**Problem Set 1**

Name: Min Guo

NID: N10971010

NetID: mg4517

**Problem1.**

(1) **Q**: (MRS Ex 1.9 p. 13, regarding the INTERSECT algorithm in figure 1.7.) For a conjunctive query, is processing posting lists in order of size guaranteed to be optimal? Explain why it is, or give an example where it is not.

**A**: Yes, processing posting lists in order of document frequency is guaranteed to be optimal.

By intersecting the two smallest posting lists, all the intermediate results must be smaller than the smallest posting lists. So, we have the least total work to do.

(2) **Q**: The time given in MRS for the procedure INTERSECT(p1,p2) in figure 1.6 is *x+y* where *x* is the length of p1 and *y* is length of p2. If you use additionally a hash table in which the key is the pair < word, docID >, and you record with each word the length of its posting list, then this can be made significantly faster. How do you use the hash table, and how fast does the revised algorithm run?

**A**: Instead of processing two posting lists, the longer posting list can be replaced by hashtable. We can put the documentID of the longer posting list into a hashtable. We can loop the documentID in the shorter posting list, and check if it is in the hashtable. If it is in the hashtable, this documentID can be stored. In this situation, the total time of processing is linear time x where x is the length of p1 with the shorter length of posting list.

**Problem2.**

**Q**: If you Google my office telephone number with the query 212-998-3123 then my home page and various course pages of mine turn up among the top results. Do a little experimentation to see how much you can vary the format and still get this match. Discuss the difficulties with tokenizing worldwide telephone numbers.

**A**: By searching the formats 212 998 3123, 212-998-3123, (212)998 3123 and (212)9983123, we can get better result. However, the formats 2129983123, 212-9983123 is worse.

Since different country has different punctuations for telephone number, tokenizing worldwide telephone numbers must follow a standard punctuation rules if we want to get better match.

Normally, the worldwide telephone number is tokenized by its punctuations. Its standard format is (area code)–(local code)–(local number). More accurate standard format we follow, it is much easier to be matched.

If we only provide all of the digital number without any separation, it will be tokenized as a unit. There is no way to distinguish which part are area code, local exchange, and local number.

We can use white space or other marks replace”-”, the telephone number can be tokenized the same result.

**Problem3.**

(1) **Q**: Propose a reasonable objective function, which would take into account the amount of change. (You should describe it as a cost function to be minimized, where the cost is 0 if the index has the current version.)

**A**:

(2) **Q**: Describe in general terms what a crawler would need to do in order to try to keep the objective function in (1) small. What kinds of information would it need to compute? How would it use that information in deciding on a refresh rate for a web page?

**A**:

(3) **Q**: Failing to refresh page A may not only cause A to be misinidexed; if page A acquires a link to page B and this is the only link to B, then B cannot be crawled until A has been refreshed. Discuss how this consideration could be incorporated in a refresh strategy.

**A**: We can compare the refresh frequencies of A and B. If the refresh frequency of page A is higher than page B, we can refresh page A as its refresh strategy. If the refresh frequency of page B is higher than page A, we can refresh page A as B’s refresh strategy. By doing so, we can always get the fresh pages both A and B.