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: Initialize the intermediate result by loading the posting list of least frequency term, all the other posting lists sorted by increasing frequency. Every time we chose the smallest posting list from the rest lists to intersect with intermediate result. By doing this, the intermediate result always has the smallest posting list. 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 O(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 punctuations or separator. Its standard format is (area code)–(local code)–(local number). More accurate standard format we follow, it is much easier to be matched.
We can use white space or other marks replace"-", the telephone number can be tokenized to the same result.
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. Moreover, it cannot recognize that whether it is a telephone number or not. In that case, it can be any information that contains the numbers.

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: We can consider the amount of change as the same format as age

$$Change(\lambda, t) = \int_0^t P(page\ changed\ at\ time\ x - V_0)(t - x)dx$$

If the amount of change smaller than V_0, we assume that the page has the current version.
(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: We need compute the change amount from previous version to current version, keep the change amount no bigger than V_0. The crawler should always compare the change amount with V_0 to keep it smaller than V_0.
Suppose that the change value from previous version to current version is V, the value $P(V - V_0)$ can be treated as a refresh rate for a web page.

(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 change frequencies of A and B. If the change 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.
Or we can store all the links (including B) in as a list; refresh all the pages (including B) every time. We treat B as an independent page.