CS 6980: Introduction to Data Science © Spring 2018

Mini-Project 4: Data Science & Information Theory

Instructor: Daniel L. Pimentel-Alarcón Due 03/19/2018

One fundamental aspect of data science is quantifying how much *information* data contains. This is one of the main questions addressed by information theory [1]. In this mini-project you will use information theory to determine whether really *an image is worth more than a thousand words*.

To this end, we will compute the *entropy* (measure of information) of an entire book (containing more than $1,000$ words), and compare it against the entropy of an image.

(a) Download the text of a book of your choice. For example, I downloaded *The Count of Monte Cristo*, by Alexandre Dumas. Load your file, and compute the number of times that each character x_{i} appears. *Hint:* you can do this in three lines of code, using the Matlab functions `fileread` and `hist`.

(b) Estimate the probability $p(x_{\mathrm{i}})$ of each character.

(c) Compute the entropy of a character as:

$$H(x) \;=\; \sum_{x_{\mathrm{i}}} p(x_{\mathrm{i}}) \log_2 \left(\frac{1}{p(x_{\mathrm{i}})} \right), \tag{4.1}$$

(d) Your result from (c) tells you the information encoded in each character of the book. Now multiply this by the number of characters in the book, to obtain the overall entropy of the book.

Now we will compute the entropy of an image.

(e) Download an image of your choice. For example, I downloaded the image in Figure 4.1 below. Load your image, and compute the number of times that each pixel intensity x_{i} appears. *Hint:* you can also do this in three lines of code.

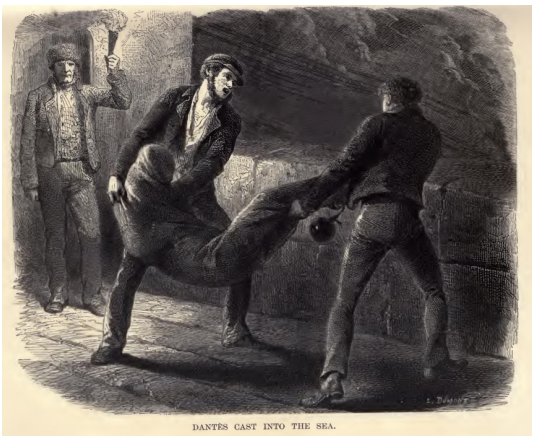

Figure 4.1: Image containing an illustration from *The Count of Monte Cristo*, depicting Edmond Dantès being thrown at the sea from the Château d'If.

(f) Estimate the probability $p(x_i)$ of each pixel intensity.

(g) Compute the entropy of a pixel as in (4.1):

(h) Your result from (g) tells you the information encoded in each pixel of the image. Now multiply this by the number of pixels in the image, to obtain the overall entropy of the book.

(i) Which contains more information, the book or the image?

(j) Do you think this is a fair comparison? Why, or why not?

References

[1] T. M. Cover and J. A. Thomas, *E*lements of Information Theory, John Wiley & Sons, Inc, 1991.