GSLetterNeo Vol.97
2016年8月

テキストデータ群の重ね合わせによるヒストリと特徴のインタラクティブ表示(3)

松原 伸人、土屋 正人
matubara@sra.co.jp, m-tsuchi@sra.co.jp

◆ はじめに

大量のデータをインタラクティブに操作する、Web ブラウザ上で動作するアプリケーションのプロトタイプ開発を行っています。データとして GSLetterNeo を使って、

- **ドキュメントタイトル、著者等を表示するレイヤー**
- **特徴語群を表示するレイヤー**
- **希少語群を表示するレイヤー**

の 3 層を、画面奥に向かって重ねて配置することで、ドキュメントの時間関係と特徴がひとまとまりに見えるようにしたプロトタイプのインタフェースと実装を、Vol.95 と Vol.96 で紹介しました。

今回はデータを解析するために必要な辞書を中心に紹介します。

◆ 形態素解析[1]と辞書

各単語のドキュメント内での使用頻度と、単語が使われているドキュメント数から、tf-idf が計算できます。

tf-idf はドキュメント内でたくさん使われていて、他のドキュメントではそれほど使われていない単語の度合いを表します。

tf-idf - [Wikipedia]
https://ja.wikipedia.org/wiki/Tf-idf

希少語群の検出では、tf-idf の計算式の分子 tf を逆にして単語の出現頻度の低い単語を計算しています。

tfidf = tf / idf
minority = (1 - tf) / idf

実際に形態素解析を行ってみると分かりますが、解析対象に合う辞書の作成が必要になります。

今回は IPA 辞書をベースにして利用して、解析結果を見ながら必要に応じて辞書に単語を追加したり、不明な語の定義の変更を行ったりしています。

辞書に 1 個単語を追加すると、解析結果が単語の計測結果にも影響します。

◆ 解析結果と計測結果の可視化

このような辞書作成して単語計測結果を見るためのツールが図 1～図7です。

このツールは HTML と JavaScript と CSS でプログラムしてあります。

GSLetterNeo のテキストの解析結果を見ていく過程で作成し、使っていました。

画面左側にあるドキュメントリストで対象ドキュメントをクリックすると、画面中央のテキストエリアにテキスト内容を表示、画面右側に解析結果を表示します。

テキストエリア内で、太字で大きな字になっている単語は tfidf 値が高い単語で、値が大きいほど大きな字で表示されるようになっています。

---

1 文法的な情報の注記の無い自然言語のテキストデータ(文)から、対象言語の文法や、**辞書**と呼ばれる単語の品詞等の情報にもとづき、形態素(*Morpheme*, おおまかにいえば、言語で意味を持つ最小単位)の列に分割し、それぞれの形態素の品詞等を判別する作業(Wikipedia より引用)。

図 1　選んだ単語にアンダーラインをつけて示す

図 4　ドキュメント数順

右側のリストの単語をクリックすると、テキストエリア上の単語にアンダーラインをつけて示します(図1)。

クリックするたびに登場する順に単語を移動して見ることができます。

図 2　単語計測結果を見るツールの画面
（単語の登場順）

図 5　tfidf順

図 3　出現数順

図 6　希少順

画面下部にある「spawn top 10 words」ボタンを押すと、登場順、出現数順、ドキュメント数順、tfidf順、希少順にそれぞれスコアの高い 10 語を抽出し、JSON 形式で解析結果を出力します（図 7）。

出力した JSON データを用いて、特徴語レイヤーと希少語レイヤーを作成しています。

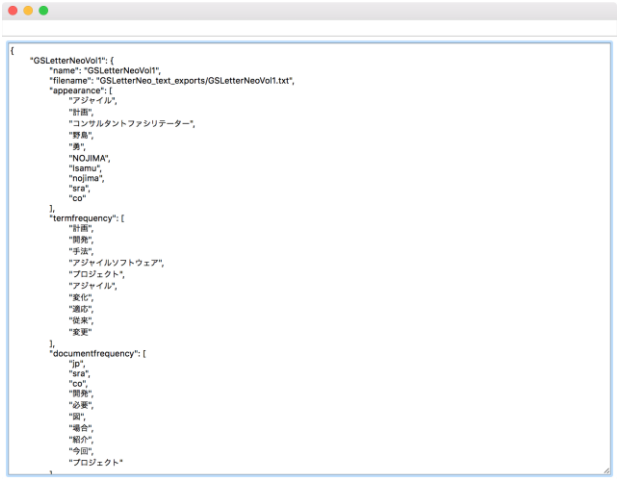

図 7　解析結果を JSON 形式で出力した画面

◆ 終わりに

解析結果のリストを見るだけでは、なぜこういう分解がなされたのかわからない場合も、元のテキストと合わせて見られると理由がなんとなくわかり、試しに辞書に単語を追加してみると、正しそうな結果になることが多かったです。

多くの場合は、もともと辞書には載っていないような単語が現れた時に、分解がうまくいかないようです。

GSLetterNeo Vol.97
2016 年 8 月 20 日発行
発行者●株式会社 SRA 先端技術研究所
編集者●土屋正人

バックナンバを公開しています●http://www.sra.co.jp/gsletter
ご感想・お問い合わせはこちらへお願いします●gsneo@sra.co.jp

株式会社SRA
〒171-8513 東京都豊島区南池袋２－３２－８