〈文書比較基本仕様書〉

本書は、日本語の文書比較機能の基本仕様について纏めるものである。比較する２つの文書において、比較元を文書A、比較先を文書Bとする。すなわち、文書Aが文書Bに編集されたと仮定して、比較処理を行う仕様である。

- 本機能は、エクセルに入力された２つの文字列(文書)の相違を検出するものである。今回は、英文、やプログラムコードなどの、日本語文章以外には対応しておりません。
- 文書を、単語区切りに管理し、比較処理を実装する仕様とする。本書では、同一文字種の文字が続いたものを、ひとくくりの単語と定義する。文字種とは、漢字、片仮名、数字、英字、記号、制御コードのことである。また、分割された単語には、文書の先頭から、インデックスを振っていき、管理するものとする。

  Ex. 私は、昨日、3回ほどパソコンを起動した。

  これを定義に基づき、単語に分割すると、以下となる。

  [私][は][、][昨日][、][3][回][ほど][パソコン][を][起動][した][。]

  インデックスを振り、プログラム内部では、以下のように管理する。

  #0[私] #1[は] #2[、] #3[昨日] #4[、]

  #5[3] #6[回] #7[ほど] #8[パソコン] #9[を] #10[起動] #11[した] #12[。]

- 文書Aと文書Bを比較したDIFF結果には、以下の5通りがあるものとする。文書の相違が【移動】の場合は、移動位置を表すために、グループ･インデックスを設定する仕様とする。
  - **同じ** … 文書Aと文書Bで一致する文言。
  - **挿入** … 文書Aには存在しなかったが、文書Bで追加挿入された文言。

    文書A: 私は、パソコンを起動した。

    文書B: 私は、昨日、パソコンを起動した。

    比較結果: 私は、(+ 昨日、) パソコンを起動した。
  - **削除** … 文書Aには存在したが、文書Bで削除された文言。

    文書A: 私は、昨日、パソコンを起動した。

    文書B: 私は、パソコンを起動した。

    比較結果: 私は、(- 昨日、) パソコンを起動した。
  - **変更** … 文書Aの文言が、文書Bのデータに変更された文言。

    文書A: 私は、昨日、パソコンを起動した。

    文書B: 私は、おとどい、パソコンを起動した。

    比較結果: 私は、(* 昨日 → おとどい)、 パソコンを起動した。
  - **移動** … 文書Aのある文言の位置が、文書Bにおいて、他の文言の前後に移動。

    文書A: 私は、パソコンを、昨日、起動した。

    文書B: 私は、昨日、パソコンを、起動した。

    比較結果: #0 私は、 #6 昨日、 #3 パソコンを、 #8 起動した。
- グループ・インデックスは、DIFF結果の【同じ】【挿入】【削除】【変更】【移動】と判定された文言の先頭の単語の**文書Aにおける単語インデックス**である。ただし、**【挿入】の場合は、文書Aに該当する単語が存在しないので、-1**を設定する。
- 文書AとBの、一致率を算出できるものとする。尚、一致率の計算方法については、後述する。

<比較の最小単位である単語について>

文書比較は、文書を**同一文字種の一括りの単語**として分解し、その分解された単語を比較の最小単位として、実行する仕様とする。

例えば、以下のような文書 A と B があったとする。

文書 A: 私は、ノートに ABC と書いた。

文書 B: 私は、ノートに ABCDEF と書いた。

上記を単語単位(同一文字種)に分割すると、それぞれ以下となる。

文書 A: [私][は][、][ノート][に][ABC][と][書][いた][。]

文書 B: [私][は][、][ノート][に][ABCDEF][と][書][いた][。]

これを、比較の最小単位を、単語として、比較すると、

”ABC”が ”ABCDEF”に【変更】されたと、判定される仕様である。ですので、”DEF”が【追加】されたとは判定しない仕様である。

<一致率>

一致率とは、文書 A と文書 B が、どの程度、同じ文書であるかを示す指標のことである。文書 A を基準としたときの、文書 B との差分(相違の文字数)を差し引いた割合である。

一致率の計算は、簡単には、以下の計算式により算出する。

一致率 ＝ (文書 A の文字数 - 相違の文字数) ÷ 文書 A の文字数

相違の文字数は、以下を合計したものとする。

- 【削除】判定となった、文書 A の文字数。
- 【挿入】判定となった、文書 B の文字数。
- 【変更】判定となった、文書 A の文字数と文書 B の、文字数の大きい方の文字数。
- 【移動】判定となった、文書 A の入れ替わった文言の文字数を比較し、大きい方の文字数。

