Slovenská technická univerzita
Fakulta informatiky a informačných technológii
Ilkovičova 3, 812 19 Bratislava

Martinus
Zadanie 1.

Vyhľadávanie informácií

Peter Válka
18/10/2018

Cvičiaci: Ing. Jakub Mačina
Študijný odbor: Inteligentné Softvérové Systémy
Ročník: 2. Ing
Akademický rok: 2018/2019

1. Úvod

Cieľom zadania bolo získať zložitý dataset z internetu. Zdroj informácií pre vytvorenie datasetu bol zvolený internetový portál martinus.sk, ktorý poskytuje informácie o knihách ktoré ponúka v obchodoch. Ako programovací jazyk crawlera a parsera bol zvolený Python, pretože ponúka intuitívne knižnice pre prácu s získavaním informácií z martinusu. Na sprostredkovanie rýchleho vyhľadávania nad datasetom bola zvolená technológia Elasticsearch, do ktorej parser ukladá informácie o knihách.

2. Crawler

Cieľom programu je získať informácie o knihách, ktoré sú ponúkané v martinuse. Crawler funguje v 2 módoch. Prvý mód inkrementuje ID knihy v URL a podľa toho postupne získava a zapisuje knihy do priečinka. Druhý mód je viac sofistikovanejší. Začína výberom náhodnej knihy. Následne v 1. kroku vyberie z dostupných kníh ktoré má k dispozícii jednu (v prvej iterácii je to tá náhodne zvolená) a vyšle požiadavku o stránku o knihe na server. Následne získa ID ďalších kníh zo sekcie Odporúčané, zapíše si ich do radu pre dostupné knihy a pokračuje ďalšou iteráciou. Algoritmus si zapisuje knihy, ktoré prešiel a teda sa nezacyklí. Pokiaľ mu došli dostupné knihy, vygeneruje sa náhodné ID novej knihy a pokračuje sa od začiatku.

3. Parser

Program číta stránky, ktoré uložil crawler do priečinku, získa z nich potrebné informácie o knihe, vytvorí slovník knihy a nakoniec uloží knihu do Elastic indexu. Atribúty knihy a ich Elastic typy sú nasledovné:

- name (text): Meno knihy
- description (text): Opis knihy
- pages (integer): Počet strán knihy
- comments (nested): Zložený typ, každý komentár má obsah komentára a ohodnotenie knihy
- author (text): Meno autora knihy
- rating (float): Celkové ohodnotenie knihy

- category (array): Pole kategórií knihy
- catalogue_number (integer): Katalógové číslo knihy
- year (integer): Rok publikácie knihy
- language (text): Jazyk knihy
- publisher (text): Vydavateľstvo
- cover (text): Väzba knihy
- review (text): Martinusove ohodnotenie knihy
- isbn (keyword): ISBN číslo knihy
- price (float): Cena knihy

4. Elastic Index

Vytvorenie indexu:

Ako všeobecný analýzer bola použitá slovenčina, ktorá obsahuje aj synonymický slovník, ktorý pomáha pri vyhľadávaní. Query pre vytvorenie indexu sa nachádza v prílohe

Mapovanie Indexu:

Pri určitých atribútoch bola použitá metóda n-gramov, ako napr opis knihy, lebo samotný opis knihy nebol jedine v slovenčine, ale mohol byť aj v angličtine alebo češtine. Vytvorenie mapovania indexu sa nachádza v prílohe

5. Elastic Query

Nad indexom som vytvoril následne 6 rôznych query. Query používajú (rozsah, “must not”, agregáciu, boostovanie parametrov, filter, viac atribútové vyhľadávanie, “More like this” typ query, fuzzy vyhľadávanie, regexp, vnorená query). Všetky query sú zobrazené v prílohe.

