CS380 – Homework #2 (Questions)

1. Since the submission is another report, and the first question asks for a table, do you just want a table for the first page? Or, do we need to describe the dataset right before the table so that it looks more like a report? If we need to describe the dataset right before the table, can we just use the same description we used in Homework #1?
    a. Just table is fine. Since I am telling you what to put in, it seems repetitive to have you define the content

2. In the first question, when creating the table for one document's 50 highest-ranking tokens in my collection, do you want to see the calculations / computations done to obtain the IDF weights, TF weights, and TF-IDF weights?
    a. I do not need to see the calcs. The columns have plenty of data for me to reproduce if I care to.

3. Does it matter which document I choose from my collection when I fill in the table in my answer for the first question? Does it have to be the longest or the shortest document? Do I randomly choose which document from my collection to use?
    a. Choose a document. I would say compute the centroid and use that, but since I have not discussed centroids and that would be a lot more work, just pick a document.

4. In the first question, the assignment sheet says that one of the columns has to be "the number of documents in the collection". All 50 rows in the table should have the same number in this column, right?
    a. Correct — kind of boring but it gives me data I would need to check your calcs

5. In the first question, the assignment sheet says that one of the columns has to be "the frequency of the most common token in the document". This would just be one value, right? Are all 50 rows of this column also supposed to have the same number?
    a. Again yes. Again boring but needed if I am to check.

6. In the second question, when we have to explain how our IR engine works, what exactly are you looking for? Are you looking for how the TF-IDF weights contribute to what results turn up for a given query?
    a. More like how you implemented the process of finding documents. Did you use an inverted index. If so, how is the inverted index implemented.

7. For the script, how many queries should be sufficient?
    a. An excellent question. The problem is that the answer is strongly dependent on your data and how you write your queries. So I cannot give a specific answer. I would guess that something less than 10 should be enough.

8. What might make the script interesting or significant?
    a. At the very least, not finding the same set of documents for every query. You might print out a "match score" for each document. You might comment on why the match scores differ and why some high ranking matches are good / bad. Things like that.

9. How long do you expect this assignment to take?
    a. 15 hours or less. Certainly not more. I will probably more the due date to later in the week. Wednesday or Thursday. Since I have not given you much other work 15 hours seems not unreasonable.

10. I was wondering if you could explain how to calculate the IDF weight for a token in a query. I'm just a little confused because if I treat a query as a single document in a collection wouldn't it be 0, but I think I'm calculating it wrong.
    a. IDF is not dependent upon the query
    Let N = number of docs in collection
    ni = number if documents in which term i appears in collection.
    Then for term i its IDF = log(N / ni)
    Once the collection is set, the IDF for every term can be computed and does not change.