Programming using Python

Assignment 1

Week 1 (13/10/2018 - 21/10/2018)

G) What is ASCII AND UTF-8?

1. ASCII
   - Stands for American Standard Code for Information Interchange.
   - It is a code for representing 128 English characters as numbers, with each letter assigned a number from 0 to 127. For example, the ASCII code for uppercase *M* is 77. Most computers use ASCII codes to represent text, which makes it possible to transfer data from one computer to another.
   - Text files stored in ASCII format are sometimes called ASCII files. Text editors and word processors are usually capable of storing data in ASCII format, although ASCII format is not always the default storage format. Most data files, particularly if they contain numeric data, are not stored in ASCII format. Executable programs are never stored in ASCII format.
   - The standard ASCII character set uses just 7 bits for each character. There are several larger character sets that use 8 bits, which gives them 128 additional characters. The extra characters are used to represent non-English characters, graphics symbols, and mathematical symbols.

2. UTF-8
   - Stands for Unicode Transformation Format. The '8' means it uses 8-bit blocks to represent a character. The number of blocks needed to represent a character varies from 1 to 4.
   - It was designed for backward compatibility with ASCII. Code points with lower numerical values, which tend to occur more frequently, are encoded using fewer bytes. The first 128 characters of Unicode, which correspond one-to-one with ASCII, are encoded using a single octet with the same binary value as

ASCII, so that valid ASCII text is valid UTF-8-encoded Unicode as well.

- UTF-8 is safe to use within most programming and document languages that interpret certain ASCII characters in a special way, such as "/" in filenames, "\" in escape sequences, and "%" in print
- The Unicode Standard has become a success and is implemented in HTML, XML, Java, JavaScript, E-mail, ASP, PHP, etc. The Unicode standard is also supported in many operating systems and all modern browsers.
- UTF-8 is a compromise character encoding that can be as compact as ASCII (if the file is just plain English text) but can also contain any unicode characters (with some increase in file size).