CS 345: Homework 2 (Due: Oct 12, 2015)

Rules. Individual homeworks; see Handout 1 (aka Syllabus). Hardcopies may be submitted no later than start of class the day they are due; electronic copies by 23:59:59 the same day.

Problem 1. (21 points) **Do the w_1, w_2 experiment**

Last time we tried this problem the best set of index-terms to estimate reliably the corpus size of Google and Bing was `pasta` and `physics`. The textbook's attempt of `tropical` and `Lincoln` does not work any more. And it is very likely that pasta and physics won't work either! We ask you to try find the three set of queries that do not include any of the previous index-terms nor anything similar to them (eg. mathematics instead of physics or polar instead of tropical) for w_1, w_2 and w_1 w_2 respectively so that the estimate for both engines is reasonable/reliable and over 20G but less than 40G. The same set will query Google and Bing. The w_1, w_2 and the reported hits will be included in E1, E2, E3 and the estimate of the Corpus size of Google and Bing will be shown in E4. Provide screenshots.

No	Query Term(s)	GOOGLE No of hits	BING No of hits
E1	w_1 is $\langle$ $\rangle$		
E2	w_2 is $\langle$ $\rangle$		
E3	w_1 w_2 is $\langle$ $\rangle$		
E4	Corpus size estimate is		

Table 1: Table for HW2 Problem 1

Problem 2. (21 points) **Quality Assurance**

We perform the same query **CS345 Web Search** in Bing and Google to find course related material. Results are on the course web-page, Section B with three files of links for Google and three for Bing with a total of 10 links each. A relevant results is considered one that relates to "CS345 Web Search course at NJIT".
(a) (8 points) Find and count the retrieved documents shown on the three figures (per engine) which is obvious, the relevant documents and then compute the precision for the query. This info goes into the `Values` column of the table below. If Google is better, give one point to Google or Bing gets one point when you complete the `Points` column. If it is a tie, give each engine one point. The last 4 rows below are computed after you complete part (b) to follow.

```
                                        Values      Points
                                      |Bing  |Google |Bing  |Google|
 # Retrieved docs for Query           |      |       |      |      |
 # Relevant docs  for Query           |      |       |      |      |
   Precision      for Query           |      |       |      |      |
  20%  recall interpolated precision  |      |       |      |      |
  60%  recall interpolated precision  |      |       |      |      |
  100% recall interpolated precision  |      |       |      |      |
  3-point effectiveness(20,60,100)    |      |       |      |      |
=========================================================================
 Number of point wins  (sum)          |  -   |  -    |      |      |
```

Who is the winner? Number of points won by the winner?

(b) (13 points). Use the information available to determine and tabulate the results below and also use them for the next problem in order to establish effectiveness. In computations round up to the next integer (i.e. take a ceiling of a percentage point i.e. a 33.3% would become 34%).

```
Relevant Documents
Q2 :  Bing                      Google
    Rlvnt|Recall Prcsion    Rlvnt|Recall Prcsion
d1 |     |      |       ||      |       |
d2 |     |      |       ||      |       |
d3 |     |      |       ||      |       |          Compute Effectiveness at 20% 60%, 100% recall rate (interpolate)
d4 |     |      |       ||      |       |          Q2:   Bing                        Google
d5 |     |      |       ||      |       |          Recall  Prec  InterpP       Recall Prec InterP
---|-----------------------------------------     0%     -                     0%     -
d6 |     |      |       ||      |       |          20%                        20%
d7 |     |      |       ||      |       |          40%                        40%
d8 |     |      |       ||      |       |          60%                        60%
d9 |     |      |       ||      |       |          80%                        80%
d10|     |      |       ||      |       |         100%                       100%
---|-----------------------------------------   Effectiveness:              Effectiveness:
```

Problem 3. (24 POINTS) **(Paper by Brin and Page)**
Read the paper (pdf and HTML through link L1 in section C5 of the course web-page). For Hashin consult wikipedia or your favorite CS114 or CS435 textbook. Justify your answers based on the info of the papers.
`http://en.wikipedia.org/wiki/Hash_table`
might also help.

(a) According to the paper, what was the size (bytes) of Google's Lexicon around 1998?

(b) According to the paper how many bits for a `docID`? Quote the paper.

(c) Describe the data structures used by Google for indexing only (not all of them are listed in the architecture figure).

(d) In what data structure(s) does Google use binary search?

(e) Does Google (1998) use a hash table for the dictionary? What do they use? Why ?

(f) How many bytes are assigned to each hit (of the hitlist)? How many types of hits? What are they (types of hits)?

Date Edited: 9/17/2015