第25回 Lucene/Solr勉強会
オンライン／初心者編３
Apache Solr のスキーマ設定

10/19/2021 @kojisays

株式会社ロンウイット

はじめに（再掲）

- 約2年ぶりの開催。
    - 内容が高度化、準備が大変。間隔が空くとますますハードルが上がる。コロナ禍で勉強会もオンラインが一般的に。オンラインで短めのものを、これまでよりも頻度高めで。
- 今後数回に分けて初心者向けの内容を用意。
    - 関口から初心者の方々へ。今回のみならず毎年？
- その後は事例や新機能の発表など。
    - 勉強会＝双方向で。

ネタをお持ちの方等、アンケートでぜひお知らせください

前回受講アンケートより

- 事例発表1社様確定。あともう１～２社様いかがでしょう？

- managed-schema vs schema.xml

- Apache Solr と Elasticsearch

  - CodeZine 記事：「中の人」が教える！ 奇跡の巨大IT系ボランティア団体ASFの組織運営とは？ https://codezine.jp/article/detail/10991

本日の内容

- Solrスキーマの基礎知識
  - サンプルデータ（livedoorニュースコーパス）
- 日本語検索の留意点
  - 日本語検索表記揺れの分類
  - 基本テクニック
  - シノニム検索

スキーマが想定する文書

- livedoorニュースコーパス
  - NHN Japan 社（現 LINE 社）のlivedoorニュース記事からHTMLタグを消去したコーパス。
- タイトル、カテゴリ、日付、本文、記事URL
- 検索のみならず、NLP研究でも徐々に人気に。
- CC BY-ND 2.1（当時）

livedoorニュースコーパス（例）

カテゴリ string

master　extensions / data / livedoor / data / dokujo-tsushin / **dokujo-tsushin-4778030.txt**

```
1  http://news.livedoor.com/article/detail/4778030/ ← URL string, ユニークキー
2  2010-05-22T14:30:00+0900 ← 日付 date
3  友人代表のスピーチ、独女はどうこなしている? ← タイトル text_ja
4  もうすぐジューン・ブライドと呼ばれる6月。独女の中には自分の式はまだなのに呼ばれてばかり……という「お祝い貧乏」状態(
5
6  「お願いがあるんだけど……友人代表のスピーチ、やってくれないかな?」
7
8  さてそんなとき、独女はどう対応したらいいか?
9
10 最近だとインターネット等で検索すれば友人代表スピーチ用の例文サイトがたくさん出てくるので、それらを参考にすれば、
11
12 「一晩で3人位の人が添削してくれましたよ。ちなみに自分以外にもそういう人はたくさんいて、その相談サイトには同じように
13
14 しかし「事前にお願いされるスピーチなら準備ができるしまだいいですよ。一番嫌なのは何といってもサプライズスピーチ!」と
15
16 「私は基本的に人前で話すのが苦手なんですよ。だからいきなり指名されるとしどろもどろになって何もいえなくなる。そうす
17
18 サプライズスピーチのメリットとしては、準備していない状態なので、フランクな本音をしゃべってもらえるという楽しさがある。
19
20 一方「ありきたりじゃつまらないし、ネットで例文を検索している際に『こんな方法もあるのか!』って思って取り入れました」とい
```

text_ja 本文

Solrスキーマの基礎知識

- schema.xmlではなくmanaged-schema を使う。スキーマレスモードは使わない。
  managed-schema は編集してはいけない（ことになっている）。
- <fieldType …/>で与えられているフィールド型を使って必要なだけ <field …/>を定義する。
- いわゆる全文検索フィールドは <fieldType …/> 内に <analyzer …/> を設定する。
  - Analyzer := 0個以上の <charFilter /> + 1個の <tokenizer /> + 0個以上の <filter />
  - トークンテキスト以外に、ポジションとオフセットが重要な Attribute。
- <uniqueKey />はオプション扱いだが、業務システムでは必須。
- <copyField …/> を使ってインデクシング時にフィールドをコピー。
- Solr管理画面のAnalysisを使って単語分割の様子をチェック。

livedoorニュースコーパスのスキーマ設定

```
<field name="url" type="string"
       indexed="true" stored="true"/> ①

<field name="category" type="string"
       indexed="true" stored="true"/> ②

<field name="title" type="text_ja" ③
       indexed="true" stored="true" multiValued="false"/>

<field name="body" type="text_ja" ④
       indexed="true" stored="true" multiValued="true"/>

<field name="date" type="pdate"
       indexed="true" stored="true"/> ⑤

<fieldType name="pdate" class="solr.DatePointField" ⑥
           docValues="true"/>

<fieldType name="string" class="solr.StrField"/> ⑦
```

