CS 345: Homework 4 (Due: Before midnight of Nov 29, 2016)

Rules. Individual homeworks; see Handout 1 (aka Syllabus).

Problem 1. (10 points) **Follow-up from HW1**
By providing screenshots, or other hard evidence explain how often your weg-page gets crawled by Google. Be reminded that in Problem 1 of Homework1 you setup the framework to collect this information. You might still have time to do so between the time of the posting of this homework (early november) and the due time (late november).

Problem 2. (15 points)
Use the best approach possible to evaluate a query
`tA AND tB AND tC AND tD AND tE`
where the sorted doclists of `tA,tB,tC,tD,tE` are of length $12000, 80, 3000, 40, 1000$ respectively. What is the total number of operations performed? Explain and express the answer in terms of the length of the doclists.

Problem 3. (10 points) **Data Center**

This is the Exercise from the last page of Subject 5. NSA built a 1,000,000sqf Utah Data Center at a cost of $1.5billion, supported by a 65-MW electrical substation (Wired, 2012/3/15, James Bamford, URL in L9 in Section C of the course web-page). Some other information is available in (Forbes, 2013/07/24, Kashmir Hill, URL in L9 in Section C of the course web-page). Based on the information of those two articles and the information in Subject 5 (read it carefully and USE it), estimate the MAXIMUM SERVER NUMBER of that facility by providing and JUSTIFYING TWO ARGUMENTS that roughly end-up to the same conclusion. Your number should be a multiple of 10,000. (Thus ignore article estimates and do not write something like 24,567 servers!) Make sure your answer is kept brief, no more than a paragraph per argument. Use 2016 data (256GiB RAM per server, 2HDDx5TiB each=10TiB).

Hint. An incorrect argument is for example the following: 1,000,000 servers each one having 256GiB RAM and 10HDDx10TiB. First I am telling you to use 2HDDx5TiB, and second your configuration at 400W conservative consumption per server, and no network switch connectivity or PUE considerations would need way too much more energy than the one available.

Problem 4. (20 points)

```
Document 1 (docID 1):
A sentence is a group of words. Sentences have a subject and a verb and maybe an object.

Document 2 (docID 2):
In a sentence find the verb and then find the subject.

Document 3 (docID 3):
To find the subject ask who or what followed by the verb.

Document 4 (docID 4):
To find the object ask  'subject verb who or what'?
```

Using the following stopword list (last line) for the four documents shown above and after case-folding and stemming (plural), show the form of the occurrence lists if (a) Doclist is used, (b) Counts is used, (c) Positions is used. For Positions what would an implementation of vocabulary and the inverted list table look like ?
`an a and by have in is of or the then to who what`,

Problem 5. (12 POINTS)

(a) A cluster of 4400 servers is to be used for Web Searching (configuration of Problem 1, with 2 HDD and RAM shown there). Using Google 2003 (Barroso paper) material for types of servers, and the discussion in class about index size how many machines would you use for Doc Servers, Index Servers, Cache servers, Web servers, Ad-servers, Spell-check Servers and DNS servers? Justify your answer. If you need a compression factor for the corpus use a 2.5; if you need a size for the index use the size of the compressed corpus! Make sure copies are multiples of 3 (per Barroso paper). Assuming 90% of queries hit a cache server and can be satisfied within 0.25seconds/query how many queries can your configuration deal with? Justify your arguments.

(b) With reference to the Barroso (2003) paper, in 2003 (based on the paper) did Google store compressed or uncompressed Web pages (data)?

(c) With reference to the Barroso (2003) paper, in 2003 did Google store the index itself compressed or not?

If you can't collect the 67 points from the previous problems, this is a bonus problem. But note you can still only get 67 points out of this homework.

Problem 6. (20 POINTS)

Using the cluster of 4,400 servers of Problem 5, how would you interconnect them using two types of switches. A TYPE-A switch has 36ports that operate at 10Gbits each, plus 4 uplink ports at 40Gbits. A TYPE-B switch ahs 36 ports that operate at 40Gbits each. To keep things simple, you are allowed to use a rack of say 36 servers (think of using two racks of 18 servers and a single switch is connected to all servers of those two racks).

(a) How many levels of switching would your configuration have? In class the simple TOR-EOR configuration had one level of switching: a TOR layer-1 connected to an EOR layer-2. The Pod configuration had two levels of switching: the TOR layer-1 connecting to the aggregate layer-2, and then the aggregate layer-2 connecting to the core switches of layer-3.

(b) What type of a switch are you going to use for a TOR switch? How many do you need? How many racks of about 36 servers are you going to need?

(c) How would you organize the second layer (not level) or more of switches? Keep things symmetric.

(d) For each layer of switches, given number of downlink ports, number of uplink ports and number of unused ports.

(e) What is the TYPE-A switch total? What is the TYPE-B switch total?

Hint and Note. For this problem treat all servers the same and do not try to build a rack out of index-servers or a rack only of cache-servers.

Hint. Use 3 layers of switches and group racks into pods of 8. Try not to use more than 180 switches. Make sure that the racks is a power of 2.

Date Edited: 11/04/2016