Rozlišovacie stránky

Vyhľadávanie informácií

Dávid Kecskés

Cieľ projektu

Nájsť všetky rozlišovacie stránky na Wikipédii a vyparsovať z nich:

- Nadpis
- Odkaz
- Popis

Dáta	**Jazyk**	**Vyhľadávanie**
Slovenská Wikipédia	Java	Elasticsearch

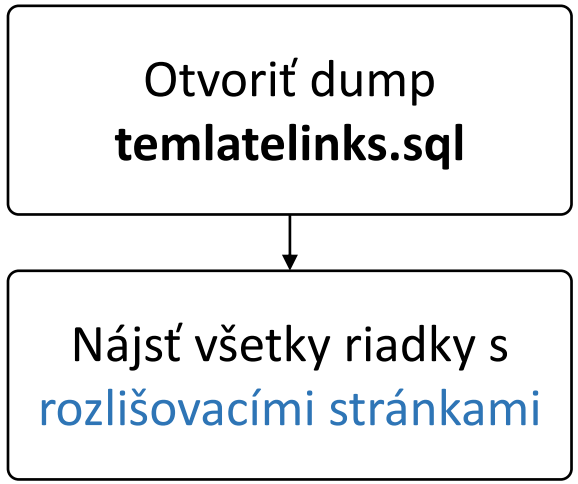
Riešenie

Otvoriť dump
temlatelinks.sql

templatelinks.sql

INSERT INTO 'templatelinks' VALUES (891,10,'Rozlišovacia_stránka',0),
(1932,10,'Rozlišovacia_stránka',0),(4345,10,'Rozlišovacia_stránka',0),
(4762,10,'Rozlišovacia_stránka',0),(5684,10,'Rozlišovacia_stránka',0),
(7653,10,'Rozlišovacia_stránka',0),(8912,10,'Rozlišovacia_stránka',0),
INSERT INTO 'templatelinks' VALUES (1891,10,'Rozlišovacia_stránka',0),
(21932,10,'Rozlišovacia_stránka',0),(34345,10,'Rozlišovacia_stránka',0),
(44762,10,'Rozlišovacia_stránka',0),(55684,10,'Rozlišovacia_stránka',0),
(67653,10,'Rozlišovacia_stránka',0),(78912,10,'Rozlišovacia_stránka',0)

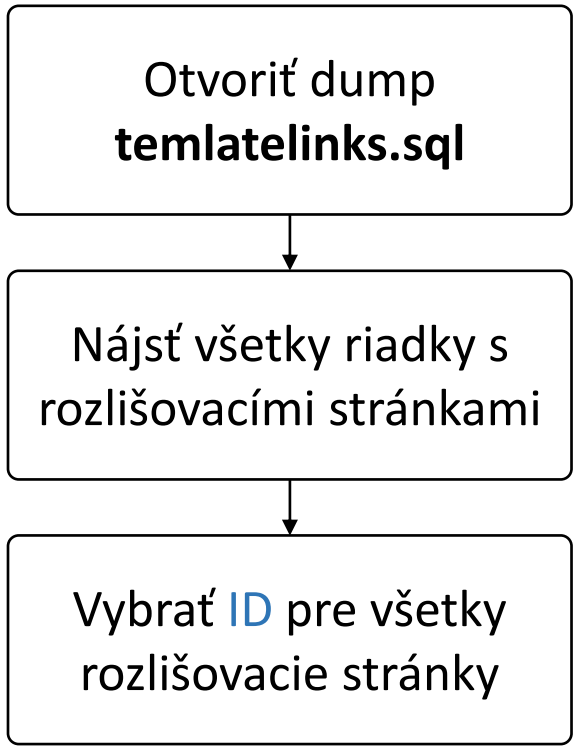
Riešenie



templatelinks.sql

INSERT INTO 'templatelinks' VALUES (891,10,'Rozlišovacia_stránka',0),
(1932,10,'Rozlišovacia_stránka',0),(4345,10,'Rozlišovacia_stránka',0),
(4762,10,'Rozlišovacia_stránka',0),(5684,10,'Rozlišovacia_stránka',0),
(7653,10,'Rozlišovacia_stránka',0),(8912,10,'Rozlišovacia_stránka',0),
INSERT INTO 'templatelinks' VALUES (1891,10,'Rozlišovacia_stránka',0),
(21932,10,'Rozlišovacia_stránka',0),(34345,10,'Rozlišovacia_stránka',0),
(44762,10,'Rozlišovacia_stránka',0),(55684,10,'Rozlišovacia_stránka',0),
(67653,10,'Rozlišovacia_stránka',0),(78912,10,'Rozlišovacia_stránka',0)