- **Cenový priemer knihy**

Query je zložená z 2 častí. Prvá časť vypočíta priemernú cenu knihy v indexe. Druhá časť následne vráti iba knihy, ktoré majú cenu vyššiu ako cenový priemer a majú kategóriu “dejiny”.

- **Knihy, ktoré majú komentár s hodnotením 5 hviezdičiek**

Query nájde knihy, ktoré majú komentár s 5 hviezdičkami, pričom je použitá metóda “nested”.

- **Viac atribútová query s boostingom**

Query hľadá výrazy vo viacerých atribútoch, pričom ak nájde zhodu v atribúte názov knihy a meno autora, výsledné skóre je boostnuté o 10. Query dodatočne obsahuje klauzulu should, kde sú rôzne skupiny kategórií a daná kniha musí mať aspoň jednu svoju kategóriu v skupine.

- **“More like this” query**

Query potrebuje katalógové číslo knihy a snaží sa nájsť všetky knihy, ktoré sa podobajú zvolenej knihe.

- **Fuzzy query**

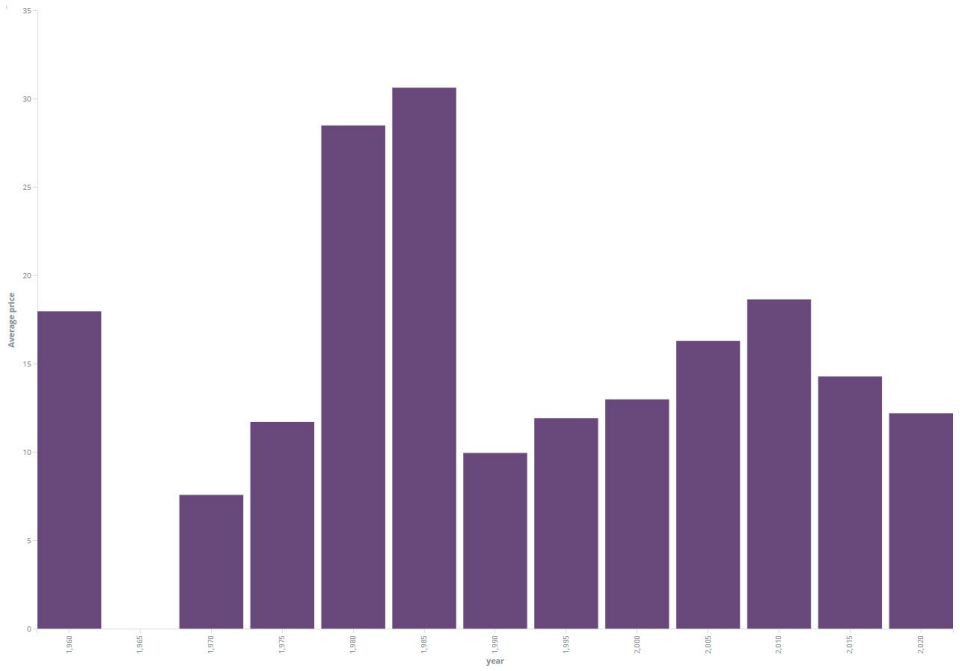
Query hľadá výrazy v atribútoch (name, author, description, comments.comment_text), pričom token nemusí sa presne zhodovať s hľadaným výrazom. Dĺžka zhodného prefixu je nastavená na 2.

