Hand in problem 2 in Information Theory (EIT 080)

VT 1, 2012

Problem 2

In this problem you should estimate the probability function for the letters included in the file LifeOnMars.txt and construct an optimal source code for it. The code is used to compress the file and the average codeword length is compared with the expected codeword length for the code and the entropy for the estimated source.

Source

In the file LifeOnMars.txt you will find the lyrics of the tune *Life on Mars?*, by David Bowie from the album *Hunky Dory*, 1971. Use the file and estimate the source probability function for the letters included, i.e.

(a, b, . . ., z, ', [space], [new line]).

Hint In the ASCII table, the characters above correspond to the numbers

(97, 98, . . ., 122, 39, 32, 10).

To get the ASCII number for a character in MATLAB, the command

```
> cast(c,'uint8')
```

can be used. You can import the file as a string (vector of characters) into MATLAB with e.g.

```
> fid = fopen('LifeOnMars.txt');
> Txt = fscanf(fid,'%c');
> fclose(fid);
```

Optimal code

Construct an optimal binary source code for the estimated probability distribution above. What is the total length of the encoded text in the file LifeOnMars.txt?

Compare

- The average number of code bits per source symbol for the file.
- The expected number of bits per source symbol for the estimated probability function.
- The entropy for the estimated probability function.

Finally, consider the distribution of 0s and 1s in the encoded sequence. What is the corresponding entropy?

Hand in details

The problem can be solved by computer scripts (e.g. MATLAB) and/or derivations on paper. The solutions you hand in should show all steps, and the code for used scripts should be handed in as appendix to the solution. (That is, it is not enough with the code).

Hand in to `adnan.prlja@eit.lth.se` and/or `stefan.host@eit.lth.se`, preferably electronically. Do not forget to write your name and STIL or student ID.