KuromojiとSynonymとパッチ(LUCENE-9123)

Jun Ohtani @johtani
Elasticsearch勉強会
2020/06/18

自己紹介

- フリーランスエンジニア
- Apache Solr入門(第2版まで)や
  データ分析基盤構築入門の著者の一人
- KibanaのAnalyze API UI pluginの作者

アジェンダ

- 日本語と類義語
- Ananlyzerとは?
- Kuromojiのおさらい
- Synonymのおさらい
- Kuromoji+Synonymの問題点
- LUCENE-9123

日本語と類義語

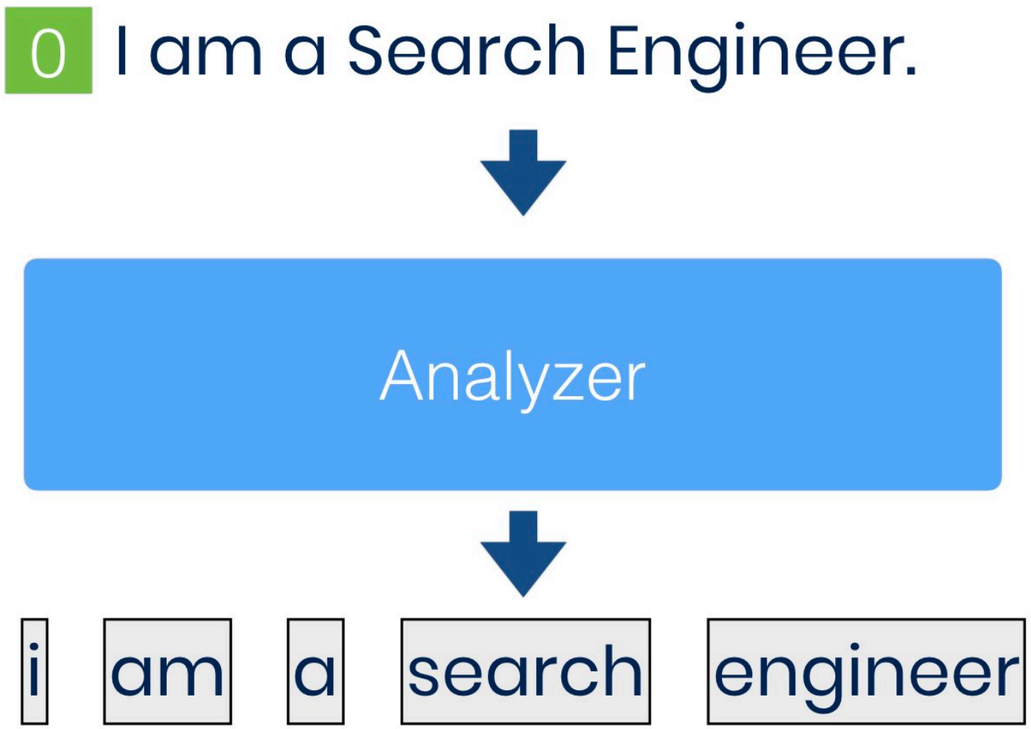
- こんなことありますよね?
  - 関西国際空港を関空で検索したい
  - 国際連合を国連で検索したい
  - コンビニエンスストアをコンビニで検索したい

Elasticsearchでやるには?

