全文検索エンジン　Apache Solr

野村総合研究所
オープンソースソリューション推進室

野村総合研究所のOpenStandia（オープンスタンディア）は、おかげさまで、2006年のサービス開始から2011年までの5年間で契約数累計が1,000件を突破いたしました！

オープンソースまるごと
OpenStandia™
Open Source Technology

NRI 未来創発
Dream up the future.

株式会社 野村総合研究所 オープンソースソリューション推進室
Mail：ossc@nri.co.jp　Web: http://openstandia.jp/

サブスクリプション＆サポートサービスの範囲

Apache Solrサブスクリプションに含まれるもの

1) 安定稼働版パッケージ

Apahce Solr（Tomcatにデプロイ済）

2) その他

日本語処理プラグイン「CharFilter」「Tokenizer ］
日本語用「想定検索（もしかして検索）」
日本語用「リアルタイムクラスタリング」
日本語用「検索語サジェスチョン」
日本語用「パーソナライズ検索」
Solrモニタリング機能
自然言語処理ライブラリ（NLP4L）
各種スクリプト

3) マニュアル

日本語マニュアル

4) サポートサービス

Lucene/Solr自体の問題に対するバグフィックス対応
Lucene/Solr以外の問題に対するアナウンス、Solr稼働を行う為の周辺ミドルウェア（例：Tomcat）などで発見された脆弱性とその対応に関するご案内（※.1）
専用サポートポータル（WEB経由での問合せ対応）
※.1 Tomcatのセキュリティパッチを提供しません。（別途、Tomcatのサポートをご購入ください）
１２インシデントまでの対応

5) 利用可能インスタンス

５インスタンスまで利用可能

検索語サジェスチョン

サジェスチョン機能を使うと、ユーザは入力した検索語と関連性の高い他の検索語のヒントを得られるようになります。
最初の数文字を入力しただけで検索語のサジェストが得られますので、タイプ量を削減できるようになります。
また、ヒット件数が0件のものはサジェストされないので、「ヒット0件」となることがありません。
Solrにはもともと検索語をサジェストするための支援機能がありますが、日本語ではIMEが介在するためにその機能は使えません。
本プラグインを使用すれば、日本語環境に対応した検索語のサジェスチョンが可能となります。

1）よみがなデータをロードします。

```
新しい,1203,形容詞,自立,*,*,形容詞・イ段,*
新し,1203,形容詞,自立,*,*,形容詞・イ段,文
新しから,1203,形容詞,自立,*,*,形容詞・イ段
新しかろ,1203,形容詞,自立,*,*,形容詞・イ段
新しかっ,1203,形容詞,自立,*,*,形容詞・イ段
新しく,1203,形容詞,自立,*,*,形容詞・イ段,
新しくっ,1203,形容詞,自立,*,*,形容詞・イ段
新しゅう,1203,形容詞,自立,*,*,形容詞・イ段
新しゅぅ,1203,形容詞,自立,*,*,形容詞・イ段
新しき,1203,形容詞,自立,*,*,形容詞・イ段,
新しけれ,1203,形容詞,自立,*,*,形容詞・イ段
新しかれ,1203,形容詞,自立,*,*,形容詞・イ段
新しけりゃ,1203,形容詞,自立,*,*,形容詞・イ
新しきゃ,1203,形容詞,自立,*,*,形容詞・イ段
新し,1203,形容詞,自立,*,*,形容詞・イ段,ガ
おおきい,2982,形容詞,自立,*,*,形容詞・アウ
おおきし,2982,形容詞,自立,*,*,形容詞・アウ
おおきから,2982,形容詞,自立,*,*,形容詞・ア
おおきかろ,2982,形容詞,自立,*,*,形容詞・ア
おおきかっ,2982,形容詞,自立,*,*,形容詞・ア
```

2）サジェスト用データを作成します。

```
国際社会　日本　役割,11
国際社会　意味,11
国際社会と日本,11
国際社会への復帰とその後の発展,11
中小企業緊急雇用安定助成金,9
中小企業診断士,9
中小企業庁,9
中小企業の会計に関する指針,9
```

