System programming

General practice: UTF-8 encoding

Problem

Modern text files use UTF-8 coding of Unicode values to represent special characters. Each character (symbol) has a Unicode value, which is represented on varying length. In the file, some of the characters are just 1 byte, while others are 2 or 3 or 4 bytes. In this way the number of bytes of a text file is not the same as the numbers of characters stored in the file. The following table summarizes how can we identify the length of different coded characters:

code length of one character	bit pattern of the code
1-byte long code	`0???????`
2-bytes long code	`110????? ????????`
3-bytes long code	`1110???? ???????? ????????`
4-bytes long code	`11110??? ???????? ???????? ????????`

For example, the following 10 bytes represents only 4 characters:
11010110 10100011 01011100 11110100 10001000 10111010 10000101 11100100 10001110 10101100

Goal

A software is needed to tell how many characters (symbols) are stored in a file. The filename is given as a command-line argument.

Requirements and steps of the implementation

The filename is given as a command-line argument.

The program must check whether the given name is a filename or not. If the command-line argument does not refer to an existing file (e.g., misspelling or directory name) the program must return by 1.

The content of the file must read with only one instruction (not byte-by-byte). That is why the size of the file must read from the related i-node and then you must allocate the required size memory dynamically. This field is loaded by the file content using binary file handling. Close the file.

The first character of all codes must be analyzed by bitwise operations to find out the size of the code (the number of bytes). The size of the code tells where the first byte of the next code is. Count the number of characters (UTF-8 codes). Free the allocated memory.

Optional: Print the codes of each character in hexadecimal form to the screen separated by a space. (Hint: Each byte as unsigned char value in "`%02x`" format.)

Optional: Determine the length of words int the text. Word is a consecutive character sequence separated by white spaces (`' '`, `'\t'`, `'\n'`, `'\r'`)