類似語チェック機能仕様

本書は、FootcelWords における類似語チェック機能の仕様について、簡潔に纏めるものである。類似語チェック機能の設計の狙いは、用語辞書作成という観点から、言葉の表現を重要視する必要があり、辞書内で使用される用語･語句の統一性を保つことを目的としている。

類似語機能チェックには、チェックの程度が寛容な、寛容モードと、厳密にチェック処理を行う厳密モードの 2 つがあるものとする。以下にそれぞれのモードに固有な処理と、共通な処理について、記述する。ここでは、文字列 A と文字列 B の 2 つの文字列の類似チェックを行うものと仮定し、説明する。

共通機能

- 文字列 A と文字列 B がまったく同じであれば、類似していないと判定。(例 A:サッカー B:サッカー)
- 文字列 A と文字列 B を構成する各文字が、文字列長(長い方)の半分の個数より多く一致していなければ、類似しているとみなさない。一致個数 > 半分。
- 文字列 A と文字列 B に含まれる各文字が、互いに 1 文字も一致するものが無い場合、類似していないと判定。(例 A:サッカー B:テニス)
- 基本ポリシーとして、半角と全角は区別しない。設定により区別できるように拡張出来ることを推奨する。(例 A:サッカー B:ｻｯｶｰ は同じ文字列とみなす)
- 基本ポリシーとして、英字の大文字、小文字は同一視する。設定により区別できるように拡張出来ることを推奨する。(例 A:FIFA B:fifa は同じ文字列とみなす)
- 文字列 A、B のいずれかの文字列長が 1 以下(空文字列含む)の場合は、類似していないと判定。(A:得点 B:点)
- “アイウエオ”と”ァィゥェォ”のそれぞれの文字は、異なる文字として扱う。
- “あいうえお”と”ぁぃぅぇぉ”のそれぞれの文字は、異なる文字として扱う。
- 文字列の各文字の並び順を考慮する。
- 片仮名、平仮名の違いは、異なる文字と判定する(同一視しない)。

寛容モード

- 文字列 A と文字列 B の文字列長の差が 3 文字以上であれば、類似していない判定する。

  (例 A:サッカー B:サッカーボール)
- 文字列 A と文字列 B の文字列長がいずれも、、**異なる文字が 2 文字以内まで、類似と判定**。

厳密モード

- 文字列 A と文字列 B の文字列長の差が 2 文字以上である場合は、類似していないと判定する。

  (例 A:サッカー場 B:サッカー球技場)
- 文字列 A と文字列 B の文字列長が、いずれも 3 文字以上の場合、**異なる文字が 1 文字以内であれば、類似と判定**する。

以上