Riešenie



templatelinks.sql

INSERT INTO 'templatelinks' VALUES (891,10,'Rozlišovacia_stránka',0),
(1932,10,'Rozlišovacia_stránka',0),(4345,10,'Rozlišovacia_stránka',0),
(4762,10,'Rozlišovacia_stránka',0),(5684,10,'Rozlišovacia_stránka',0),
(7653,10,'Rozlišovacia_stránka',0),(8912,10,'Rozlišovacia_stránka',0),
INSERT INTO 'templatelinks' VALUES (1891,10,'Rozlišovacia_stránka',0),
(21932,10,'Rozlišovacia_stránka',0),(34345,10,'Rozlišovacia_stránka',0),
(44762,10,'Rozlišovacia_stránka',0),(55684,10,'Rozlišovacia_stránka',0),
(67653,10,'Rozlišovacia_stránka',0),(78912,10,'Rozlišovacia_stránka',0)

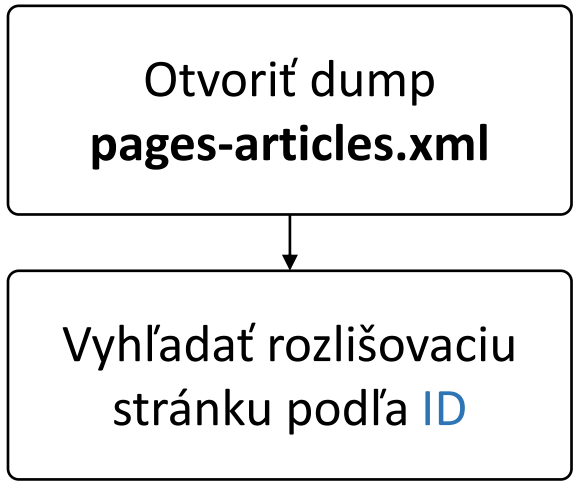
Riešenie

Otvoriť dump
pages-articles.xml

pages-articles.xml

```
<page>
  <title>História</title>
  <id>891</id>
  <text>
    '''História''' môže byť:
    * [[dejiny]]
    * [[historická veda]]
    * priebeh deja nejakej udalosti
    * časť historickej vedy spracúvajúcej len písomné pramene (opak [[prehistória|prehistórie]])
  <text>
</page>
```

Riešenie



pages-articles.xml

```
<page>
  <title>História</title>
  <id>891</id>
  <text>
    '''História''' môže byť:
    * [[dejiny]]
    * [[historická veda]]
    * priebeh deja nejakej udalosti
    * časť historickej vedy spracúvajúcej len písomné pramene (opak [[prehistória|prehistórie]])
  <text>
</page>
```

Riešenie

Vybrať názov rozlišovacej stránky

pages-articles.xml

```
<page>
  <title>História</title>
  <id>891</id>
  <text>
    '''História''' môže byť:
    * [[dejiny]]
    * [[historická veda]]
    * priebeh deja nejakej udalosti
    * časť historickej vedy spracúvajúcej len písomné pramene (opak [[prehistória|prehistórie]])
  <text>
</page>
```

Riešenie

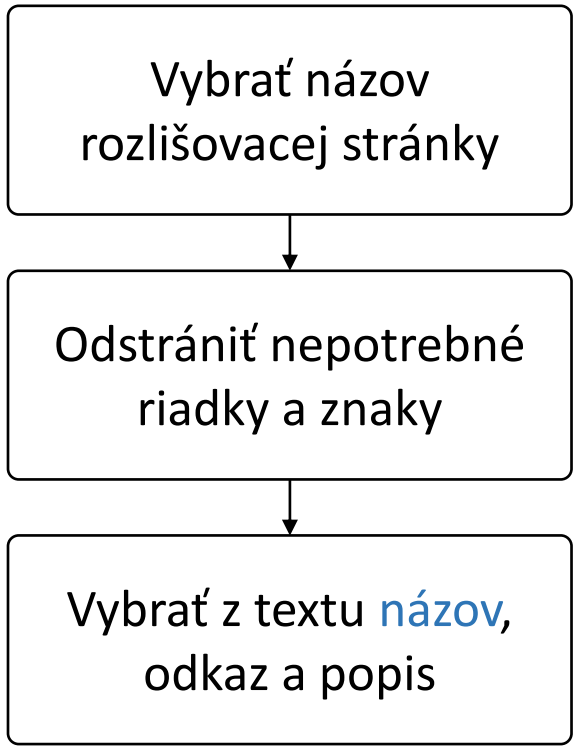
pages-articles.xml

<page>
  <title>História</title>
  <id>891</id>
  <text>

    [[dejiny]]
    [[historická veda]]

