Slovenská technická univerzita v Bratislave
Fakulta informatiky a informačných technológií
Ilkovičova 2, 842 16 Bratislava 4

Matej Puk

IMDB vyhľadávač

Vyhľadávanie informácií

Študijný program: Softvérové inžinierstvo

Študijný odbor: 9.2.5. Softvérové inžinierstvo

Ročník: 2.

Cvičenie: Ut 11^{00}, miestnosť -1.65

Ak. rok: 2017/2018

Obsah

Zadanie

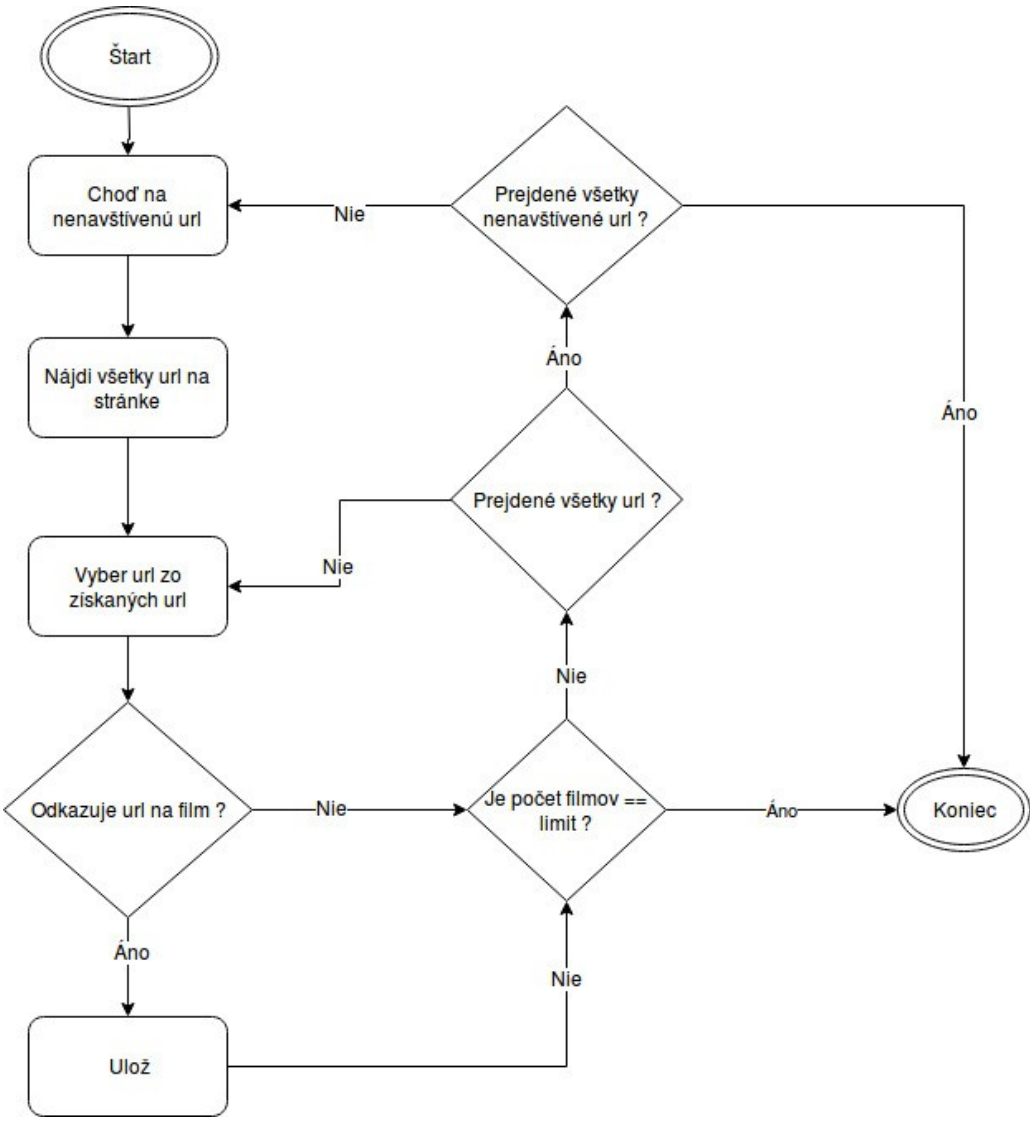
Cieľom zadania bolo vytvorenie crawlera, parsera pre získanie dát z vybratého internetového zdroja. Ako zdroj dát sme vybrali internetovú filmovú databázu **IMDB**. Následne po získaní dát sme vytvorili webovú aplikáciu, ktorá ponúka vyhľadávanie v získaných dátach na základe rôznych kritérií ako sú napríklad žáner filmu alebo jeho hodnotenie. Nakoniec sme vytvorili niekoľko vizualizácií získaných dát v nástroji **KIBANA**.

Celý projekt je implementovaný v jazyku **python**. Závisloti aplikácie sú spravované pomocou nástrojov **pip** a **bower**. Pre prepojenie python skriptov a webovej aplikácie s technológiou **Elasticsearch** bol použitý python modul **elasticsearch**.

Crawler

Prvou súčasťou zadania je **crawler**. Jeho implementácia sa nachádza v súbore **crawler.py**. Crawler je implementovaný pomocou triedy **Crawler**. Pri inicializácii je tejto triede zadaný štartovacia internetová stránka, z ktorej sa začína vyhľadávanie. V našom prípade je to stránka **http://www.imdb.com/genre/**. Následne je proces crawlovania spustený metódou **crawl()**, ktorej je zadaný argument **limit**, ktorý reprezentuje počet filmov, ktoré má crawler nájsť.

