プログラミング技法 予習課題11

IKEDA Kaito

2021/5/31

1 「文字コードセット」とその「エンコーディング法」について，両者の違いが分かるように説明してください.

- 文字コードセットは，1 つ以上の言語の特定のニーズに基づいてあらかじめ定義された文字の集まりであり，どのコードセットを使用するかの選択は，ユーザのデータ処理要件に依存する.
- それに対し，エンコーディング法とは，コード化された文字からエンコードされた文字体系を生成するために制定された業界ルールである．つまり，文字セットにコーディングされたエンコーディング法を適用することで，エンコーディングができる.

2 C 言語で strlen は，一般的には文字列の長さを返す関数とされている．この関数に漢字文字列を与えた際の挙動について説明してください．例えば，strlen(“日本語”) としたときどういう結果を得ますか. それは何故ですか.

- まず結論から述べると，私の環境で実行した出力結果は 9 であった.
- そして，私が実行した環境では UTF-8 を採用していた.
- 次にインターネットで日本語を表せる文字コードについて調べてみると，SHIFT_JIS・UTF-8 あたりがメジャーだということを知り，2 つの文字コードについて具体的な調査を行った.

2.1 SHIFT_JIS

- SHIFT_JIS の文字列には，1 バイト文字と 2 バイト文字が混在していることがわかった.
- また，SHIFT_JIS の文字列の中で，ASCII 文字・半角カナ・制御コードが 1 バイト文字であり，その他（今回の漢字含む）文字が 2 バイト文字である.
- よって，SHIFT_JIS の環境であった場合，出力結果は 6 とでていたと考察できる.
- ファイルを SHIFT_JIS で作成し直し，実行した結果 6 が出力された.

2.2 UTF-8

- UTF-8 は，文字により使用するバイト数が異なるマルチバイト文字であることがわかった.
- また，1 バイト文字から 6 バイト文字までの幅があるが，しかし現在は 5 バイト文字・6 バイト文字は存在しないので，最大で 4 バイト文字となる.
- そして，UTF-8 の文字列の中で，半角英数字などは 1 バイト文字であり，日本語は主に 3 バイト文字で表現されることがわかった.

- よって，UTF-8 の環境であった場合，出力結果は 9 とでていたと考察できる．
- 最初に実行したファイルでは，UTF-8 を採用しており，出力結果は 9 であった．

上記内容を概括すると，文字コードによって出力結果が異なるということがわかった．

3 たまに，Web ブラウザで Web ページを見ているときに文字化けという現象が起こります．文字化けが起こる理由と文字化けを起こさないように Web ページを作るにはどうしたらいいですか．

以下 2 点を徹底することで，文字化けを回避することができる．

- HTML ファイルは，文字コード UTF-8 で保存する．
- HTML の meta タグに charset="UTF-8"を指定する．

もちろん，UTF-8 でないと文字化けするということではないが，最近の HTML ファイルはほとんど UTF-8 で作成されており，また近年の GoogleChrome では UTF-8 ページを前提としているため，エンコード設定を変更するメニューが消えつつあることを踏まえると，UTF-8 で記述するのが望ましいと考える．また，近年の VSCode 等のエディタでは，ファイル生成時に自動的に charset="UTF-8"が記述されるようになっている．