- **Regexp query**

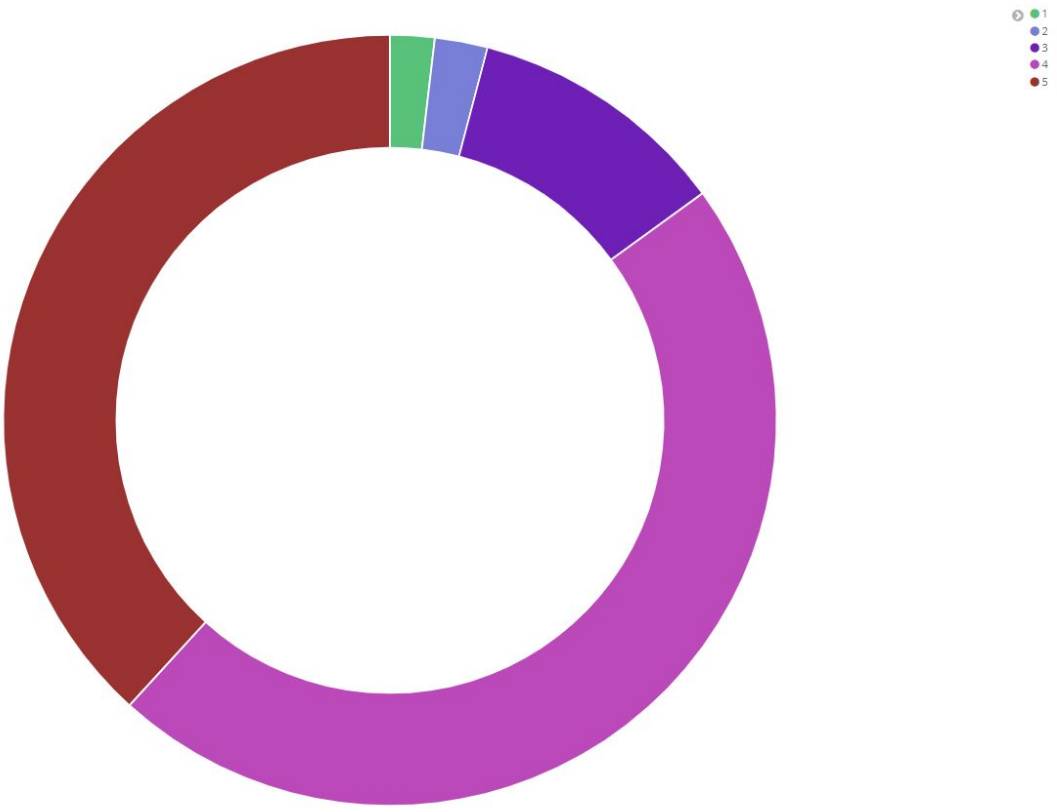
Query vyhľadá knihy z obdobia 1995- 2015, pričom nemôže mať kategóriu beletria a v názve knihy musí mať slovo začínajúce sa na bri.