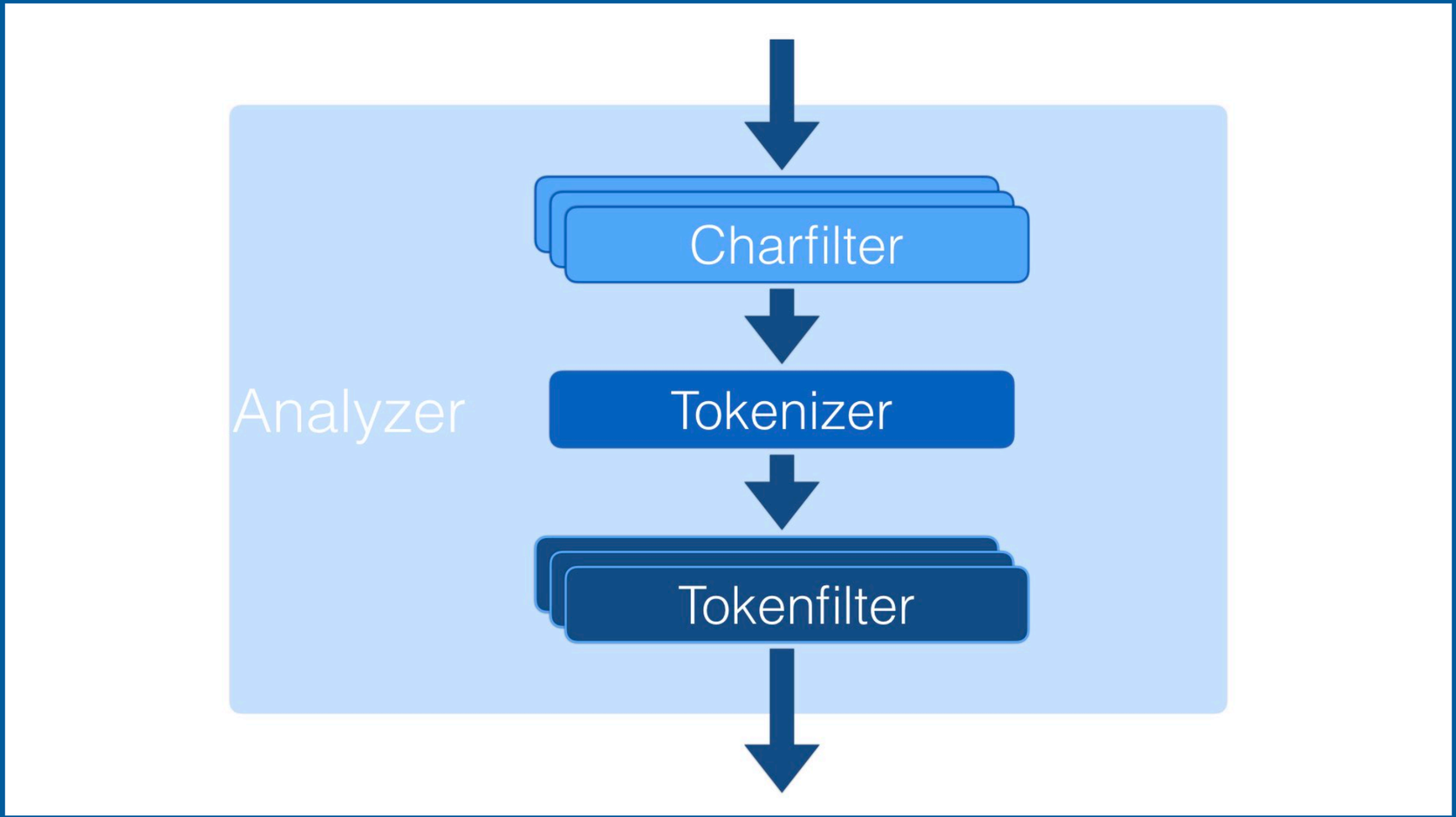
- Analyzer/Text Analysisでこんな感じ
  - 日本語区切り
    - Kuromoji Tokenizer
  - 類義語展開
    - Synonym Token Filter

Analyzerって
なんだっけ?

Analyzerとは?



