

Rozlišovacie stránky

Vyhľadávanie informácií

Dávid Kecskés

Cieľ projektu

Nájsť všetky rozlišovacie stránky na Wikipédii a vyparsovať z nich:

- Nadpis
- Odkaz
- Popis

Dáta

Slovenská Wikipédia

Jazyk

Java

Vyhľadávanie

Elasticsearch

Riešenie

Otvoriť dump
temlatelinks.sql

temlatelinks.sql

```
INSERT INTO 'temlatelinks' 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 'temlatelinks' 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
temlatelinks.sql

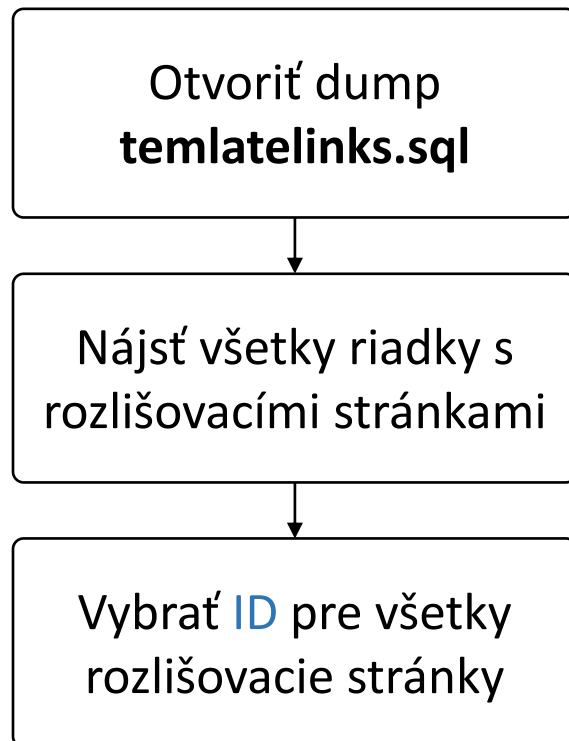


Nájsť všetky riadky s
rozlišovacími stránkami

temlatelinks.sql

```
INSERT INTO 'temlatelinks' 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 'temlatelinks' 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



temlatelinks.sql

```
INSERT INTO 'temlatelinks' 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 'temlatelinks' 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]]

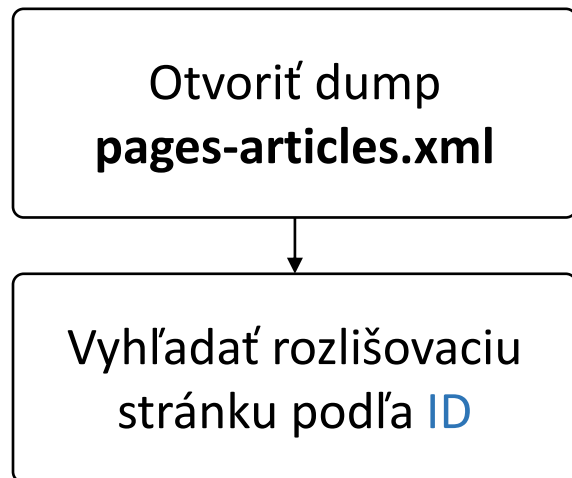
* priebeh deja nejakej udalosti

* 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

Vybrať názov
rozlišovacej stránky



Odstrániť nepotrebné
riadky a znaky

pages-articles.xml

