CS 345: Homework 5 (Due: Dec 02, 2014)

Rules. Hardcopies no later than start of class the day they are due; electronic copies by NOON-time the same day.

```
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'?
```

Problem 1. (5 points) **(This is a followup of Problem 1 of 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 (mid november) and the due time (early december).

Problem 2. (10 points)
(a) Provide v-codes for $74,260$ in hexadecimal.
(b) You are given an inverted list for a term t that looks like

`(1,1)(1,5)(1,11)(1,15)(1,17)(1,20)(2,10)(2,18)(2,30)(2,40)(3,5)(3,11)(3,25)(4,90)(4,91)`

Use the discussion of Subject 4 or 7 to v-byte encode this inverted list by providing the following information. (Note that v-byte encoding is also discussed on page 150 of the textbook. However, the example of the textbook incorrectly applies gap encoding to a count field shown here in bold-face, contradicting the algorithm stated there or in the notes: *1,2,1,6,1,***2***, 6, 11, etc.*)
(b1) Give the flattened (no parenthesis) form of the to be v-byte encoded list just before the v-byte encoding is applied when the list is flattened but still in decimal notation, and
(b2) after the v-byte encoding has been applied and the list given in hexadecimal notation.

Problem 3. (10 points)

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

Problem 4. (10 points)
Use the best approach possible to evaluate a query
`tA AND tB AND tC AND tD AND tE`
where the doclists of `tA,tB,tC,tD,tE` are of length $10000, 40, 2000, 20, 1000$ respectively. What is the total number of operations performed? Explain and express the answer in terms of the length of the doclists.

Introduction to Problem 5. This problem involves the design of 2003 search technology using 2014 servers/switches. The problem is split into Problem 5 and Problem 6. Read again the `Google's Search Engine Architecture` for data statistics, and also the paper on the **Google Cluster Architecture** by Barroso, Dean, and Holzle. **Your answers must be well-documented based on the papers and notes and quantitavely-justified.**

Use the 2014 PC configuration (PC14 from now on) of HW5 (16GiB, 4TiB). Read HW5 solutions. Do not power overload racks (10-15KW max). Assume you have 40-port switches only at 1GigE. Use round numbers for estimates or racks and switched. 24 or 18 or 32 are not round numbers to do calculations. If you use racks of 10 PC14, to accommodate 42TiB of disc-space you don't need 1.1 racks (11 machines). You need 2 with auxiliary free space. Disk-drive performance falls for disk 85-90% full.

Step 1. You need to collect by estimating certain statistics of your search infrastructure: docs, index, and servers and switches.

Step 2. You need to design a network supporting 2000 PC14, and 2 billion documents. Assume a doc is 1000kB, but indexable text content is only 100kB.

Step 3. You need to allocate resources on them. Allocation is at rack-level. You may need two types of racks: mainly servers, mainly switches.

Step 4. Estimate total number of servers, switches, queries per second.

If the number in Step 2 sound small, one could multiply by 10. But don't!

Problem 5. (11 POINTS) **(Cluster building : Part 1)**

Q1 (2 points). Use numbers from Google (1998) for initial estimates (Step 1), then finalize for the data (in Step 2). (Round numbers up to the next multiple of 10 ; i.e. a 24M becomes a 30M, 9.7GB becomes 10GB.) We provide answers for 1,1,1.4.

1.1 Number of documents indexed in corpus, **Answer: 24M becomes 30M.**

1.2 Size of uncompressed and compressed corpus text content,

1.3 Index size (including lexicon,anchors,links,Doc-Index),

1.4 Doc-index size **Answer:** $10GB$ **in doc-index rounding up 9.7GB**

Q2 (3 points). Now for the 2G document Corpus (Step 2), answer the following.

2.1 Number of documents indexed in corpus, **Answer: 2G**

2.2 Size of uncompressed and compressed corpus (use the ratio from 1.2 above),

2.3 Size of uncompressed and compressed corpus text content,

2.4 Index size (includes lexicon,anchors,links,Doc-Index), if you use 2.3 and 1.2 vs 1.3 data.

Q3 (6 points). Build cluster.

3.1 Primarily server rack: number of servers per rack.

3.2 Primarily switch rack: number of EoR, aggregate, or core switches per rack.

3.3 Rack power consumption for each rack type of 3.1, 3.2 (rough estimate).

3.4 Switch topology for ToR switches, EoR switches or any other type you need.

3.5 Number of server racks and total equipment used for them. Make sure total server size is 2000.

3.6 Number of switch racks and total switch equipment used.

3.7 Grand total server size and switches **Answer : Server total is 2000 and 264 switches!**

Problem 6. (14 POINTS) **Cluster building : Part 2**

Assume you don't have access to the GFS and thus you are responsible for data replication. Make sure your number are compatible with Google (1998) data. Cite them (eg Figure 72, line 5, column 2).!

Q4 (4 points). Use the data from the previous problem to answer the following questions. Some of them might have been answered in Q3 above. We name the first question Q4.

4.1 Corpus Doc size compressed and Index size for 2G documents from 2.2/2.4.

4.2 Degree of Replication for Corpus Docs compressed and Index.

4.3 Number of server racks per 1 copy of Corpus Docs and per 1 copy of Index.

4.4 Number of doc-server racks for all copies of Corpus Docs and index-server racks for all copies of Index.

Q5 (4 points).

5.1 How many server racks to be used as Web servers and as cache servers? (Use data from Q4.)

5.2 How many queries can be satisfied by the cache servers per second? (Assume that 5 cache servers, can collectively answer a query in 0.25 wall-clock seconds.)

5.3 Collect data and write number of rack per type (doc-server, index-servers, etc) utilizing 4.4 and 5.1 and 5.2,

5.4 How many queries can the cluster support per second? (Combine 5.2 with the capabilities of Web-server and Index-Servers; make sure Doc-server can keep up.)

Q6 (6 points).

These two questions refer to the Barroso (2003) paper.

(6.1) (3 points) In 2003 (based on the paper) did Google store compressed or uncompressed Web data?

(6.2) (3 points) In 2003 (based on the paper) did Google store the index itself compressed or not?

Problem 7. (30 POINTS) **30 More Bonus Points over the 60 point total**

Apply the HITS algorithm and PageRank algorithm on the following graph, and determine the ranking values (authority/hub for former, and rank for latter). Iterate as many times as needed for the error to be less than 10^{-3}. Use the synchronous update version. (This problem must be received by the deadline. No extension.)

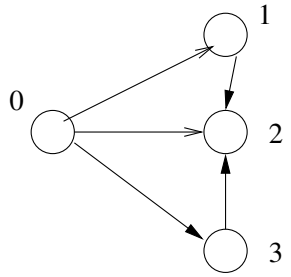

Figure 1: Graph for HITS and PageRank computation.

Date Posted: 11/18/2014