Lab09

9.2 - 1.c

The string contains 29091 symbols and the bytearray has 30491 bytes. This has to do with how the UTF-8 interprets the Swedish “ÅÄÖ”. The “ÅÄÖ” in UTF-8 take 2 bytes each instead of 1 byte.

9.2 - 2.d

Using an optimal encoding we can not achieve a better compression with an entropy with less than 4.6287 bit/symbol without exploiting the statistical redundancy.

9.2 - 4.c

* The zip code is 19926 bytes long.
* The zip code is 159408 bits long.
* TheCopy contains 29091 symbols.
* The zip algorithm managed to compress theCopy to 7.981177 bit/symbol.

9.2 - 4.d

* The unshuffled byteArr got 7.974185931720011 bit/symbol when zipped.

9.2 - 4.e

* The data source’s entropy is the smallest one.
* The zlib-encoding of theCopy has the highest entropy.
* The compressed data in the zip requires less bits to store the same data like the data source, which leads to a higher entropy (bit/symbol) plus the compressed data is shuffled randomly which increases the entropy when compressed.

9.2 - 5.b

* The first string got compressed down to 69 bytes.
* The second string got compressed down to 79 bytes.

9.2 - 5.c

* The zip algorithm searches for patterns in the text before compressing it which allows it to effectively compress the data without the need to store repeated stuff.