```
<page>
```

```
<title>História</title>
```

```
<id>891</id>
```

```
<text>
```

```
[[dejiny]]
```

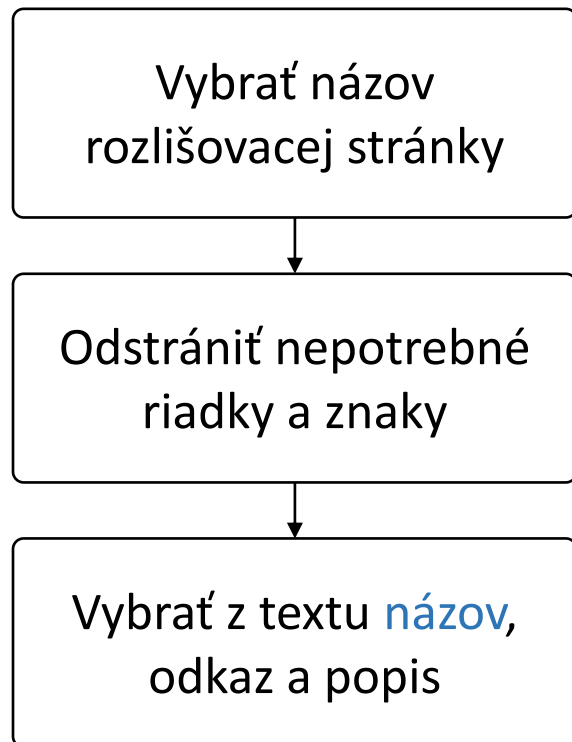
```
[[historická veda]]
```

```
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](#)]]

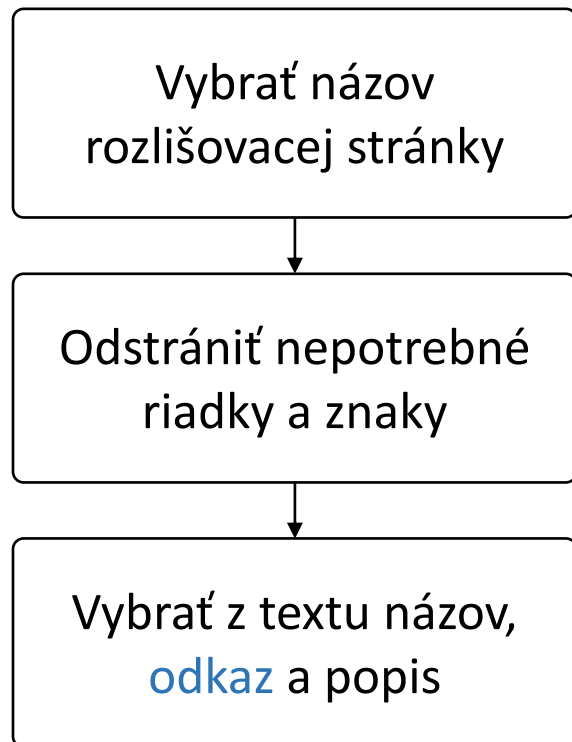
[[[historická veda](#)]]

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](#)]]

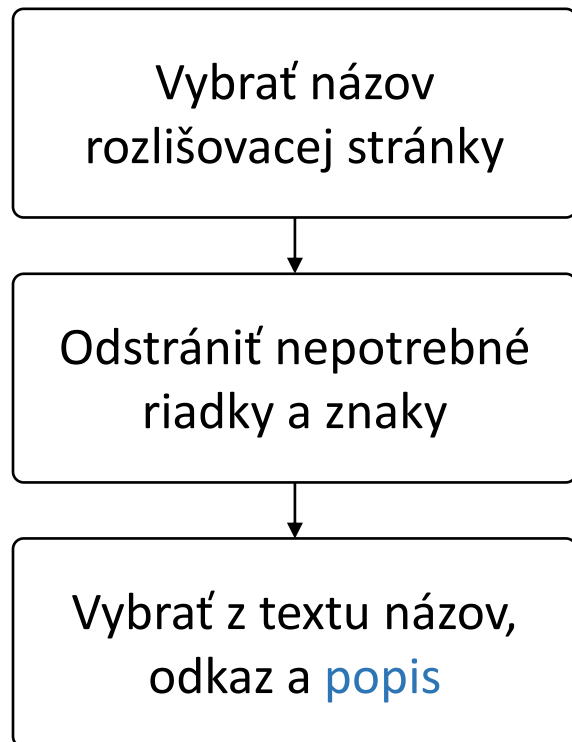
[[[historická veda](#)]]

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

Pridať index pre
Elasticsearch



Uložiť údaje do
JSON súboru

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

Pridať index pre
Elasticsearch



Uložiť údaje do
JSON súboru

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

Discover Tool interface showing search results for "História".

