

Exercise Sheet 7

Submit until Tuesday, December 8 at **2:00pm**

Exercise 1 (5 points)

Copy your web application code from Exercise Sheet 6 (or from the master solutions) to a new folder *sheet-07*.

Adjust your web application to search on the input file *cities-eu.corrupted.txt* provided on the Wiki (all cities in Europe, with their population numbers). This file is UTF-8 encoded but has corrupted bytes.

Enhance your code by a method that repairs the records from the input file by replacing a small set of bytes such that all records are valid UTF-8. The problematic bytes should be replaced by a replacement character (Unicode U+FFFD). It is not necessary that you achieve the minimal number of changes possible (that might be a very hard problem), but avoid extreme / naive solutions like overwriting all bytes with a value < 128 (which would be valid UTF-8).

For example, in the input file the record in line 43 should become “Abáig✦r” and the record in line 256 “✦caja”.

Exercise 2 (10 points)

Enhance your web application such that it works properly with special characters, in particular German umlauts. In particular, it should work with the file *cities-eu.corrupted.txt* provided on the Wiki. For example, if you type *Munchen* or *Müchen*, you should get *München* as a completion.

Note 1: this requires that you properly decode and encode URLs in your communication between the web app and the server. For example, z%C3%BCrich should be decoded to zürich.

Note 2: in your server code, choose a string representation such that the edit distance between any two characters (for example, ü and u) is 1. In Python and Java, this is trivial when you use standard strings. In C++, you can use `std::wstring`.

Exercise 3 (5 points)

Add one or more features of your choice that make meaningful use of cookies. As a minimum, add a drop-down menu that let's the user choose the number of hits to be displayed in response to each query. Remember that choice in a cookie, and make sure that the next time the user visits the website (using the same browser) the number of hits from the last session is used.

Add your code to a new sub-directory *sheet-07* of your folder in the course SVN, and commit it. Please write an *experiences.txt*, especially those of you who haven't written one yet. Make sure to include at least one unicode character you have never seen before. If you have access to a machine / server that is permanently connected to the internet, feel free to add a link to your web app in the table on the Wiki.