Solrでのスキーマ設定（つづき）

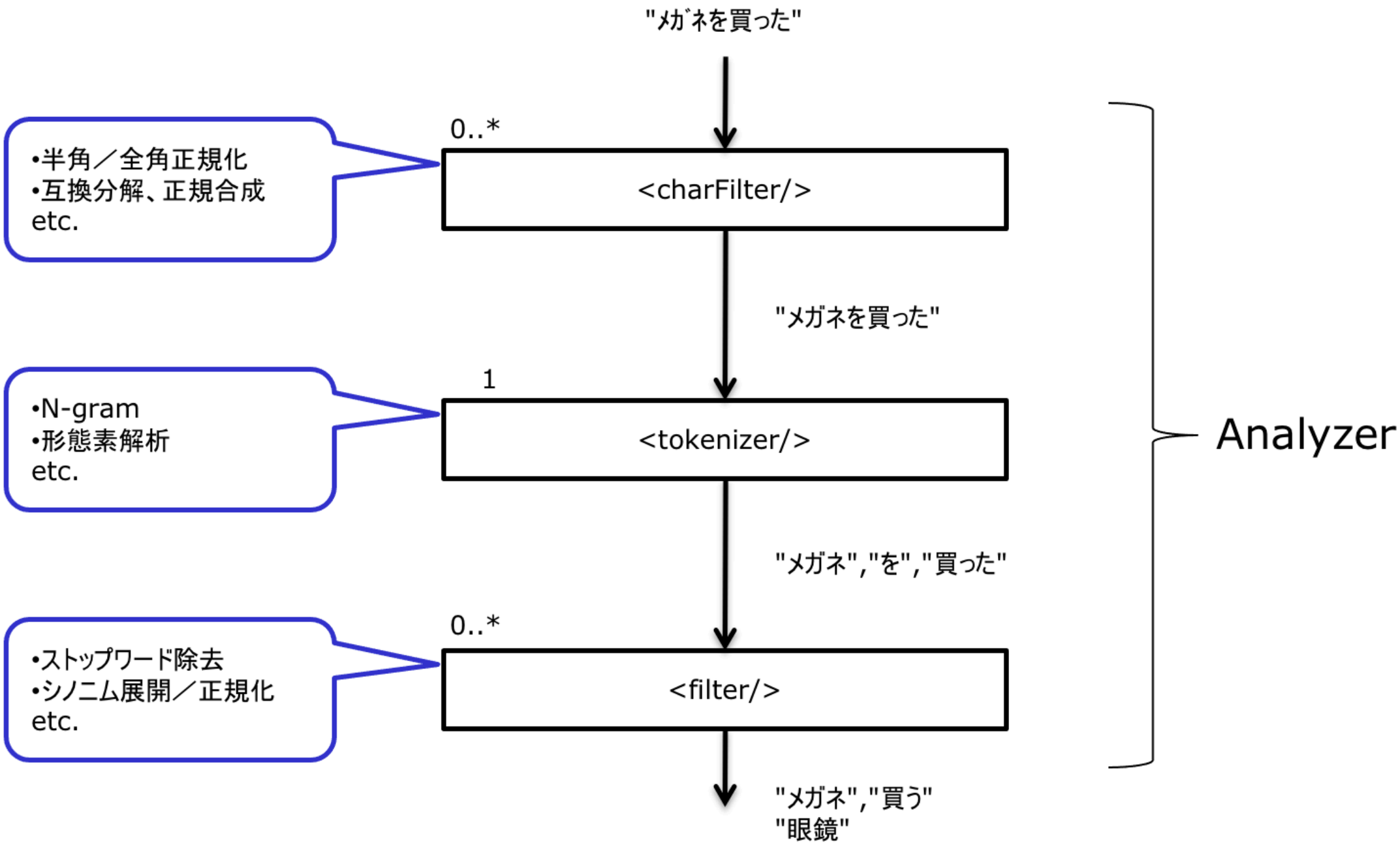
```
<fieldType name="text_ja" class="solr.TextField"
           positionIncrementGap="100"
           autoGeneratePhraseQueries="true">

  <analyzer type="index">
    <charFilter class="solr.MappingCharFilterFactory"
                mapping="mapping-ja.txt"/>
    <tokenizer class="solr.JapaneseTokenizerFactory"/>
    <filter class="solr.JapaneseBaseFormFilterFactory"/>
    <filter class="solr.JapaneseKatakanaStemFilterFactory"
            minimumLength="4"/>
  </analyzer>

  <analyzer type="query">
    <charFilter class="solr.MappingCharFilterFactory"
                mapping="mapping-ja.txt"/>
    <tokenizer class="solr.JapaneseTokenizerFactory"
               mode="search"/>
    <filter class="solr.JapaneseBaseFormFilterFactory"/>
    <filter class="solr.JapaneseKatakanaStemFilterFactory"
            minimumLength="4"/>
  </analyzer>

</fieldType>

<uniqueKey>url</uniqueKey>
```