    časť historickej vedy spracúvajúcej len písomné pramene (opak [[prehistória|prehistórie]])
  <text>
</page>

Riešenie



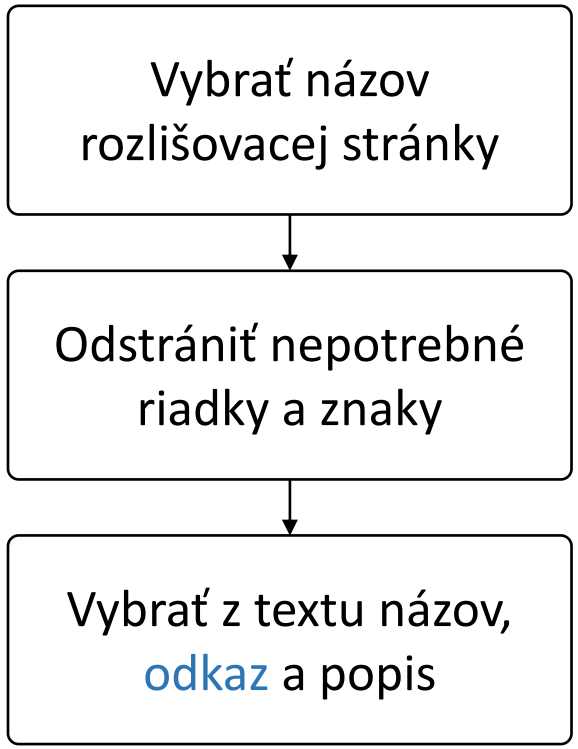
pages-articles.xml

```
<page>
  <title>História</title>
  <id>891</id>
  <text>

      [[dejiny]]
      [[historická veda]]

      časť historickej vedy spracúvajúcej len písomné pramene (opak [[prehistória|prehistórie]])
  <text>
</page>
```

Riešenie



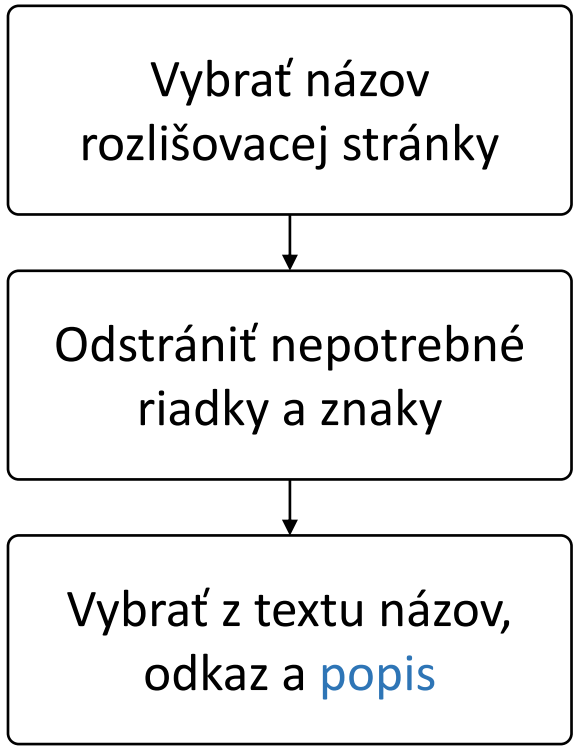
pages-articles.xml

```
<page>
  <title>História</title>
  <id>891</id>
  <text>

      [[dejiny]]
      [[historická veda]]

      časť historickej vedy spracúvajúcej len písomné pramene (opak [[prehistória|prehistórie]])
  <text>
</page>
```