パーソナライズ検索

営業社員の検索結果

```
<response>
  <lst name="responseHeader">...</lst>
  <result name="response" numFound="6" star
    <doc>
      <arr name="statement">
        <str>Luceneのパンフレット</str>
      </arr>
    </doc>
    <doc>
      <arr name="statement">
        <str>Luceneの価格表</str>
      </arr>
    </doc>
    <doc>
      <arr name="statement">
        <str>Luceneの技術資料</str>
      </arr>
    </doc>
```

技術社員の検索結果

```
<response>
  <lst name="responseHeader">...</lst>
  <result name="response" numFound="6" s
    <doc>
      <arr name="statement">
        <str>Luceneの技術資料</str>
      </arr>
    </doc>
    <doc>
      <arr name="statement">
        <str>Luceneの設計資料</str>
      </arr>
    </doc>
    <doc>
      <arr name="statement">
        <str>Luceneの企画資料</str>
      </arr>
    </doc>
```

ルールベースの固有表現抽出(１)

固有表現抽出とは、人名、地名、組織名、商品名、部品名、キャラクター名などの固有名詞を、自然言語の文書から抽出するタスクです。固有表現抽出をSolrと組み合わせると、特にフィールドの少ない企業内検索に多大な威力を発揮します。
以下は、付属のサンプルデータから「政策」や「構想」などの計画を抽出し、他フィールドにコピーする例です。

ファセットで、固有表現別のドキュメント数を表示します。

```
<lst name="facet_fields">
  <lst name="plan_sm">
    <int name="安全保障">11</int>
    <int name="社会保障制度">11</int>
    <int name="社会保障">10</int>
    <int name="行政改革">10</int>
    <int name="公務員制度改革">8</int>
    <int name="年金制度">7</int>
    <int name="拉致問題">7</int>
```

固有表現をあらかじめ登録しておきます。

エネルギー政策
外交・安全保障政策
核軍縮・不拡散政策
環境政策
競争政策
経済政策
経済財政政策
農地政策
都市政策

ルールベースの固有表現抽出(2)

固有表現(弊社ではLucene/Solrクラス名)を
あらかじめ登録しておきます。

ArabicAnalyzer
ArabicLetterTokenizer
ArabicLetterTokenizerFactory
ArabicNormalizationFilter
ArabicNormalizationFilterFact y
ArabicNormalizer
ArabicStemFilter
ArabicStemFilterFacto
ArabicStemmer
BulgarianAnalyzer
BulgarianStemFilter
BulgarianStemFilterFa
BulgarianStemmer
BrazilianAnalyzer
BrazilianStemFilter
BrazilianStemFilterFa
BrazilianStemmer

ファセットで、Lucene/Solrクラス名別のドキュメント数を表示して、絞り込み検索を促します。

ようやく Lucene 2.9.0 がリリース | 関口宏司のLuceneブログ

zer)。 新しいハイライタであるFastVectorHighlighterクラス の追加。なお先
日、FastVectorHighlighterについての解説記事を Lucene in Action...| コメント
Solr1.4ではFastVectorHighlighterが使えるようになるのでしょうか？SOLR-1268に投
票しておきました。 |...Hatcher,Otis Gospodnetic,Mike McCandless
FastVectorHighlighterについて解説記事を寄稿しました。 + RECOMMEND
http://lucene.jugem.jp/?eid=342
関口宏司のLuceneブログ

著者 で絞り込む
Koji Sekiguchi (4)

クラス名 で絞り込む
FastVectorHighlighter (440)
QueryParser (437)
NGramTokenizer (436)
PhraseQuery (436)
EdgeNGramTokenFilter (435)
IndexWriter (81)
Analyzer (78)

想定検索(もしかして検索)

日本語に対応した「もしかして」検索機能です。

以下は、歴代内閣総理大臣の所信表明演説をデータソースにしたとき、
日本語のスペルミスやIMEの変換ミスに対してどのような候補を「もしかして」として表示できるかを示しています。

入力ミスの例	「もしかして」で出力される例	備考
笑止高齢化	少子高齢化	
いんたーねっｔ	インターネット	最後の"o"を入力せず変換が正しくできないまま検索した例
しゃ皆保険	社会保険	
滑稽銀	国会議員	
オーストリア	オーストラリア	データソースに「オーストリア」がないため
partner湿布	パートナーシップ	
滅入るマガジン リーマン	メールマガジン サラリーマン	2単語以上の誤りにも同時に対応