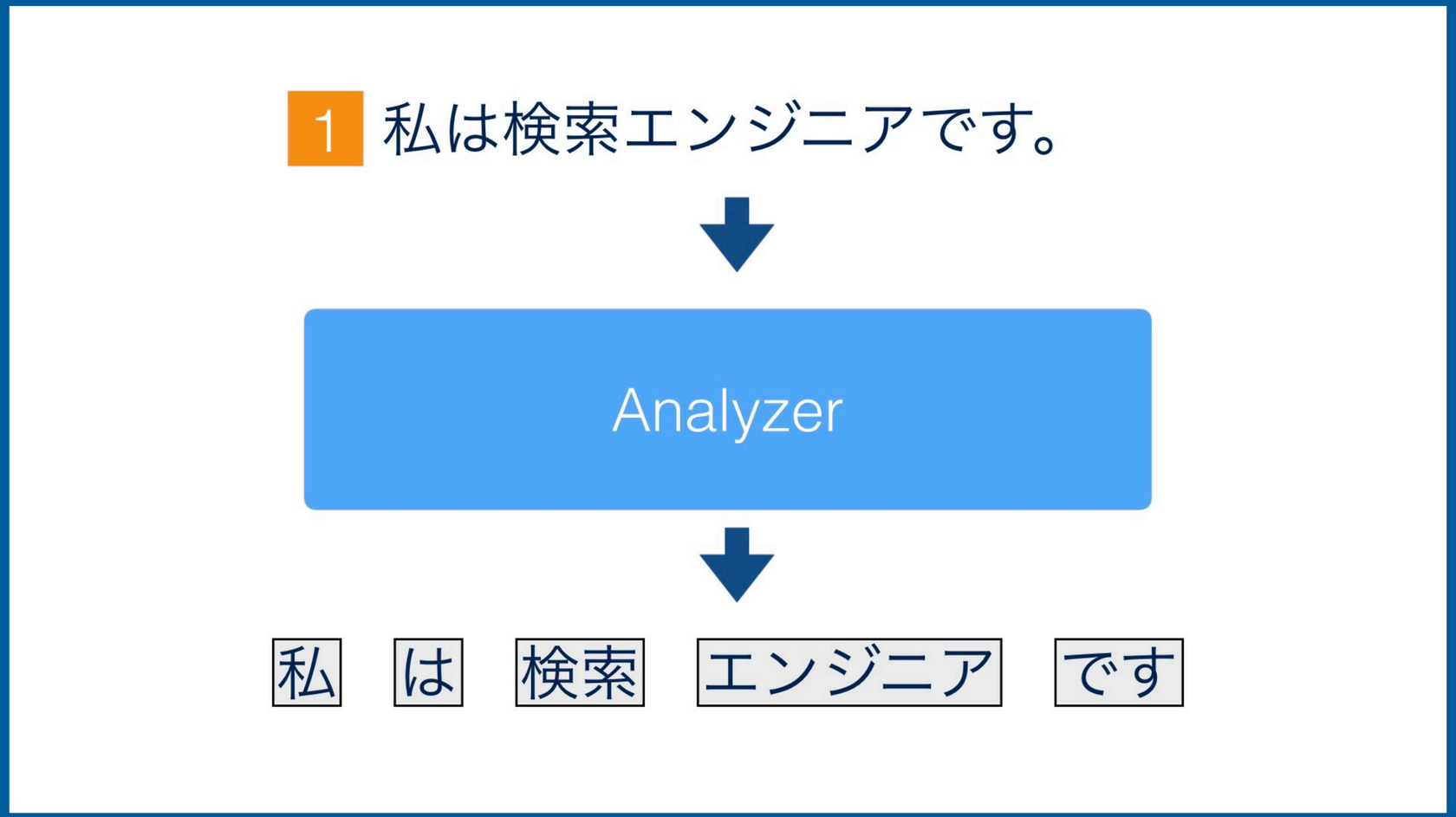
Analyzer - Elasticsearch



Kuromojiのおさらい

- Luceneの日本語用形態素解析ライブラリ
  - 辞書を用いて日本語の文章を単語に分割してくれる
- Elasticsearchでも公式プラグインとして利用可能
  - インストールが楽
  - Elastic Cloudなどですぐ使える

Kuromojiの挙動(Analyzer)



Synonymのおさらい

- 類義語を使った検索用のTokenFilter

```
{
  "en_synonym": {
    "type": "synonym_graph",
    "synonyms": ["ipod, i-pod, i pod"]
  }
}
```

Synonymの挙動

type name	tokens[0]	tokens[1]	tokens[2]	tokens[3]
tokenizer **whitespace**	token **I** position 0	**want** 1	**i-pod** 2	
filter **en_synonym**	token **I** position 0	**want** 1	**ipod** 2 **i** 2 **i-pod** 2	**pod** 3

Elasticsearchでやるには?

- Analyzer/Text Analysisでこんな感じ
  - 日本語区切り
    - Kuromoji Tokenizer
  - 類義語展開
    - Synonym Token Filter

じゃあ組み合わせると?

```
{"analyzer": {
    "kuromoji_synonym": {
      "type": "custom",
      "tokenizer": "kuromoji_tokenizer",
      "filter": ["ja_synonym"] } }
,"filter": {
    "ja_synonym": {
      "type": "synonym_graph",
      "synonyms": [
        "株式会社, コーポレーション"
] } } }
```

おや、エラーが

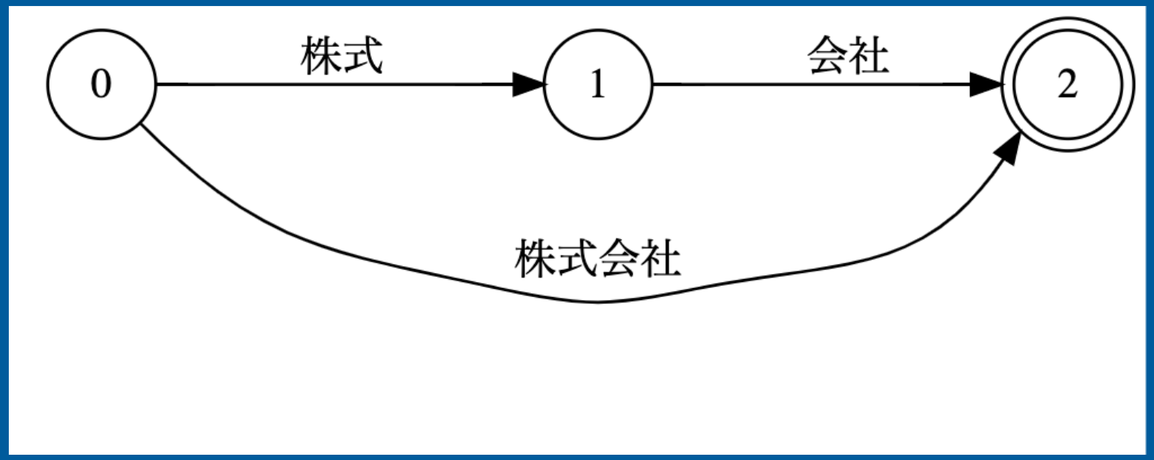

```
{
  "type" : "illegal_argument_exception",
  "reason" : "failed to build synonyms",
  "caused_by" : {
    "type" : "parse_exception",
    "reason" : "Invalid synonym rule at line 1",
    "caused_by" : {
      "type" : "illegal_argument_exception",
      "reason" : "term: 株式会社 analyzed to a token (株式会社) with position increment != 1 (got: 0)"
    }
  }
}
```