6. Kibana Vyzualizácie

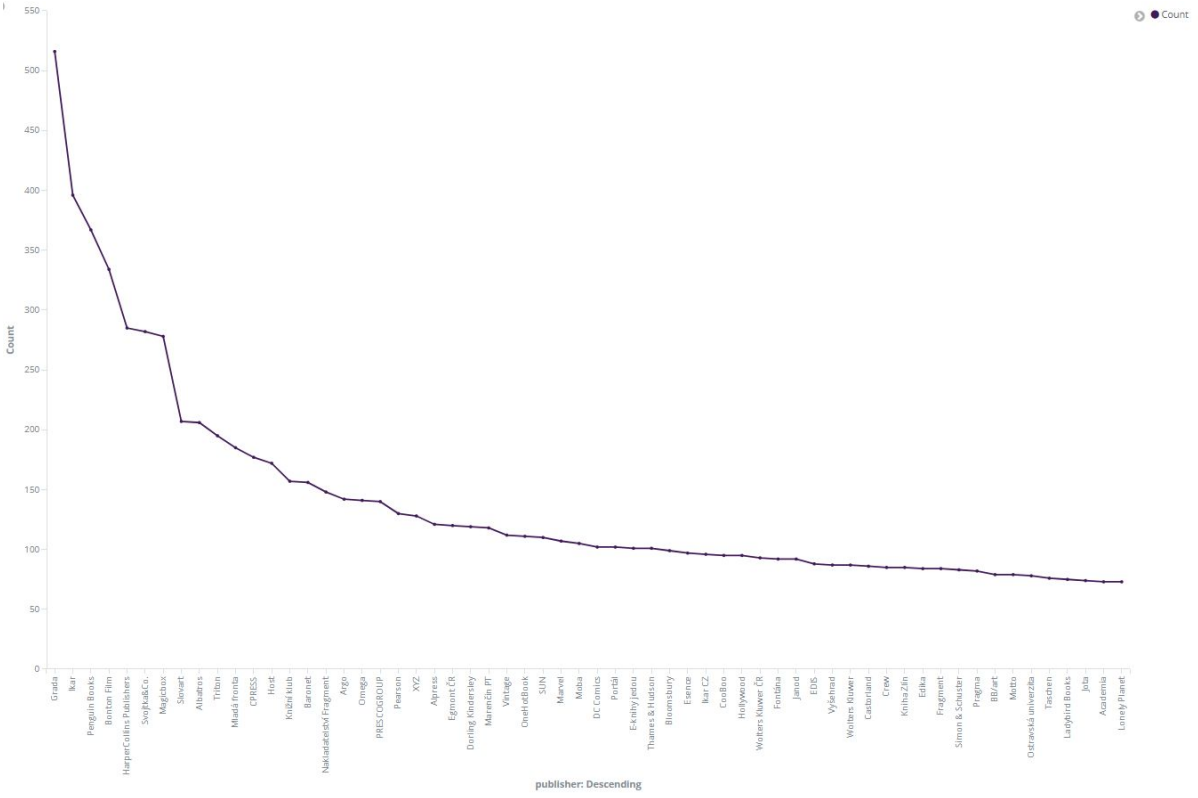
- Cenový priemer kníh



- Hodnotenie kníh



- Graf znázorňujúci počet kníh vydaných vydavateľstvami



7. Príloha: Zdrojový kód

- Vytvorenie Indexu

```
PUT martinus_book_index
{
  "settings": {
    "number_of_shards": 1,
    "analysis": {
      "filter": {
        "lemmagen_lexicon_sk": {
          "type": "lemmagen",
          "lexicon": "sk"
        },
        "lemmagen_lexicon_with_ext": {
          "type": "lemmagen",
          "lexicon": "sk.lem"
        },
        "lemmagen_lexicon_path": {
          "type": "lemmagen",
          "lexicon_path": "lemmagen/sk.lem"
        },
        "synonym_filter": {
          "type": "synonym",
          "synonyms_path": "synonyms/sk_SK.txt"
        },
        "stopwords_SK": {
          "type": "stop",
          "stopwords_path": "stop-words/stop-words-slovak.txt",
          "ignore_care": true
        }
      },
      "analyzer": {
        "slovencina": {
          "type": "custom",
          "tokenizer": "standard",
          "filter": [
            "lowercase",
            "lemmagen_lexicon_sk",
            "lemmagen_lexicon_with_ext",
            "lemmagen_lexicon_path",
```

```
            "synonym_filter",
            "stopwords_SK",
            "asciifolding"
          ]
        }
      }
    }
  }
}
```

