

キーワード抽出や索引生成に 欠かせないツール

Rosette 日本語形態素解析システム (RBL-JA: Rosette Base Linguistics Japanese) はポータブルで高性能な日本語テキスト単語分節(分かち書き)エンジンです。ウェブページ、電子メール、各種データベースの文書など、どのような種類の文書にも RBL-JA は威力を発揮します。

日本語テキストは通常、スペースなしで書かれているうえ、漢字・平仮名・カタカナ・ローマ字など 4 種類の書記体系が利用されているため、コンピュータでの解析が容易ではありません。しかし、RBL-JA はそのような日本語の解析を的確かつ効率的に実現します。

日本語のテキストの解析には、まず文章をトークン(単語)へと分割する必要があります(分かち書き)。さらに単語には様々な派生形がありますが、その処理を容易にするため一つの基本形に統一して処理したり(ステミングや基本化)、また複合名詞を構成要素別に分けたり(分割)、また様々な書式を統一するための相互変換、たとえば半角入力⇄全角入力等(正規化)を実行します。

中米の小国コスタリカは、四国と九州を合わせたほどの面積しかない小国だが、国立公園や自然保護区域の数は60箇所以上あり、それらを活かしたエコ・ツアーがとて盛んなところだ。

#	Item	Part of Speech	Stem	Compound
20	国立公園	NC		国立 - 公園
21	や	PL		
22	自然保護	NC		自然 - 保護
23	区域	NC		
24	の	PL		
25	数	NC		
26	は	PL		
27	60	NN		
28	箇所	NU		
29	以上	NC		
30	あり	V	ある	
31	,	PUNCT		
32	それら	NR		
33	を	PL		
34	活かした	V	活かす	
35	エコ・ツアー	NC		

RBL-JA はこれらの処理をおこなうためのアルゴリズムを持ち、さらに品詞情報と頻度情報をタグ付けた約 50 万語の辞書(* 半年毎に更新)を有し、正確に日本語テキストを単語に分割します。

RBL-JA は、外来語の普及で増加しているカタカナ文字列の分かち書きも的確に行えるよう設計されています。このソフトウェア・ライブラリはポータブルで、デスクトップ PC から、1分間に数百文書进行处理するような高速マルチ CPU ウェブサーバまで、多くのプラットフォームで動作可能です。

応用分野

RBL-JA は日本市場で既にご活躍の企業、また今後日本市場参入を目指す企業で幅広くご利用いただけます。特にインターネット用ソフトウェア、企業用アプリケーション、また、消費者直結型の e-ビジネスにおいても高い利用価値を持っています。

特に膨大な日本語文書の索引付けを必要とする情報検索、または、分節処理や語幹抽出、品詞分析を必要とする自然言語処理に、その威力を発揮します。

電子メールやブラウザ、ウェブでよく使用される文字コードにも対応しています。また、Rosette 言語判別システムと一緒に使用することで、言語・文字コードの自動判別の対応範囲を広げることができます。

特長

- 分節(分かち書き) / トークン化
- ユーザー定義辞書の利用が可能
- 品詞の付与、タグ付け
- 複合語の抽出・分解
- キーワード抽出、名詞句抽出
- 活用語の基本形(終止形)を出力
- 句読点、スペース、数字の識別およびフィルタリング
- ストップワードの検知、ユーザー定義が可能
- 平仮名、カタカナ、漢字、ローマ字などの日本語各種文字表記に対応
- カタカナ文字列の分かち書きが的確におこなえる設計
- 全角・半角文字に対応
- 各トークンの語句を含む辞書(標準辞書もしくはユーザー定義辞書)の辞書 ID を、解析結果に表示
- ユーザー定義辞書に、顔文字などの特殊文字を含む語句の登録可能

辞書データ

- 約 50 万語収録
(新語 360 語以上をさらに追加)
- 日本人と西洋人の人名、地名、企業名を含む
- 平仮名、カタカナ、漢字、ローマ字表記を含む

表記ゆれ対応モジュール (別売オプション)

表記ゆれ対応モジュール (JOA) は、日本語をはじめ各国言語の語句の表記ゆれを標準的な形に正規化する辞書ベースのソフトウェアです。アルゴリズムによる正規化処理は誤りが多いので、JOA はそのアプローチは取らず、辞書編集者が実際のテキストから抽出した多くの表記ゆれパターンをもとに作成した辞書を使用します。

現状の JOA のデータは特に汎用の Web 検索を念頭に編纂してあります。漢字の字体だけでなく、カタカナの表記ゆれも正規化し、検索に役立つように設計されています。JOA データには約 5,000 のカタカナ語句 (ペア) が含まれています。

対応プラットフォーム

以下のプラットフォーム対応の SDK を提供します。その他のプラットフォームのサポートも、ご要望に応じ対応します。

AIX 5.2, PowerPC
FreeBSD 4.8, IA32
FreeBSD 6.0, IA32/AMD64
HP-UX 11.0, PA-RISC
HP-UX 11.22, IA64
Linux Debian 3.1, IA32/AMD64

Linux Fedora Core 4, IA32/AMD64
Linux Fedora Core 5, IA32/AMD64
Linux Red Hat ES 2.1, IA32
Linux Red Hat ES 3.0, IA32/AMD64
Linux Red Hat ES 4.0, IA32/AMD64
Linux Suse EL 10, IA32/AMD64

Solaris 8, SPARC32/64
Solaris 9, SPARC32/64, IA32
Solaris 10, SPARC32/64, IA32/AMD64
Windows NT/XP/2003, IA32/AMD64
Windows Vista/2008, IA32/AMD64

カタカナ表記ゆれの例を以下に示します。左が正規形、右の複数の表記が左の文字列に正規化されます。

ダンスセラピー ←	ダンスセラピ / ダンステラピ / ダンステラピー
エキスポ ←	エクスポ
バーミューダー ←	バーミューダ / バミューダ
ファミコン ←	ファミリーコンピュータ / ファミリーコンピューター
ベネチア ←	ベニス / ベネツィア / ヴェネチア / ヴェネツィア

また、漢字の新旧字体もサポートしています。辞書には旧字体の漢字を含む語句が登録されており、それらを新字体の表記に正規化します。

こちらのオプションについての詳細は、下記までメールにてお問い合わせください。

お問合せ

さらに詳しい製品情報ならびに評価版のご利用をご希望の方は下記へご連絡ください。

info@basistech.jp

電話 03-3511-2947



詳細

www.basistech.jp

お問合せ

info@basistech.jp

電話番号

03-3511-2947

ベイシス・テクノロジー株式会社

〒102-0084 東京都千代田区二番町 9-6

U.S. 本社 (ボストン)

One Alewife Center, Cambridge, MA 02140

ワシントン D.C. 支社

13800 Coppermine Road, Herndon, VA 20171