何が問題?

- 1. 類義語の設定を読み込む
  - "株式会社, コーポレーション"
- 2. 類義語の設定をTokenizerを使って解析
  - "株式会社" -> ["株式", "株式会社", "会社"]
  - "コーポレーション" -> ["コーポレーション"]
- 3. 解析結果の単語列を類義語として保持

何が問題?

```
{
  "tokens" : [
    {
      "token" : "株式",
      "position" : 0
    },
    {
      "token" : "株式会社",
      "position" : 0,
      "positionLength" : 2
    },
    {
      "token" : "会社",
      "position" : 1
    }
  ]
}
```



Kuromoji Tokenizerの設定

- mode - 複合語と未知語の扱いに関するモード
  - normal - 内部の辞書に基づき形態素解析。
  - search - デフォルト。長い名詞がある場合により短い単語に分割
  - extended - searchに加え、未知語は1文字ずつに分割

Kuromoji Tokenizerの設定

- mode - 複合語と未知語の扱いに関するモード
  - normal - 内部の辞書に基づき形態素解析。
  - search - デフォルト。長い名詞がある場合により短い単語に分割
  - extended - searchに加え、未知語は1文字ずつに分割

modeをnormalにしてみる


```
{
  "token" : "株式会社",
  "start_offset" : 0,
  "end_offset" : 4,
  "type" : "word",
  "position" : 0
}
```

エラーはでない

```
{
  "analyzer": {
  "kuromoji_synonym": {
    "type": "custom",
    "tokenizer": "ja_tokenizer",
    "filter": ["ja_synonym"]
}}
```

```
{"tokenizer": {
  "ja_tokenizer": {
    "type": "kuromoji_tokenizer",
    "mode": "normal"  }},
 "filter": {
  "ja_synonym": {
    "type": "synonym_graph",
    "synonyms": [
      "株式会社, コーポレーション"
    ]  }}}
```

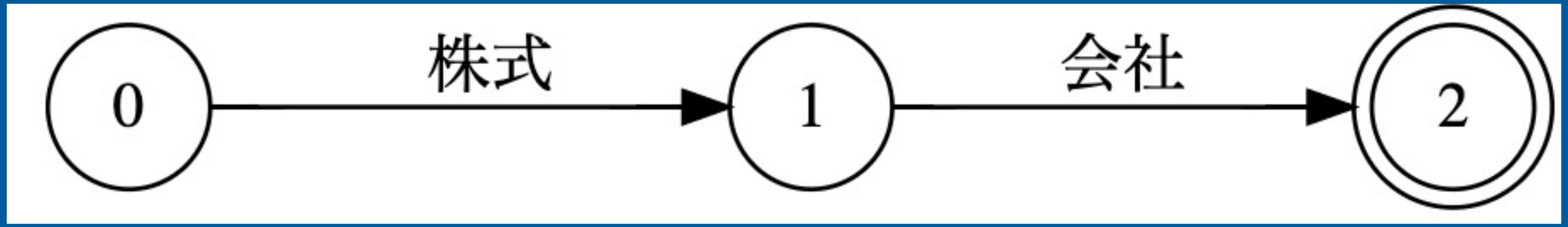
エラーはでないけど

- 株式と会社も出てこない => これらの単語で検索できない

```
{
  "token" : "株式会社",
  "start_offset" : 0,
  "end_offset" : 4,
  "type" : "word",
  "position" : 0
}
```

LUCENE-9123

- LUCENE-9123
- modeがsearchかextendedのときに、複合語をトークン列に出力しない。



Elasticsearchでは?

- 7.9.0以降に利用可能に

```
{
  "ja_tokenizer": {
    "type": "kuromoji_tokenizer",
    "discard_compound_token": true
  }
}
```

Elasticsearchでは?

- 7.9.0以降に利用可能に

```
{
  "ja_tokenizer": {
    "type": "kuromoji_tokenizer",
    "discard_compound_token": true
  }
}
```

まとめ

- 7.8までは"mode": "normal"でエラーが出ない
- 7.9以降では"discard_compound_token": trueオプションで