0個以上のcharFilter

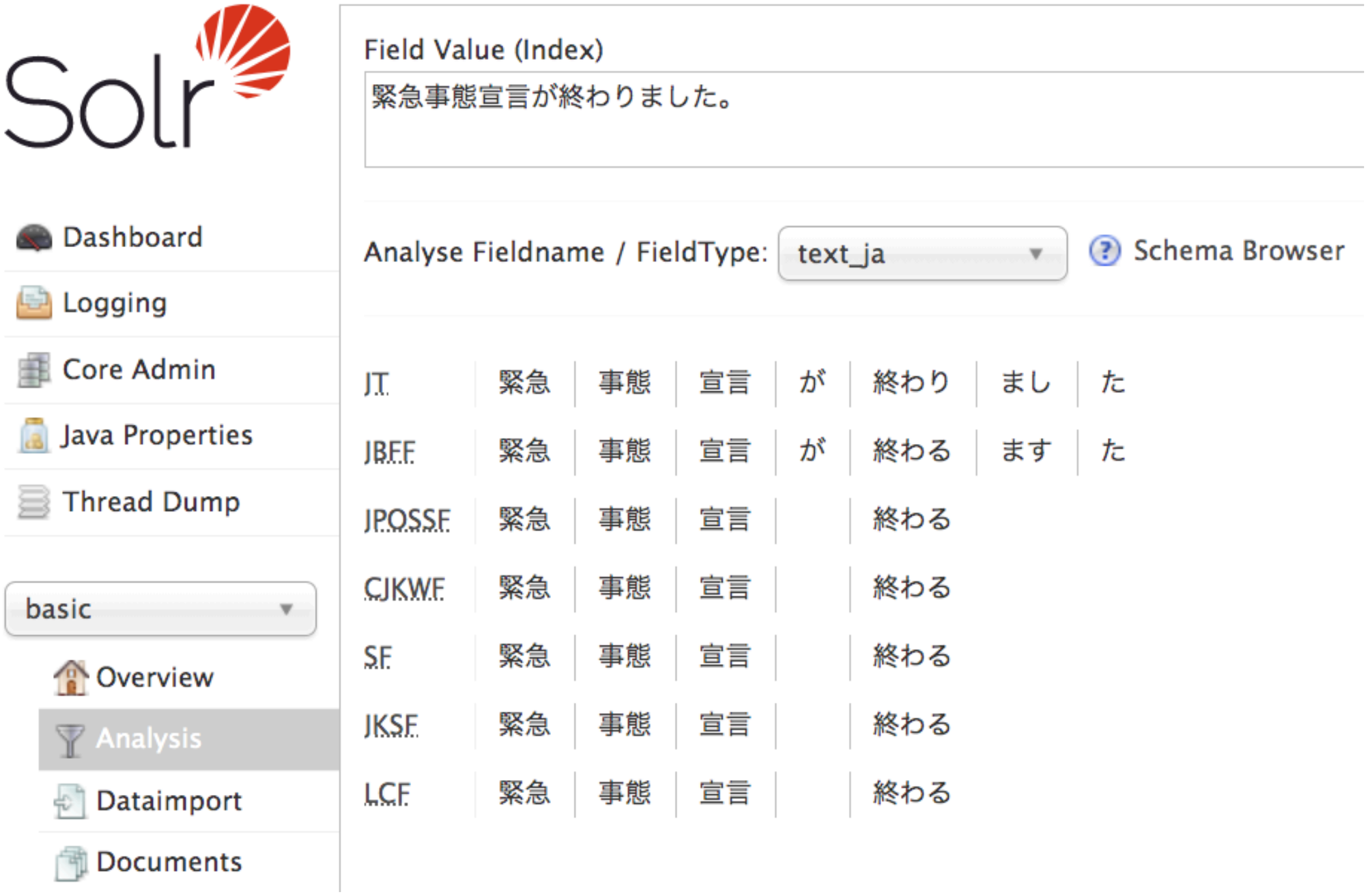
1個のtokenizer

0個以上のfilter

Analyzerの構造



Solr管理画面のAnalysis



トークンのAttribute

- Analyzerはトークンテキスト以外にAttributeを出力
  - ポジション（インクリメント）とオフセット

入力テキスト　こんにちは、世界。

↓

JapaneseTokenizer

↓

オフセット	0	1	2	3	4	5	6	7	8	9
トークンテキスト	こ	ん	に	ち	は	、	世	界	。	
ポジション	1					2	3		4	

日本語検索表記揺れの分類

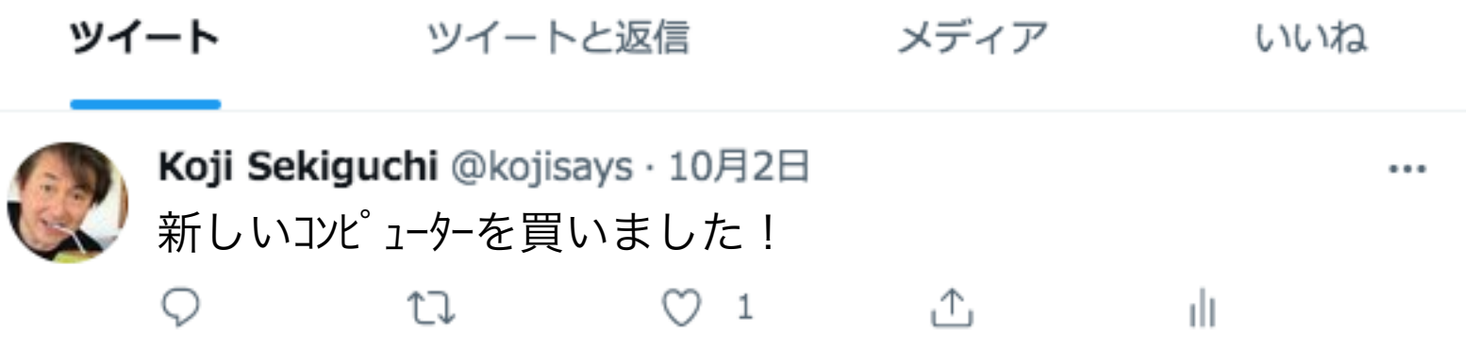
文字の正規化	半角全角／"ｶｰﾄﾞ"<->"カードﾞ"<->"カード" 新旧漢字／"慶應"<->"慶応"
同義語 **類義語**	同義語／"ピンポン"<->"卓球" 類義語／"言う"<->"話す"
頭文字略語 **省略語**	頭文字略語／"WHO"<->"World Health Org" 省略語／"木村拓哉"<->"キムタク"
外来語	"interface"<->"インターフェイス"<->"インタフェース"
漢字送り仮名	"引っ越し"<->"引越し"<->"引越" "受け付け"<->"受付け"<->"受付"
西暦・和暦	"2019年"<->"平成31年"<->"令和元年"

日本語検索の留意点

- MappingCharFilterで文字を正規化。
  - オフセット補正のしくみがあるので、ハイライトずれを起こさない。
  - 半角カナ⇒全角カナに正規化。「斉藤」で「斎藤」「齋藤」「齊藤」を検索。
- JapaneseTokenizerのAttributeの活用。
  - 読みで検索やソート、JapaneseKatakanaStemFilter で再現率向上
- SynonymGraphFilter で tokenizerFactory を忘れずに設定。
- N-gramと形態素解析を組み合わせる。
- autoGeneratePhraseQueries
  - true: 精度アップ、false: 再現率アップ

