Hand in problem 2 in Information Theory (EITN45)

VT 2, 2017

Problem

In this problem you should estimate the probability function for the letters included in the file LifeOnMars.txt and construct an optimal source code based on this estimation. Then use the code to compress the text, and compare the average codeword length with the entropy of the estimated distribution.

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.

$$\Big\{\mathsf{a}, \mathsf{b}, \ldots, \mathsf{z}, \text{'}, [\text{space}], [\text{new line}]\Big\}.$$

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

$$\Big\{97, 98, \ldots, 122, 39, 32, 10\Big\}$$

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

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

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);
```

Then you have the letters as a vector of characters.

Optimal code

Construct an optimal binary source code for the estimated probability distribution above. Derive the total length of the encoded text in the file LifeOnMars.txt, and compare with the uncoded case (compression ratio)? For the uncoded case assume that each letter is coded with 8 bits.

Compare

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

Finally, consider the distribution of 0s and 1s in the encoded sequence.

Hand in details

You should hand in your solution of the problem in paper format. The line of reasoning should be clear from the solution and it should contain

- Distribution for the letters
- Code table and comparisons above
- If you use computer to solve parts of the problem you should also hand in the code for these scripts as appendix. (It is not intended that normal computer aided derivations, like inline calculations, should be handed in. But if you for example write a script that finds the distribution of letters, this is part of the solution and should be handed in as code list.)

You can hand in your solution either written on computer or hand written, or as a combination. The important thing is that it explains in a clear way how you have solved the problem. Used scripts should be handed in as appendix to the solution.

The solutions are handed in individually to Stefan or Umar, or in the course mailbox at the third floor in the northern staircase from the main entrance in the E-building. Do not forget to write your name and STIL or student ID.

In Matlab there are functions `huffmanenco` and `huffmandict`. These are not allowed in your solution. If you solve the coding part using a computer you have to write the code yourself.