Riešenie



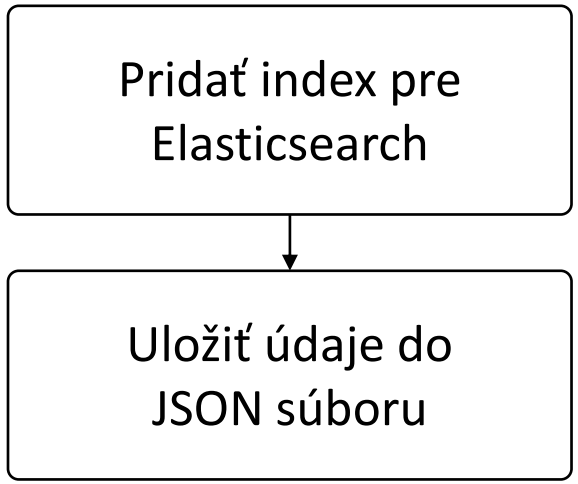
pages-articles.xml

```
<page>
  <title>História</title>
  <id>891</id>
  <text>

      [[dejiny]]
      [[historická veda]]

      časť historickej vedy spracúvajúcej len písomné pramene (opak [[prehistória|prehistórie]])
  <text>
</page>
```

Riešenie



result.json

{"index":{"_index":"vinf","_id":0}}

{"name":"História","title":"Dejiny","anchor":"dejiny","description":"História | Dejiny"}

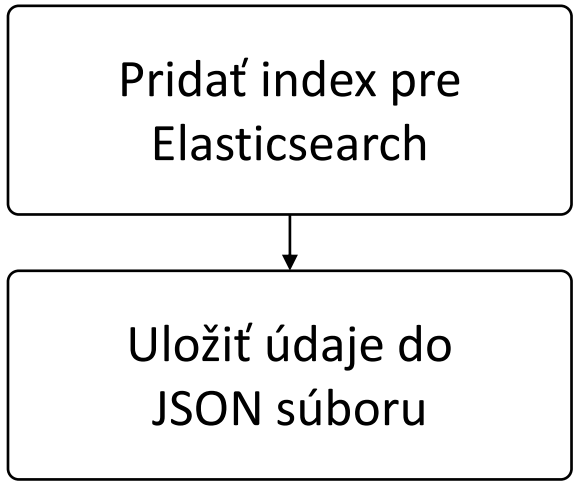
{"index":{"_index":"vinf","_id":1}}

{"name":"História","title":"Historická veda","anchor":"historická veda","description":"História | Historická veda"}

{"index":{"_index":"vinf","_id":2}}

{"name":"História","title":"Prehistória","anchor":"prehistórie","description":"História | Časť historickej vedy spracúvajúcej len písomné pramene (opak prehistórie)"}

Riešenie



result.json

{"index":{"_index":"vinf","_id":0}}

{"name":"História","title":"Dejiny","anchor":"dejiny","description":"História | Dejiny"}