リアルタイムクラスタリング

http://localhost:8080/solr/basic/clustering?q=*%3A*&indent=on&wt=json&carrot.produceSummary=false

通常の検索結果の後ろの部分に、次のようなクラスタリング結果が、それぞれのクラスタにつけられたラベルとともに表示されます（JSON形式表示）。


```
"clusters":[{
    "labels":["社会保障 雇用創出 社会保障"],
    "score":17.413701721479544,
    "docs":["176",
      "174",
      "171"]},
  {
    "labels":["中央省庁再編 行政改革 構造改革"],
    "score":23.794252515821555,
    "docs":["151",
      "147"]},
  {
    "labels":["原発事故"],
    "score":17.39688335449091,
    "docs":["179",
      "178"]},
  {
    "labels":["地球温暖化"],
    "score":13.268708473570012,
    "docs":["173",
      "169"]},
  {
    "labels":["戦略指針 イノベーション"],
    "score":23.46372853035862,
    "docs":["166",
      "165"]},
  {
    "labels":["景気対策 財政再建 中長期的 経済成長"],
    "score":25.9951706804661,
    "docs":["171",
      "170"]},
```


日本語に対応した検索結果文書のクラスタリングをリアルタイムに行うことができます。クラスタリングは、検索結果文書を適当にラベル付けされたテーマ別グループに自動分類する機能です。

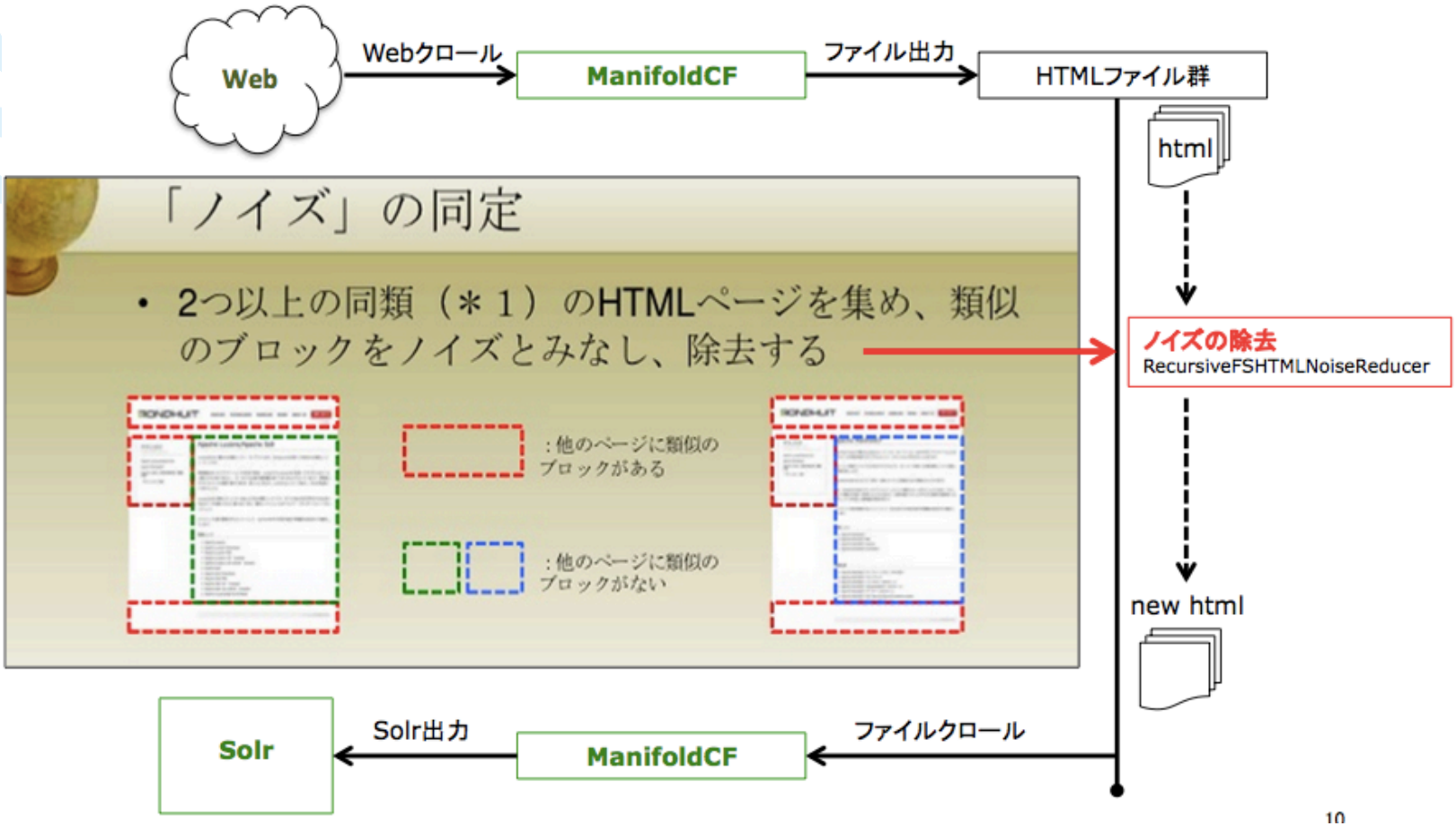
この図は、歴代内閣総理大臣の所信表明演説をデータソースにしたクラスタリングの結果です。

