WHAT IS CHARACTER ENCODING IDENTIFICATION ALGORITHM? WHAT DOES IT DO? NAME SOME!

The process of heuristically guessing the character encoding of a series of bytes that represents text. This algorithm involves statistical analysis of byte patterns. This process is not fool proof as it depends on statistical data.

Character encoding is the process of representing individual characters using a corresponding encoding system made up of other symbols or data types.

The modern use of character encoding is done according to specific computer design principles. One of these is that computers don't recognise linguistic text characters as such, but use them as data types. These data types are stored in binary at machine level, as sets of ones and zeros.

There are three writing systems- syllabic, alphabetic and logographic.

7-bit ASCII was used initially. But it could represent only 27=128 characters. This was not sufficient for all the characters of all languages. Also, romanisation or ASCIIfication was implemented. Romanisation or ASCIIfication is the conversion of writing from a different writing system to the Roman script. Methods of romanization include transliteration, for representing written text, and transcription, for representing the spoken word, and combinations of both.

Through this, ASCII equivalents were defined for characters that are not defined in the character set.

But 7-bit ASCII encoding was not sufficient for all characters of all languages. So, the 8-bit encoding came into existence. It followed alphabetic and syllabic writing systems. Also, it resulted in large number of overlapping character sets for different languages.