- Mapovanie Indexu

```
PUT martinus_book_index/_mapping/_doc
{
  "properties": {
    "cover": {
      "type": "text",
      "analyzer": "slovencina"
    },
    "category": {
      "type": "text",
      "fields": {
        "en": {
          "type": "text",
          "analyzer": "english"
        },
        "cz": {
          "type": "text",
          "analyzer": "czech"
        },
        "general": {
          "type": "text",
          "analyzer": "slovencina"
        }
      }
    },
    "pages": {
      "type": "integer"
    },
    "language": {
      "type": "keyword"
```

```
},
"publisher": {
  "type": "keyword"
},
"year": {
  "type": "integer"
},
"size": {
  "type": "keyword"
},
"weight": {
  "type": "keyword"
},
"catalogue_number": {
  "type": "integer"
},
"isbn": {
  "type": "keyword"
},
"name": {
  "type": "text",
  "analyzer": "slovencina"
},
"author": {
  "type": "text",
  "fields": {
    "en": {
      "type": "text",
      "analyzer": "english"
    },
    "cz": {
      "type": "text",
      "analyzer": "czech"
    },
    "general": {
      "type": "text",
      "analyzer": "slovencina"
    }
  }
},
"price": {
```

```
      "type": "float"
    },
    "description": {
      "type": "text",
      "fields": {
        "en": {
          "type": "text",
          "analyzer": "english"
        },
        "cz": {
          "type": "text",
          "analyzer": "czech"
        },
        "general": {
          "type": "text",
          "analyzer": "slovencina"
        }
      }
    },
    "rating": {
      "type": "float"
    },
    "review": {
      "type": "text",
      "analyzer": "slovencina"
    },
    "comments": {
      "type": "nested"
    }
  }
}
```

- **Cenový priemer knihy**

```
GET /martinus_book_index/_search
{
  "size": 0,
  "aggs": {
    "avg_price": {
      "avg": {
        "field": "price"
```

```
      }
    }
  }
}
GET martinus_book_index/_search
{
  "query": {
    "bool": {
      "must": {
        "range": {
          "price": {
            "gt": 14.557515533222107
          }
        }
      },
    "filter": {
        "match": {
          "category": "dejiny"
        }
      }
    }
  }
}
```

- **Knihy, ktoré majú komentár s hodnotením 5 hviezdičiek**

```
GET martinus_book_index/_search
{
  "query": {
    "nested": {
      "path": "comments",
      "score_mode": "avg",
      "query": {
        "bool": {
          "must": [
            {
              "match": {
                "comments.user_rating": 5
              }
            }
          ]
```

```
        }
      }
    }
  }
}
```

- **Viac atribútová query s boostingom**

```
GET martinus_book_index/_search
{
  "query": {
    "bool": {
      "must": [
        {
          "multi_match": {
            "query": "klobasa Jozko ukulele gitara motorka hudba",
            "type": "most_fields",
            "fields": [
              "name^10",
              "author^10",
              "description",
              "comments.comment_text"
            ]
          }
        }
      ],
      "should": [
        {
          "match": {
            "category": "Cudzojazyčná literatúra Anglická"
          }
        },
        {
          "match": {
            "category": "Ostatne"
          }
        }
      ],
      "minimum_should_match": 1
    }
  }
```

```
}
```

- **“More like this” query**

```
GET martinus_book_index/_search
{
  "size": 20,
  "query": {
    "more_like_this": {
      "fields": [
        "name",
        "description",
        "category",
        "author",
        "language",
        "cover",
        "publisher"
      ],
      "like": [
        {
          "_index": "martinus_book_index",
          "_type": "_doc",
          "_id": "286225"
        }
      ],
      "min_term_freq": 1,
      "max_query_terms": 500
    }
  }
}
```

- **Fuzzy query**

```
GET martinus_book_index/_search
{
  "query": {
    "function_score": {
      "query": {
        "multi_match": {
          "query": "huba stonozka ryba zalud oko",
          "type": "most_fields",
          "fields": [
```

```
            "name^10",
            "author^10",
            "description",
            "comments.comment_text"
          ],
          "fuzziness": "AUTO",
          "prefix_length": 2
        }
      }
    }
  }
}
```

- **Regexp query**

```
GET martinus_book_index/_search
{
  "query": {
    "bool": {
      "must": [
        {
          "range": {
            "year": {
              "gte": "1995",
              "lte": "2015"
            }
          }
        },
        {
          "regexp": {
            "name": "bri.*"
          }
        }
      ],
      "must_not": [
        {
          "term": {
            "category": {
              "value": "beletria"
            }
          }
        }
```

```
        ]
      }
    }
  }
```