**NPL Homework**

**9 Dec 2019**

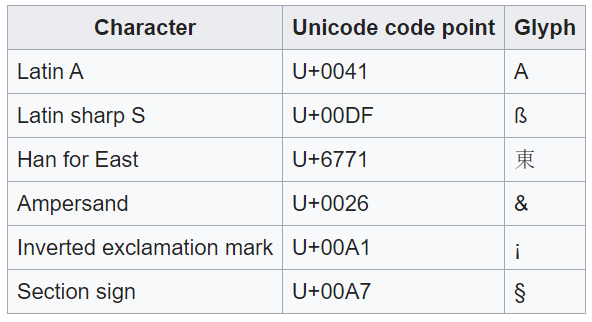
Q) What is a character encoding identification system? What are the available types?

A)

Each character to be displayed or used on a computer need to have a specific representation such that it can be understood by a computer system. Since all data is represented in binary, these characters as well need to have a unique binary representation. For this purpose, each character is given a unique numeric value. This value is then represented in binary and used by the computer.

A character encoding is the way the characters are converted in to binary. It defines the unique value of each character. It is basically a representation format for characters.

Example -



Types of character encodings:

1. *ASCII-*

The ASCII (American Standard Code for Information Interchange) character set is one of the most common ones used today for english. Most systems use this along with other character sets. It contains 128 characters and gives a 7 bit binary representation for each character.

1. *Unicode-*

Unicode is the computing standard all over the world and it is used in most writing systems all over the world. It contains representations of 137,994 characters and has a 1 – 4 byte representation.

Unicode can be implemented by different character encodings. The Unicode standard defines UTF-8, UTF-16, and UTF-32, and several other encodings are in use. The most commonly used encodings are UTF-8, UTF-16, and UCS-2.

1. *ISO/IEC 8859 -*

ISO/IEC 8859 is a joint ISO and IEC series of standards for 8-bit character encodings. The series of standards consists of numbered parts, such as ISO/IEC 8859-1, ISO/IEC 8859-2, etc. There are 15 parts, excluding the abandoned ISO/IEC 8859-12. The ISO working group maintaining this series of standards has been disbanded

1. *Extended Binary Coded Decimal Interchange Code(EBCDIC) -*

Extended Binary Coded Decimal Interchange Code(EBCDIC) is an eight-bit character encoding used mainly on IBM mainframe and IBM midrange computer operating systems. It descended from the code used with punched cards and the corresponding six-bit binary-coded decimal code used with most of IBM's computer peripherals of the late 1950s and early 1960s

1. *Mac OS Roman -*

Mac OS Roman is a character encoding primarily used by the classic Mac OS to represent text. It encodes 256 characters, the first 128 of which are identical to ASCII, with the remaining characters including mathematical symbols, diacritics, and additional punctuation marks. It is suitable for English and several other Western languages. Mac OS Roman is a superset of the original Macintosh character set, used in System 1.

1. *Indian Script Code for Information Interchange (ISCII) –*

Indian Script Code for Information Interchange (ISCII) is a coding scheme for representing various writing systems of India. It encodes the main Indic scripts and a Roman transliteration. The supported scripts are: Assamese, Bengali (Bangla), Devanagari, Gujarati, Gurmukhi, Kannada, Malayalam, Oriya, Tamil, and Telugu. ISCII does not encode the writing systems of India based on Persian, but its writing system switching codes nonetheless provide for Kashmiri, Sindhi, Urdu, Persian, Pashto and Arabic. The Persian-based writing systems were subsequently encoded in the PASCII encoding.

1. *Tamil Script Code for Information Interchange (TSCII) -*

Tamil Script Code for Information Interchange (TSCII) is a coding scheme for representing the Tamil script. The lower 128 codepoints are plain ASCII, the upper 128 codepoints are TSCII-specific. After long years of being used on the Internet by private agreement only, it was successfully registered with the IANA in 2007.

1. *JIS X 0208 -*

JIS X 0208 is a 2-byte character set specified as a Japanese Industrial Standard, containing 6879 graphic characters suitable for writing text, place names, personal names, and so forth in the Japanese language. The official title of the current standard is 7-bit and 8-bit double byte coded KANJI sets for information interchange. It was originally established as JIS C 6226 in 1978, and has been revised in 1983, 1990, and 1997. It is also called Code page 952 by IBM. The 1978 version is also called Code page 955 by IBM.

1. *GB 2312 -*

GB/T 2312-1980 is a key official character set of the People's Republic of China, used for simplified Chinese characters. GB2312 is the registered internet name for EUC-CN, which is its usual encoded form. GB abbreviates Guojia Biaozhun, which means national standard in Chinese. GB2312 (1980) has been superseded by GBK and GB18030, which include additional characters, but GB2312 remains in widespread use as a subset of those encodings.

1. *ANSEL -*

ANSEL, the American National Standard for Extended Latin Alphabet Coded Character Set for Bibliographic Use, was a character set used in text encoding. It provided a table of coded values for the representation of characters of the extended Latin alphabet in machine-readable form for thirty-five languages written in the Latin alphabet and for fifty-one romanized languages. The standard was reaffirmed in 2003 although it has been administratively withdrawn by ANSI effective 14 February 2013. It is registered as Registration # 231 in the ISO International Register of Coded Character Sets to be Used with Escape Sequences.