CS 345: Homework 4 (Due (New date!): Nov 23, 2015)

Choice. Drop two of the 12-point problems or collect 24 more points for a maximum of 90!

Problem 1. (12 points) **NSA Data Center**

This is the Exercise from the last page of Subject 5. NSA (National Security Agency) has 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, estimate the server size of that facility and amount of information that can be maintained. (Do not use the article estimates.) Make sure your answer is kept brief (no more than a paragraph). Use 2015 data (eg 64GiB RAM per server, 10TiB per HDD). **(You need to present two different arguments that reach the same conclusion and are consistent with the information listed above. Justify your argument based on the available documentation listed above.)**

Hint. An incorrect argument is for example the following: 1,000,000 servers each one having 64GiB RAM and 10 10TiB HDDs. It is unlikely the power consumption of such a server is going to be 65W only and will need no air-conditioning or other cooling.

Problem 2. (6 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 3. (12 points) **Heaps' Law Revisited**

We revisit Heaps' Law (read solutions of a HW3 problem to avoid making the same mistakes). In 2015, Google has approximately 25G documents, with average text content 110KiB (roughly 120,000B). How many distinct words does Heaps' Law imply. Justify your claims. Round to the closest million or multiple of ten-million as needed (i.e. do not dare write 123,456,789 or 12,345,567 but 130,000,000 or 13,000,000).

How does this compare to the 14,000,000 figure of Google in 1998? Can you explain the discrepancy? ("The English language is 10 times richer in words is not an explanation of interest".)

Problem 4. (12 points) **Huffman and more...**

The text below of 25 characters A, C, G, T, W is first encoded in a byte-aligned ASCII code, and then using Huffman coding. (Ignoring spaces that are provided for readability only), what are the prefix codes for each one of the five characters under Huffman coding, and what is the total number of bits used to encode just the text (consisting of A, C, G, T, W and ignoring the spaces or other auxiliary information)? How does this compare to the ASCII encoding (express bit and byte SAVINGS as a percentage).

```
TTCGA AATTT WCCGA ATTCC WTCGC
```

What is the entropy of the text? Compare it to the average number of bits per character in the Huffman coding.

Problem 5. (12 POINTS)

Provide Elias-γ, Elias-δ and Golomb-7 codes for $k = 17$. Which one is shorter?

Problem 6. (12 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:
`an a and by have in is of or the then to who what`,
for the four documents shown above and case-folding, 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 7. (12 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 $12000, 80, 3000, 40, 1010$ respectively. What is the total number of operations performed? Explain and express the answer in terms of the length of the doclists.

Problem 8. (12 POINTS)

(a) A cluster of 4000 computers is to be used for Web Searching. A rack can contain a maximum of 20 servers (configuration in Problem 1, with 2 HDD and RAM shown there). Using Google 2003 (Barroso paper), how many machines are Doc Servers, Index Servers, Cache servers, and other (Web servers, Add servers, Spell-check Servers). Justify your answer. Assuming 90% of queries hit a cache server and can be satisfied by it within 0.25seconds 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 data?

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

Date Edited: 10/21/2015