A. V. Gerbessiotis CS 345-001 Fall 2014
HW 4 Oct 21, 2014 60 points

CS 345: Homework 4 (Due: Nov 4, 2014)

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

Problem 1. (15 points) **(NSA's Bluffdale, UT, Data Center)**

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

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 2014 data for typical PC configurations (eg 16GiB RAM, 4TiB disk space). **(You need to present two different arguments that reach the same conclusion. Justify your argument based on the available documentation listed above.)**

Problem 2. (15 points) **(Heaps' Law)**
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? Subject 3, page 11 provides an estimate which is close to the number quoted in the Google paper.
Using an 100KiB for the text size of a document in 2014, 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. (15 points)

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

Problem 4. (15 points)
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).

```
AACGT TTAAA WCCGT TAACC WACGC
```

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

Date Posted: 10/21/2014