Search bar: História

Filter: vinf* (29 hits)

Selected fields: _source

Available fields: _id, _index, _score, _type, anchor, description, name, title

Search results (29 hits):

- name: História description: História | Dejiny title: Dejiny anchor: dejiny _id: 0 _type: _doc _index: vinf _score: 0
- name: História description: História | Historická veda title: Historická veda anchor: historická veda _id: 1 _type: _doc _index: vinf _score: 0
- name: História description: História | V užšom zmysle časť dejín, pre ktoré už existovali písomné pramene, resp. časť historickej vedy spracúvajúcej len písomné pramene (opak prehistórie) title: Prehistória anchor: prehistória _id: 2 _type: _doc _index: vinf _score: 0
- name: História description: História | Príbeh title: Príbeh anchor: príbeh _id: 3 _type: _doc _index: vinf _score: 0
- name: História title: História (prehliadač) anchor: história (prehliadač) description: História | Zoznam navštívených stránok v internetovom prehliadači, pozri história (prehliadač) _id: 4 _type: _doc _index: vinf _score: 0
- title: Genealógia (história) anchor: genealógia (história) description: Genealógia | Rodopis, pozri genealógia (história) name: Genealógia _id: 3276 _type: _doc _index: vinf _score: 0
- title: Jantárová cesta (história) anchor: Jantárová cesta (história) description: Jantárová cesta | Praveká a staroveká cesta, pozri Jantárová cesta (história) name: Jantárová cesta _id: 3560 _type: _doc _index: vinf _score: 0

Console

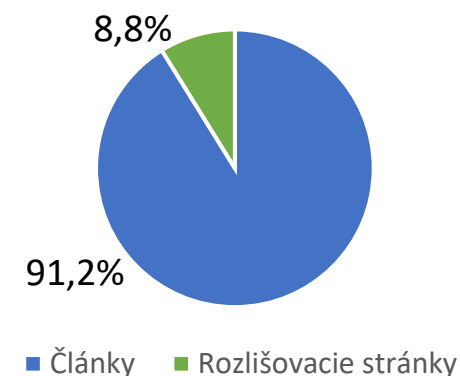
```
GET /vinf/_search?q=*História*

{"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": "prehistória",
        "description": "História | V užšom zmysle časť dejín, pre ktoré už existovali písomné pramene, resp. časť historickej vedy spracúvajúcej len písomné pramene (opak prehistórie)"
      }
    },
    {
      "_index": "vinf",
      "_type": "_doc",
      "_id": "3",
      "_score": 1.0,
      "_source": {
        "name": "História",
        "title": "Príbeh",
        "anchor": "príbeh",
        "description": "História | Príbeh"
      }
    },
    {
      "_index": "vinf",
      "_type": "_doc",
      "_id": "4",
      "_score": 1.0,
      "_source": {
        "name": "História",
        "title": "Zoznam navštívených stránok v internetovom prehliadači, pozri história (prehliadač)",
        "anchor": "história (prehliadač)",
        "description": "História | Zoznam navštívených stránok v internetovom prehliadači, pozri história (prehliadač)"
      }
    },
    {
      "_index": "vinf",
      "_type": "_doc",
      "_id": "3276",
      "_score": 1.0,
      "_source": {
        "name": "Genealógia",
        "title": "Genealógia",
        "anchor": "genealógia",
        "description": "Genealógia | Rodopis, pozri genealógia"
      }
    },
    {
      "_index": "vinf",
      "_type": "_doc",
      "_id": "3560",
      "_score": 1.0,
      "_source": {
        "name": "Jantárová cesta",
        "title": "Jantárová cesta",
        "anchor": "Jantárová cesta",
        "description": "Jantárová cesta | Praveká a staroveká cesta, pozri Jantárová cesta"
      }
    }
  ]
}
```

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

Slovenská Wikipédia



Anglická Wikipédia

