A. V. GERBESSIOTIS | CS 345-001 | Fall 2013
HW 4 | OCT 28, 2013 | 90 POINTS

CS 345: Homework 4 (Due: Nov 18, 2013 before 10:01am)

Rules. This is to be completed no later than the start of class on the day it is due. Earlier submission is also possible by email (read Handout 1).

Problem 1. (18 POINTS) **(This is Exercise 2 of Subject 5)**

Figure 1: Be careful with Problem 1:-):-)

NSA (National Security Agency) is building 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 and also `http://nsa.gov1.info/utah-data-center/`).
Based on those two articles (for links go to the protected area of the course Web-page) and pieces of information, estimate the server size of that facility and amount of information that can be maintained. Make sure your answer is kept brief (no more than a paragraph). Use 2013 data for typical PC configurations. (Two varying arguments that reach the same conclusion are better than one.)

Problem 2. (18 POINTS)
(a) Is the state of Google's dictionary (lexicon) in 1997/1998 (use paper L1 and Subject 2 for reference) consistent with Heaps' Law i.e. how many words does the lexicon maintain and how many unique words does Heaps' Law predict about the 147GB or so of the document collection? Justify your claims.
(b) Using an 80KB for the text size of a document in 2013, and a corpus of 25billion documents for Google, how many distinct words does Heaps' Law imply? Verify its consistency and justify your claims.

Problem 3. (18 POINTS)
Provide Elias-γ, Elias-δ and Golomb-11 codes for $k = 50$.

Problem 4. (18 POINTS)
The text below of 20 characters A, C, G, T 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 four characters under Huffman coding, and what is the total number of bits used to encode just the text (consisting of A, C, G, T and ignoring the spaces or other auxiliary information)? How does this compare to the ASCII encoding (express bit and byte savings as a percentage).

```
ACGT TAAA CCGT AACC ACGC
```

Problem 5 is on the next page

Problem 5. (18 POINTS) **Paper Google Cluster architecture by Barroso, Dean, and Holzle**

This problem involves the design of a (rather obsolete) cluster using 2003 technology. There will be a continuation in Homework 5.

Read the paper on the **Google Cluster Architecture** by Barroso, Dean, and Holzle. A link to the paper is available in section C5 link L2 of the Web-page or the corresponding Subject. Try to answer after reading the paper the following questions whose answers are in the paper or may need to be further thought upon; we provide some answers to some of the questions to guide you in what is expected from you. Your answers must be well-documented based on textbook, notes, or paper-based (C5/L2) material and quantitavely-justified.

Q1 (5 points). Citing the 1998 (not 2003) paper, what are the Google (1998) statistics for the following questions. (Round numbers up to the next multiple of 10 for (1.2) through (1.4), and to the next integer for 1.5.)

1.1 Number of documents indexed,

1.2 Size of uncompressed and compressed text,

1.3 Index size (includes lexicon and anchor-related data),

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

1.5 Links size used for PageRank computations.

Q2 (6 points). From the 2003 paper, what are the answers to the following questions. (Assume a power consumption of 200W per machine, and use the fastest machine configuration mentioned in the paper/Subject 5; for rack configuration use page 14, NOT page 15 information.)

2.1 Maximum number of machines per side of a rack (rack-side) and per rack,

2.2 Footprint of a rack, and power consumption (per rack rather than individual machine),

2.3 Switch connections per rack side (or rack) assuming one machine one network card.

Q3 (3 points). Related to Question 2, answer the following questions (Use the most powerful configuration quoted in the paper as instructed also in 1.2.)

3.1 How much main memory RAM per rack (in GBytes). **Answer: 80 machines with 2GB per machine is 160GB/rack.**

3.2 Disk space (in GBytes or TBytes) per rack.

Q4 (4 points). Answer the following questions. (These questions is for a single cluster of roughly 2000 machines.)

4.1 Are raw-page data stored redundantly and if so how many times.

4.2 Is the index stored redundantly and if so how many times?