<用語定義>

文書比較機能を実装するにあたり生成される、ドキュメントや、プログラムのコメントなどに記載される用語で、一部、文書比較用に、用語の意味を独自に定義しているものがある。それら用語を以下に纏める。

文字：文字列を構成する単一、すなわち文字列長が 1 の文字列。

単語：同一種の連続した文字を、ひとくくりにした文字列。

文言：複数の単語を連結した文字列のこと。

文節：句読点をひとつの区切り(終端)とした文言のこと。

文節接尾辞：日本語文章において、文節の末尾に記述されることの多い。記号のことで、
″、。，．・？！）］｝」』!),.:;?]｝。」、″などがある。

文：複数の文言から文節から構成され、句点やクエスチョンマークなどの文末特有の
接尾辞(改行を含む)で、終端する文言。
″。。..！!??))]] }} 」」』″ & vbLf

文書：複数の文節や文から構成された文字列のこと。

文頭：文書の先頭。

文末：文書の末尾。

<関数 IF>

Public Function **DfDiffSentence**(_
ByVal snt_a *As String*, _
ByVal snt_b *As String*, _
ByRef dinf *As* StructDiffInfo) *As Long*

概要：文書 A と文書 B の DIFF 比較を行い、結果を dinf 構造体に格納する。

戻り値：相違のあった個数が返却される。すなわち、0 の場合は、文書 A と B が、相違のない同じ文書であったことを、意味している。

引数 snt_a：比較する文書 A の文字列を指定する。

引数 snt_b：比較する文書 B の文字列を指定する。

引数 dinf：比較結果を保存する StructDiffInfo 型の構造体を指定する。

Public Function **DIFF**(_
ByVal rg1 As *Range*, _
ByVal rg2 As *Range*, _
Optional is_show_gidx *As Boolean* = False, _
Optional is_auto_show_gidx *As Boolean* = True) *As String*

概要：エクセルのセルから呼び出せる、DIFF 関数である。エクセルのセルに設定された 2 つのテキストに対して DIFF 比較を行い、結果を文字列データにて返却する。

戻り値：相違を文字列データにて、返却する。

引数 rg1：比較する文書 A が設定されたセルを指定する。

引数 rg2：比較する文書 B が設定されたセルを指定する。

引数 is_show_gidx：返却する文字列データにグループインデックスを含めるかどうかを指定する。省略可能である。

引数 is_auto_show_gidx ： 返却する文字列データにグループインデックスを含めるかどうか自動判定するかどうかを指定する。省略可能である。

Public Function **DfGetSameRate**(dinf *As* StructDiffInfo) *As Double*

概要 : DfDiffSentence 関数により、取得した DIFF 結果情報より、文書 A と文書 B の一致率を算出する。

戻り値 : 文書 A と文書 B の一致率を返却する。

引数 dinf : DfDiffSentence 関数により、取得した DIFF 結果情報を指定する。

以下が、比較結果の格納用の StructDiffInfo 型の構造体である。この構造体のメンバ変数は a、b はそれぞれ、文書 A と文書 B に関する情報で、比較処理に使用するデータとして利用される。最終的に、作り出される DIFF 情報の、中間バッファと位置づけられる。

```
'文書比較結果情報構造体
Public Type StructDiffInfo
    a As StructSentenceInfo_A           '文書 A の情報
    b As StructSentenceInfo_B           '文書 B の情報

    '▼▼▼ Output DIFF 結果 ▼▼▼
    words() As StructDiffResultWordInfo
    words_num As Long
    is_move As Boolean                  '文書の移動があるかどうか
    '▲▲▲ Output DIFF 結果 ▲▲▲

End Type
```

最終的に、作り出される DIFF 情報は、上記の StructDiffResultWordInfo 型の構造体 words に格納されている。この構造体は、以下のようになっている。

```
'文書比較結果単語情報構造体
Public Type StructDiffResultWordInfo
    df_inf As EnumDiffResult            'DIFF 結果
    str1 As String                      '文書 A での単語
    str2 As String                      '文書 B での単語
    str1_len As Long                    '文書 A での単語長
    str2_len As Long                    '文書 B での単語長
    idx_group As EnumGroupIdx           '文書 B で文書の入替があった場合のグループインデックス
End Type
```

DIFF 結果の df_inf には、以下の列挙値のいずれかが格納されている。

```
'文書 A と B の比較結果列挙値
Private Enum EnumDiffResult
    EDR_UNKNOWN = 0
    EDR_SAME                            'B 側と A 側は同じ
    EDR_INSERT                          'B 側に挿入された
    EDR_UPDATE                          'B 側が更新された
    EDR_DELETE                          'B 側から削除された
End Enum
```

以上