クラスタリング機能はファセット機能と並び、大量にヒットした文書から、目的の文書をすばやく見つけるための絞り込みに役立ちます。フィールドファセットと異なり、あらかじめ絞り込み用のフィールドを整備する必要がありません。そのため、R&D部門など創造性が重視される部門ではスコアの低い文書を掘り起こし、いろいろな「気づき」を与えてくれるなどの効用が知られています。

自然言語処理ツール

- HTMLNoiseReducer

  同類の複数のHTMLページからノイズを削除するツール



※Apache Solrのサポート範囲にはManifoldCFのサポートは含まれておりません。
必要な場合は、別途OpenStandiaのManifoldCFのサポートをご契約ください。

日本語処理プラグイン(CharFilter)

- MappingCharFilter

mapping-ja.txtで
全角カタカナや半角英数字への正規化を処理します。

半角・全角正規化	ｱｲｳ１２３ <=> アイウ123
新旧漢字変換	慶應大学 <=> 慶応大学

```
# 半角カタカナ=>全角カタカナ
"ｱ" => "ア"
"ｲ" => "イ"
"ｳ" => "ウ"
"ｴ" => "エ"
"ｵ" => "オ"
"ｶ" => "カ"
"ｷ" => "キ"
"ｸ" => "ク"
```

- NakaguroCharFilter

中黒の正規化を処理します。

オープンソース・ソフトウェア <=> オープンソースソフトウェア
ザ・ビートルズ <=> ザビートルズ

- KanjiNumberCharFilter

"四十七"などの漢数字を"47"という算用数字(アラビア数字)に正規化します。

変換前	変換後
四七	47
四十七	47
四百七	407
四〇七	407
四千十七	4017
一億	100000000
1億	100000000
一千億	100000000000
壱億弐千参百四拾萬	123400000

- WarekiCharFilter

和暦年を西暦年に正規化します。

変換前	変換後	備考
大化元年 大化1年	645年	元年でも1年でもOK
明治30年 明治三十年 明治参拾年	1897年	KanjiNumberCharFilterが必要
昭和64年 平成元年	1989年	昭和と平成が重なった年もOK

日本語処理プラグイン(Tokenizer)

- JaBuzzPhraseTokenizer

文章中で一定以上連続する漢字文字やカタカナ文字は専門用語と見なします。

↓

ファセット、クラスタリング、レコメンド、サジェスト(オートコンプリート)、もしかして検索、に活用します。

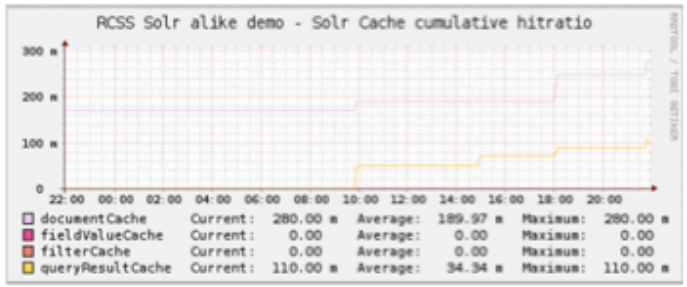
```
<tokenizer
class="com.rondhuit.solr.analysis.JaBuzzPhraseTokenizerFactory"
minAlphabetLen="4" minKatakanaLen="5" minKanjiLen="4"/>
```

