

C# でテキストファイル文字コードを自動判定する

アルゴリズム

- ・ ASCII 制御コードのうち 0x00-0x03、ないし 0x7F(DEL) が出現した場合
 - ・ 原則として非テキストファイルとみなす
 - ・ ただしファイル先頭 2 バイトで 0x00 が登場した場合は、BOM なし UTF16 の可能性を調べる
- ・ 非 ASCII コード (0x80 以降) が出現しなかった場合
 - ・ JIS エスケープシーケンスがあれば JIS、なければ ASCII
- ・ 非 ASCII コード (0x80 以降) が出現した場合
 - ・ 以下 4 種類の文字コードに該当するか、可能性を調査する
 - ・ ANSI (CP1252)
 - ・ BOM なし UTF8 (CP65001)
 - ・ EUC (CP51932、補助漢字使用時は CP20932 相当)
 - ・ 登場するコード範囲の関係で (必ず 0x80 以上となる) 上記 3 種はまとめてチェック
 - ・ ShiftJIS (CP932)
 - ・ 2 バイト目が 0x20-0x7E の ASCII になることもあるので、個別にチェック
 - ・ 文字コード体系として明らかに妥当ではない場合、可能性なしとみなす
 - ・ 半角文字 (半角カナ)・全角文字が連続していれば可能性が高いとみなす
 - ・ 逆に不連続であれば可能性は低いものとみなす
 - ・ 誤判別を起こしやすい箇所は個々に配点をチューニング
 - ・ (過去の誤判別事例などをフィードバックしています)
 - ・ 全バイト走査完了後、可能性が高いものから順にテキスト取り出しを試みる
 - ・ デコードエラーが発生しなければその文字コードで確定

ソースコード抜粋：文字コード判別クラス・判別メソッド本体

ソースコード抜粋：BOM なし UTF16 可能性判定メソッド / ShiftJIS 判定メソッド

ソースコード抜粋：文字コード定義クラス・および BOM つき UTF 判別メソッド