## CS 345: Homework 2 (Due: Sep 30, 2014)

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 NOON-time the same day.

## Problem 1. (15 POINTS) (Estimate Corpus size of Google and Bing)

- (a) Write one query in Google that contains only disjunctions and no conjunctions or negations that returns a number of documents is higher than 25,000,000,000. Repeat the same for Bing. Provide the two screenshots. (The grading will deduct points based on the number of terms you use for the query but add points if number is much higher than 25 billion. You will use substantial number of points if for any reason I detect a conjunction or negetion. The queries may be different.)
- (b) Write one query (per search engine) in Google and Bing that contains disjunctions and no conjunctions that return a number of documents for BOTH Bing and GOOGLE that is higher than 25,000,000,000. Provide two screenshots.

(The higher the number it is the more points you get; the more operators you use the more points you lose.)

## Problem 2. (15 POINTS) (Quality Assurance: Part A)

We perform two queries in Bing and Google. (Results are on the course web-page.)

Query Q1 is CS345 Web Search and we got a single page with 8 results for each search engine; number of retrieved documents is thus 8.

#### What to do.

Find and count the relevant documents and then compute the precision for Q1 and also for Q2. This info goes into the Values column. 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.

Query Q2 is CS345 Web Search NJIT. We got more hits, but we only include 20 retrieved documents in several screenshots for each search engine. To help you identify relevant links for Q2 we comment on the following. For Bing links 6,8 are not relevant but 17 is. For Google 8 and 9 are relevant but 17 is not. Note that there are 3 screenshots for the Bing results and 4 for the Google results.

What to do. Use the information available to determine and tabulate the results below and also use them for the next problem in order to establish effectiveness.

**Calculations:** Round to the closest multiple of 5. Thus 4/6 is a no brainer 65% and a 4/7 is a 55%.

	Values		Points		
	Bing	Google	Bing	Google	l
# Retrieved docs for Q1	1	1	1	1	l
# Relevant docs for Q1	1	1	1	1	l
Precision for Q1	1	1	1	1	l
# Retrieved docs for Q2	1	1	1		l
# Relevant docs for Q2	1	1	1		l
Precision for Q2	1	1	1	1	l
25% recall interpolated precision for Q1	1	1	1		l
50% recall interpolated precision for Q1	1	1	1		l
75% recall interpolated precision for Q1	1	1	1		l
3-point effectiveness(25,50,75) for Q1	1	1	1	1	l
Number of point wins (sum)	1 -	1 -	1	1	I

# Problem 3. (15 POINTS) (Quality Assurance: Part B)

Deal now just with question Q2 of the previous problem. Give a complete write up of a 6-point effectiveness computation for Query Q2.

Calculations: Round to the closest multiple of 5. Thus 4/6 is a no brainer 65% and a 4/7 is a 55%.

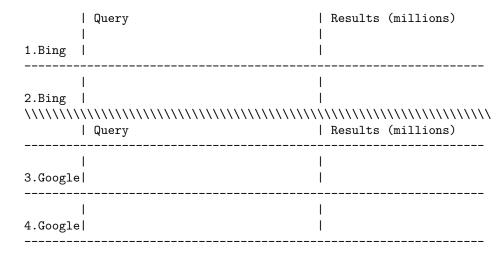
Rele	vant Doc	uments					
Q2 :	Bing				Google		
	Relevant	Recall	Precision		Relevant	Recall	Precision
d1		1	1				
d2		1	1				
d3			1			l	
d4		1	1				
d5		1	1				
d6		1	1				
d7		1	1				
d8		1	1				
d9			1			l	
d10			1			l	
d11			1			l	
d12			1			l	
d13			1			l	
d14			1			l	
d15			1			l	
d16		1	1				
d17			1				
d18			1			l	
d19		1	1			l	
d20		1	I	$\Pi$		l	l

Q2: Bing		Google		Compute Effectiveness
Recall P	rec InterpP	Recall Prec I	interP	using data from 20%, 50%, 80%;
0% -		0% -		(No decimals, round up)
20%		20%		
40%		40%		
60%		60%		
80%		80%		
100%		100%		
Effectivene	ss:	Effectiveness:		

recall

### Problem 4. (15 POINTS)

Formulate 2 queries in Google and 2 in Bing that are equivalent to "answer is all documents of the Corpus". The queries do not need to be the same (i.e. you might have 4 different queries at the end rather than 2 queries that work for both Google and Bing at the same time). Report results. Acceptable queries are those NOT introduced in class, those NOT similar to those introduced in class, and ones which give a number of documents in the tens of billions (25 billion or more, preferably 27 billion or more), and also NOT used in Problem 1! Present screenshots!



Date Posted: 9/16/2014