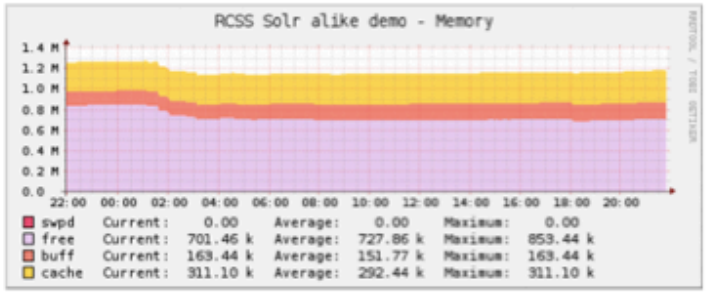
カントリー・アイデンティティ	2
共済年金	2
年間三万人	2
関係機関	2
戸別所得補償制度	2
成長戦略	2
北方四島	2
改正手続	2
医療・介護サービス	2
政府一体	2
タウンミーティング	2
政治主導	2
政策決定	2
九十年代以降	2
教育再生	2
参議院議員通常選挙	2

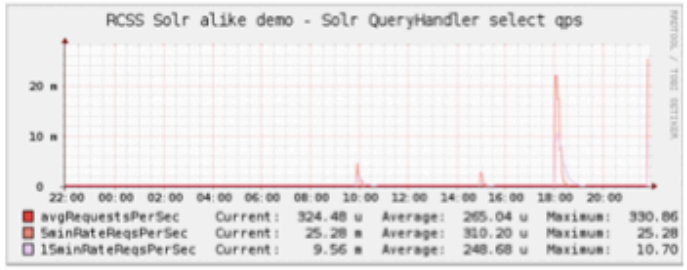
役割分担	2
消費者庁	2
事業規模	1
事業環境	1
ワールドカップサッカー大会	1
司法制度	1
事業再生	1
ワーク・ライフ・バランス	1
チャレンジド	1
事業内容	1
事業仕分	1
ワンストップ・サービス	1
事故リスク情報	1
原子力産業	1
事後チェック型	1

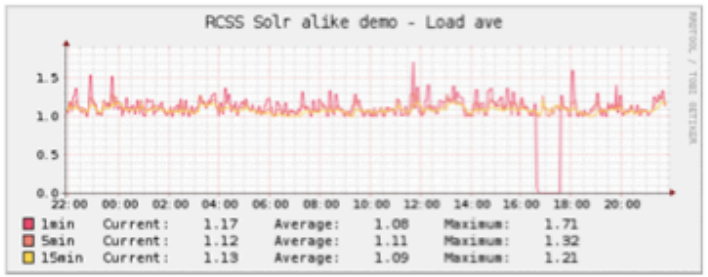
Solrサーバーのモニタリング

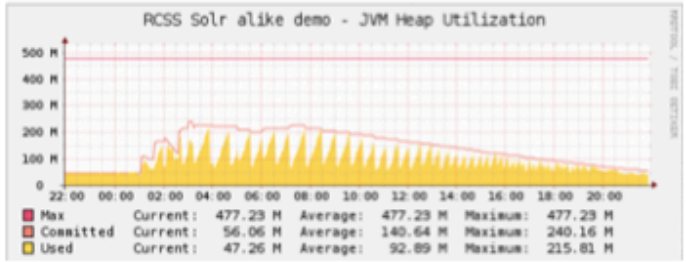
- CPU/メモリ/Disk/ネットワークなどのOSレイヤだけでなく、JVMヒープやSolrのキャッシュヒット率やリクエストハンドラ／アップデートハンドラ／レプリケーションハンドラの利用状況まで統合的にモニタリングできます。以下はモニタリング可能な項目の一部です。





