Exercise on Unicode and UTF-8

Jonas Sternisko, ADITION technologies

12.04.2018

Exercise 1: UTF-8 Encoding Scheme

Manually encode your name with UTF-8. The result should be a sequence ob binary octets. If your name does not contain any special characters, replace at least two characters with fancy forms. For example, "Jonas" can become "Jôñàŝ".

Exercise 2: UTF-8 Encoding Scheme

Write a function in a programming language of your choice which displays a unicode character as a sequence of bytes. Test it against the things you did manually. Try not to use any built-in functionality, but do it on your own.

Exercise 3:

Download the document from gitlab. Inspect it with `xxb` (or `xxb -b`). Can you guess the encoding of the document? Hint: It has a fixed size of 8 bits per word.

Write a function that reads a sequence of binary and outputs the corresponding UTF-8 character. Use the function to convert the document to UTF-8.

Now google for command line tools which could do this for you and try at least two.

Exercise 4: UTF-8 and URL-Encoding

Check Wikipedia for URL-encoding. First, reflect why we need URL-encoding at all and do not send UTF-8? Then write a function which reads a URL-encoded string and outputs the UTF-8 representation. Fetch the list of URL-encoded HTTP referrers from gitlab and use your function to conver the referrers to UTF-8 strings.