日本語検索ハイライトの要件

検索対象文書



検索キーワード

ｺﾝﾋﾟｭｰﾀｰ
コンピューター
コンピュータ
computer

[要件]
これらのいずれのキーワードでも
ヒット＆ハイライトさせたい

検索結果（ハイライトスニペット表示）

新しい **ｺﾝﾋﾟｭｰﾀｰ** を買いました！

日本語検索ハイライトの失敗例

- そもそもヒットしない
- 一部のキーワードではヒットするが、ヒットしないキーワードもある
    - 「ｺﾝﾋﾟｭｰﾀｰ」、「コンピューター」、「computer」はヒット
    - 「コンピュータ」、「Computer」はヒットしない
- ヒットはするが、ハイライトがずれる

    新しい ｺﾝﾋﾟｭｰﾀｰを 買いました！

<analyzer /> の設定の実際

```
<analyzer type="query">

  <charFilter class="solr.MappingCharFilterFactory"
              mapping="mapping.txt"/>

  <tokenizer class="solr.JapaneseTokenizerFactory"/>

  <filter class="solr.JapaneseKatakanaStemFilterFactory"
          minimumLength="4"/>

  <filter class="solr.LowerCaseFilterFactory"/>

  <filter class="solr.SynonymGraphFilterFactory"
          synonyms="synonyms.txt"
          tokenizerFactory="solr.JapaneseTokenizerFactory"
          ignoreCase="true"
          expand="true"/>

</analyzer>
```

mapping.txtの設定例

```
                :
<charFilter class="solr.MappingCharFilterFactory"
            mapping="mapping.txt"/>
                :
```

```
"ｺ" => "コ"          # 半角カナ ⇒ 全角カナへの正規化
     :
     :
"ﾋﾟ" => "ピ"

"齊" => "斉"          #  「斉藤」でいろいろな「サイトウさん」を広くヒットさせる
"斎" => "斉"
"齋" => "斉"

"髙" => "高"          #  「高橋」で「高橋」も「髙橋」もヒットさせる

"應" => "応"          #  「慶応義塾大学」で「慶應義塾大学」をヒットさせる
```

<charFilter /> と <tokenizer /> のオフセット補正（前）

0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
新	し	い	コ	ン	ヒ	゜	ュ	ー	タ	ー	を	買	い	ま	し	た	。	

・・・

MappingCharFilter

0	1	2	3	4	5	7	8	9	10	11	12	13	14	15	16	17	18
新	し	い	コ	ン	ピ	ュ	ー	タ	ー	を	買	い	ま	し	た	。	

JapaneseTokenizer

0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
新	し	い	コ	ン	ピ	ュ	ー	タ	ー	を	買	い	ま	し	た	。	

<charFilter /> と <tokenizer /> のオフセット補正（後）

0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
新	し	い	コ	ン	ピ	ュ	ー	タ	ー	を	買	い	ま	し	た	。		