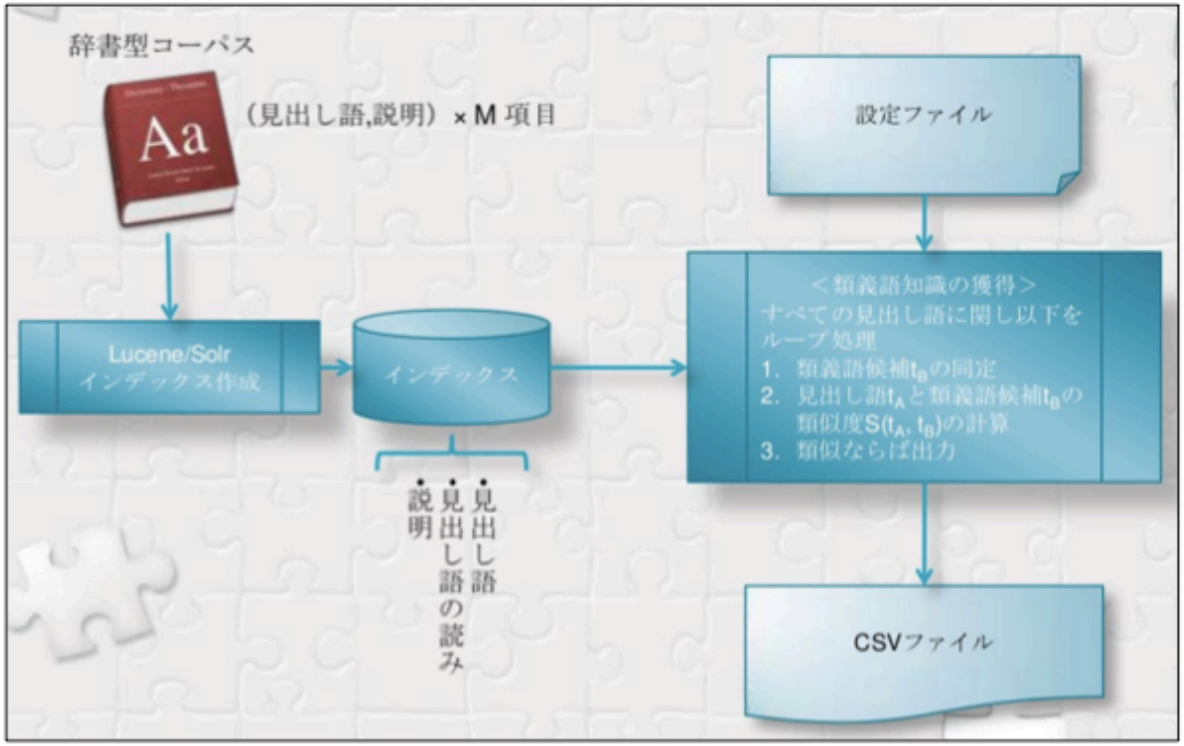






NLP4L(自然言語処理ライブラリ)

● 辞書型コーパスからの類義語知識の自動獲得



見出し語とその説明からなる「辞書型コーパス」から、Lucene/Solrでそのまま使える類義語辞書(CSVファイル)を出力します。類義語は見出し語を原型語とみなし、説明からその原型語の略語候補を抽出したあとスコアリングをし、原型語の略語として確度の高い原型語と略語のペアを出力します。
右図は、「辞書型コーパス」として日本語Wikipediaを用いた場合に、実際に獲得できる類義語知識の一例です。

原型語	略語
入国管理局	入管
文房具	文具
社員食堂	社食
国際連盟	国連
リポビタンD	リポD
ベルサイユのばら	ベルばら
木村拓哉	キムタク
ファミリーレストラン	ファミレス
ワードプロセッサ	ワープロ

NLP4L(自然言語処理ライブラリ)

- JaNBestTokenizer
  - 3つのOSS辞書が使える
    ipadic(奈良先端大)、juman(京大)、unidic(国立国語研究所)
  - 日本語の単語分割における多義性に対応している

1	2	3	4	5	6	7	8
ここ	で	はきもの		を	脱い	で	ください
		は	きもの				

  - 類義語をサポートしている(ユーザ類義語辞書もサポート)

1	2	3	4
旭丘	へ	引越し	する
旭が丘		引っ越し	

- **OpenStandiaは、「攻めのIT」を支援します。**
- **オープンソースのことなら、なんでもご相談ください！**

オープンソースまるごと

OpenStandia™

Open Source Technology

お問い合わせは、NRIオープンソースソリューション推進室へ

ossc@nri.co.jp

http://openstandia.jp/

NRI 未来創発

Dream up the future.