documents that contains the term
    int ttf = index.get(term).term_freq; //
//4.a compute idf
        idf = log10(N / (double) tdf);   // can be computed earlier
      //5 for each pair(doc_id, dtf ) in postings list
      //6  add the term score for (term/doc) to  score of each doc
        scores[p.docId] +=  (1 + log10((double) p.dtf)) * idf);

 //Normalize for the length of the doc
//7 Read the array Length[d]
//8 for each d
//9 do Scores[d] = Scores[d]/Length[d]
//10 return Top K components of Scores[]
```

بالتوفيق ان شاء الله
Good luck