MappingCharFilter

0	1	2	3	4	5	7	8	9	10	11	12	13	14	15	16	17	18
新	し	い	コ	ン	ピ	ュ	ー	タ	ー	を	買	い	ま	し	た	。	

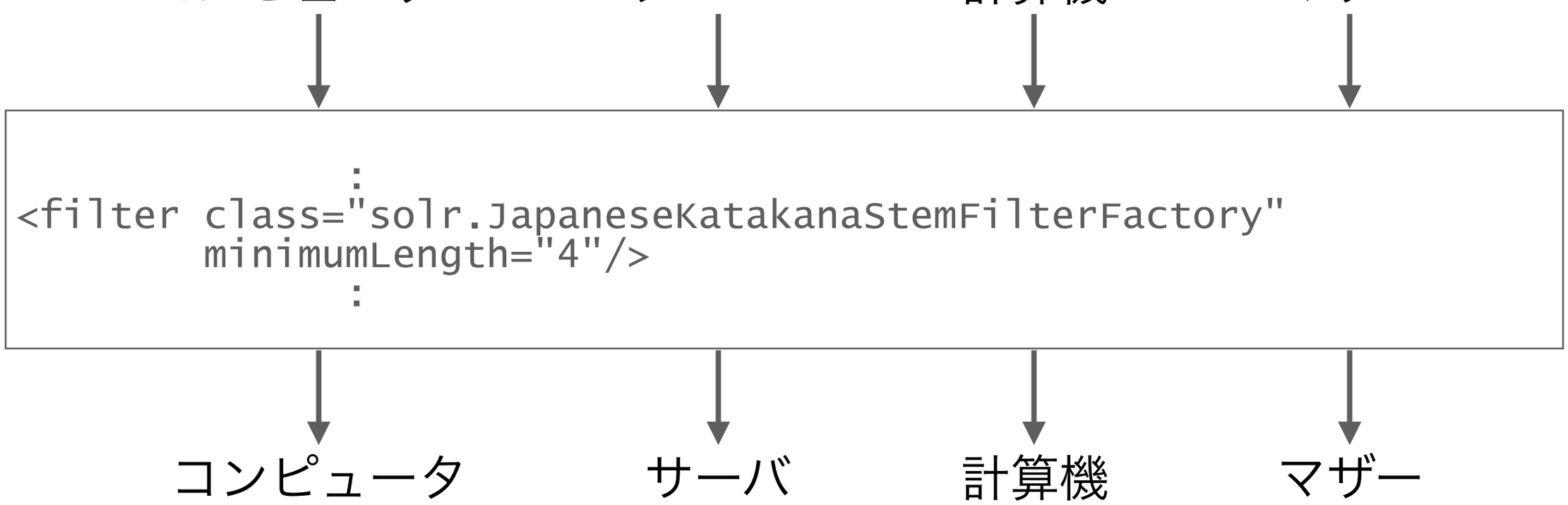
3? 3　　10? 11

0	1	2	3	4	5	6	7	8	9	11	11	12	13	14	15	16	17
新	し	い	コ	ン	ピ	ュ	ー	タ	ー	を	買	い	ま	し	た	。	

JapaneseTokenizer

JapaneseKatakana StemFilter

- JapaneseKatakanaStemFilter でカタカナ語の末尾の長音記号を除去（minimumLength以上の長さのカタカナ語に適用）



SynonymGraphFilter

- ファイルに設定したシノニム（類義語）で検索させる。
  - tokenizerFactory に上流で設定した Tokenizer を忘れずに指定すること。

```
                :
<filter class="solr.SynonymGraphFilterFactory"
        synonyms="synonyms.txt"
        tokenizerFactory="solr.JapaneseTokenizerFactory"
        ignoreCase="true"
        expand="true"/>
                :
```

synonyms.txt

```
コンピュータ, computer
マクドナルド, マック, マクド
```

N-gramフィールドを追加

```
<fieldType name="text_2g" class="solr.TextField"
           positionIncrementGap="100"
           autoGeneratePhraseQueries="true">

  <analyzer>
    <charFilter class="solr.MappingCharFilterFactory"
                mapping="mapping.txt"/>
    <tokenizer class="solr.NGramTokenizerFactory"
               minGramSize="2" maxGramSize="2"/>
   <filter class="solr.LowerCaseFilterFactory"/>
  </analyzer>

</fieldType>


<dynamicField name="*_2g" type="text_2g" />

<copyField source="title" dest="title_2g"/>
<copyField source="body" dest="body_2g"/>
```

edismax で横断検索

- タイトル（title）と本文（body）のそれぞれの形態素解析とN-gramフィールドを横断検索。

- 本文よりもタイトルフィールド、N-gramよりも形態素解析フィールドにより重みを置く。

- defType=edismax&qf=title^10 body^5 title_2g^5 body_2g

- 「タイトル ＞ 本文」はNormが効くはず。

まとめ

- Solrスキーマの基礎知識
- livedoorニュースコーパス
- トークンにはテキスト以外にポジション（インクリメント）とオフセットという特に重要なAttributeがある。
  - ハイライトずれを防ぐオフセット補正の機構が組み込まれている。
  - Analyzerの構造。
- 日本語シノニム検索とハイライトの設定ポイント

次回（予定）

- Apache Solr 活用事例
  - 皆様より。現在1社決定。
    - 10〜20分程度で事例紹介可能な方、アンケートにご記入ください。こちらから連絡させていただきます。
- 11月下旬（予定）

受講アンケート

- 次回以降の勉強会の参考とするため、ぜひ受講アンケートにご協力お願いします！（無記名可）

- 勉強会で発表できるネタをお持ちの方、本日の内容で質問のある方、次回以降で取り上げて欲しい内容のリクエストなどあれば、アンケートの自由記入欄にお書きください。

https://docs.google.com/forms/d/e/1FAIpQLSdZctPzCvb-k8gOPGR0sveAbeBMtje6XBXVplyUonl0zVxmuA/viewform?usp=sf_link

開始時間（11:00）まで
ビデオと音声をオフ（ミュート）にして
そのままお待ちください。

第25回 Lucene/Solr勉強会
オンライン／初心者編３