{"index":{"_index":"vinf","_id":1}}

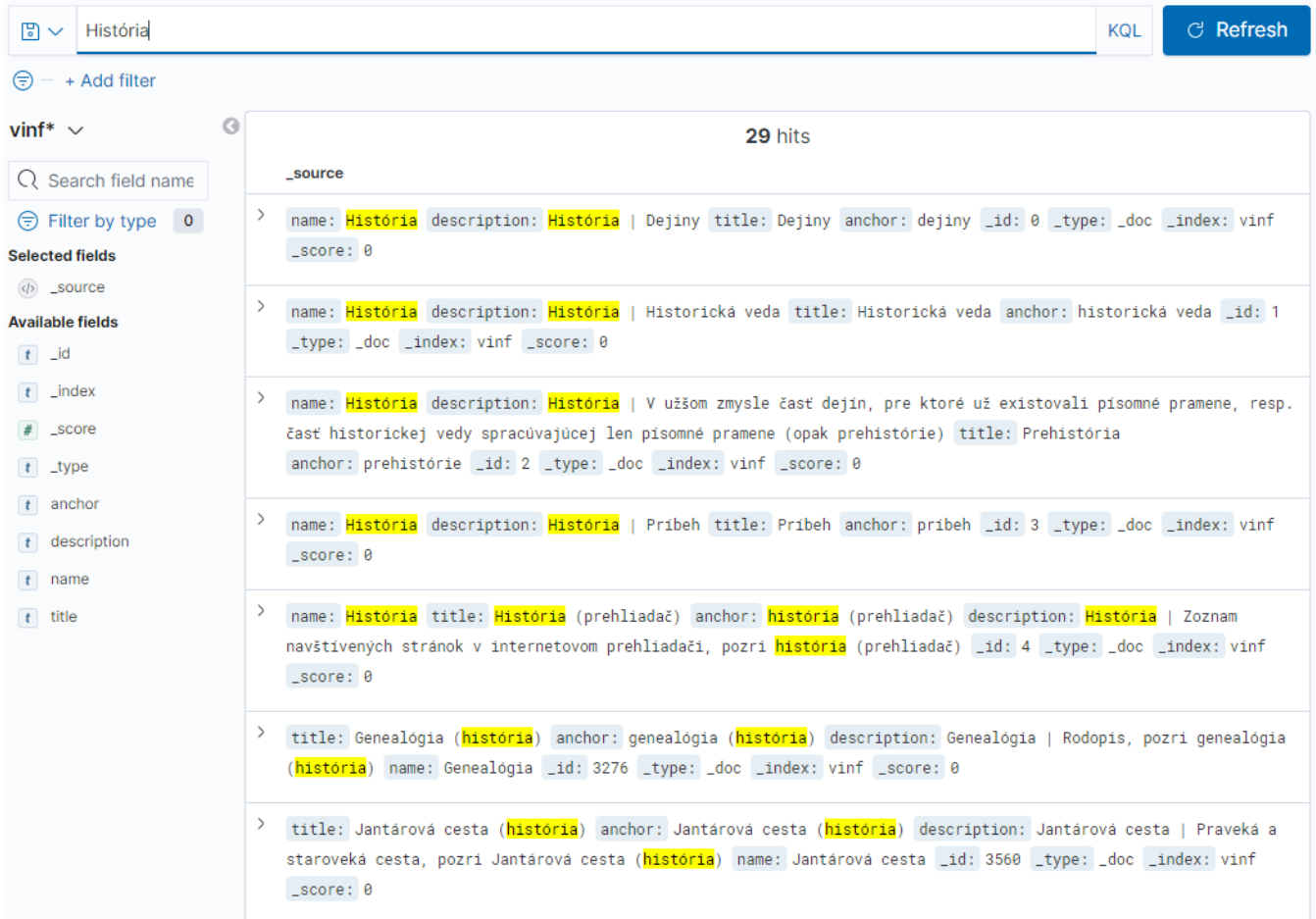
{"name":"História","title":"Historická veda","anchor":"historická veda","description":"História | Historická veda"}

{"index":{"_index":"vinf","_id":2}}

{"name":"História","title":"Prehistória","anchor":"prehistórie","description":"História | Časť historickej vedy spracúvajúcej len písomné pramene (opak prehistórie)"}

Vyhľadávanie Elasticsearch & Kibana

Discover Tool



Console

```
GET /vinf/_search?q=*História*
```


```
"hits" : {
  "total" : {
    "value" : 34,
    "relation" : "eq"
  },
  "max_score" : 1.0,
  "hits" : [
    {
      "_index" : "vinf",
      "_type" : "_doc",
      "_id" : "0",
      "_score" : 1.0,
      "_source" : {
        "name" : "História",
        "title" : "Dejiny",
        "anchor" : "dejiny",
        "description" : "História | Dejiny"
      }
    },
    {
      "_index" : "vinf",
      "_type" : "_doc",
      "_id" : "1",
      "_score" : 1.0,
      "_source" : {
        "name" : "História",
        "title" : "Historická veda",
        "anchor" : "historická veda",
        "description" : "História | Historická veda"
      }
    },
    {
      "_index" : "vinf",
      "_type" : "_doc",
      "_id" : "2",
      "_score" : 1.0,
      "_source" : {
        "name" : "História",
        "title" : "Prehistória",
        "anchor" : "prehistórie",
        "description" : "História | V užšom zmysle časť dejín, pre ktoré už existovali písomné pramene, resp.
          časť historickej vedy spracúvajúcej len písomné pramene (opak prehistórie)"
      }
    },
```

Vyhodnotenie

- Manuálna kontrola na 50 stránkach
- Počet rozlišovacích stránok: 22 748
- Priemer odkazov na stránke: 4.89
- Medián odkazov na stránke: 3

- Najviac odkazov na stránky: Újezd (172)
- Najčastejšie slovo v názve stránky: Ulica (772)
- Najkratší názov rozlišovacej stránky: Y
- Najdlhší názov rozlišovacej stránky: Majstrovstvá sveta v ľadovom hokeji 2012 - nižšie výkonnostné kategórie