Na nasledujúcom diagrame je zobrazený algoritmus metódy **crawl()**.



Obrázok 1: Algoritmus metódy crawl()

Výstupom metódy **crawl()** je set jedinečných url odkazujúcich na filmy na stránke **IMDB**.

Parser

Druhou časťou zadania je vytvorenie parsera. Implementácia parsera sa nachádza v súbore parser.py. Parser je implementovaný triedou Parser. Pri inicializácii parsera je mu ako argument zadaný zoznam filmov, ktorý je výstupom crawlera. Následne je zoznam týchto filmov prejdený metódou parse(), ktorá prejde každý film v zozname uloží jeho html kód v podobe súboru na disk aby ho pri ďalšom spracovaní nebolo nutné znova sťahovať. Následne parser prejde celý html kód nachádzajúci sa v súbore a vyextrahuje z neho nasledovné dáta:

- hodnotenie filmu,
- počet ludí, ktorý hodnotili film,
- názov filmu,
- rok vydania,
- hodnotenie obsahu,
- dĺžka,
- popis,
- žánre,
- režisér,
- dej filmu,
- tvorcovia,
- herci,
- hviezdy,
- ocenenia,
- kľúčové slová,
- krajina,
- jazyk,

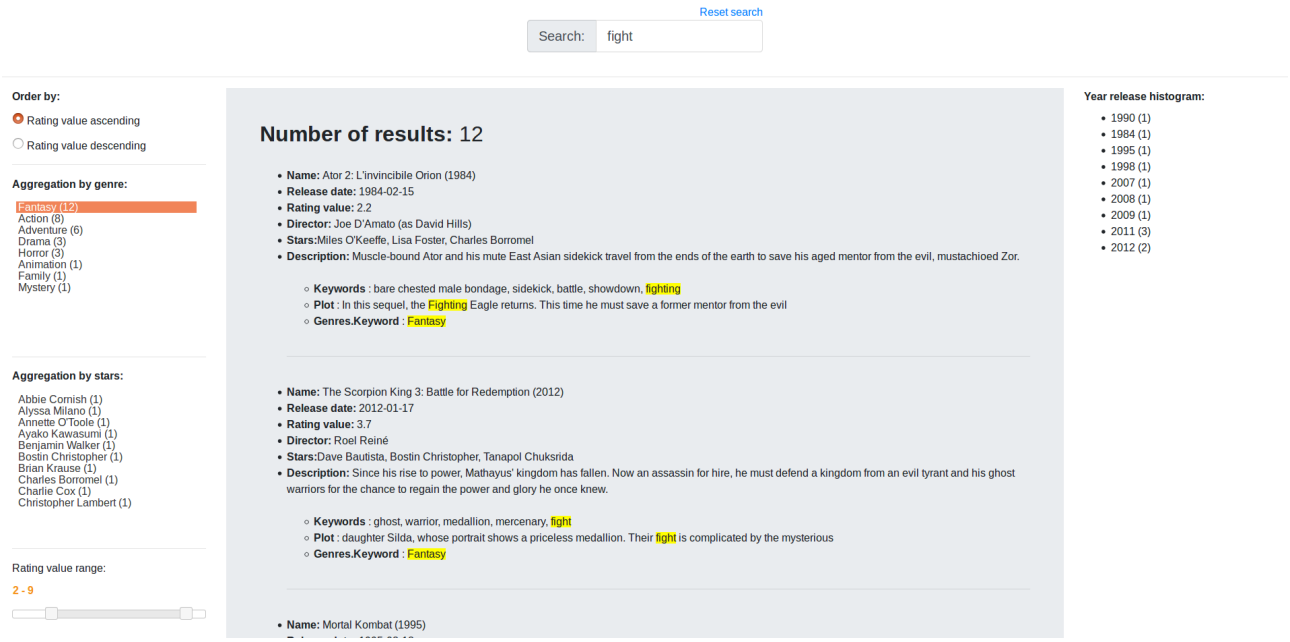
- dátum vydania.

Po vyextrahovaní dát je film uložený do databázy. Pre vyextrahovanie dát bol použití modul **BeautifulSoup**.

Webová aplikácia

Pre implementáciu webovej aplikácie bol použitý framework **Flask**. Aplikácia je implementovaná v súbore app.py. Šablóny pre vykreslenie aplikácie sa nachádzajú v adresári **templates**. Aplikácia vykresľuje hlavnú šablónu **main.html** a následne do nej vykresľuje šablónu s výsledkami vyhľadávania **hits.html**. Aplikácia sa dopytuje do databázy cez metódu **search()**. V tejto metóde sa nachádza celé telo vyhľadávacieho dopytu a aj telo dopytu pre sugescie v prípade, vyhľadávací dopyt nenájde žiadne položky.

Na nasledujúcom obrázku je zobrazený celkový vzhľad aplikácie.

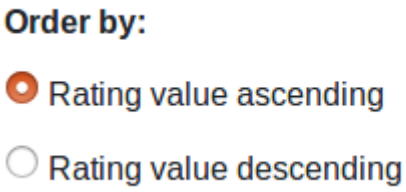


Obrázok 2: Webová aplikácie pre vyhľadávanie filmov

Na vrchu aplikácie v strede sa nachádza pole pre zadanie vyhľadávanej frázy. Zároveň je nad týmto poľom pridané tlačidlo pre celkové vynulovanie vyhľadávania.

Zoradenie podľa hodnotenia

Používateľ má možnosť zoradiť výsledky vyhľadávania podľa hodnotenia filmu a to vzostupne a zostupne.



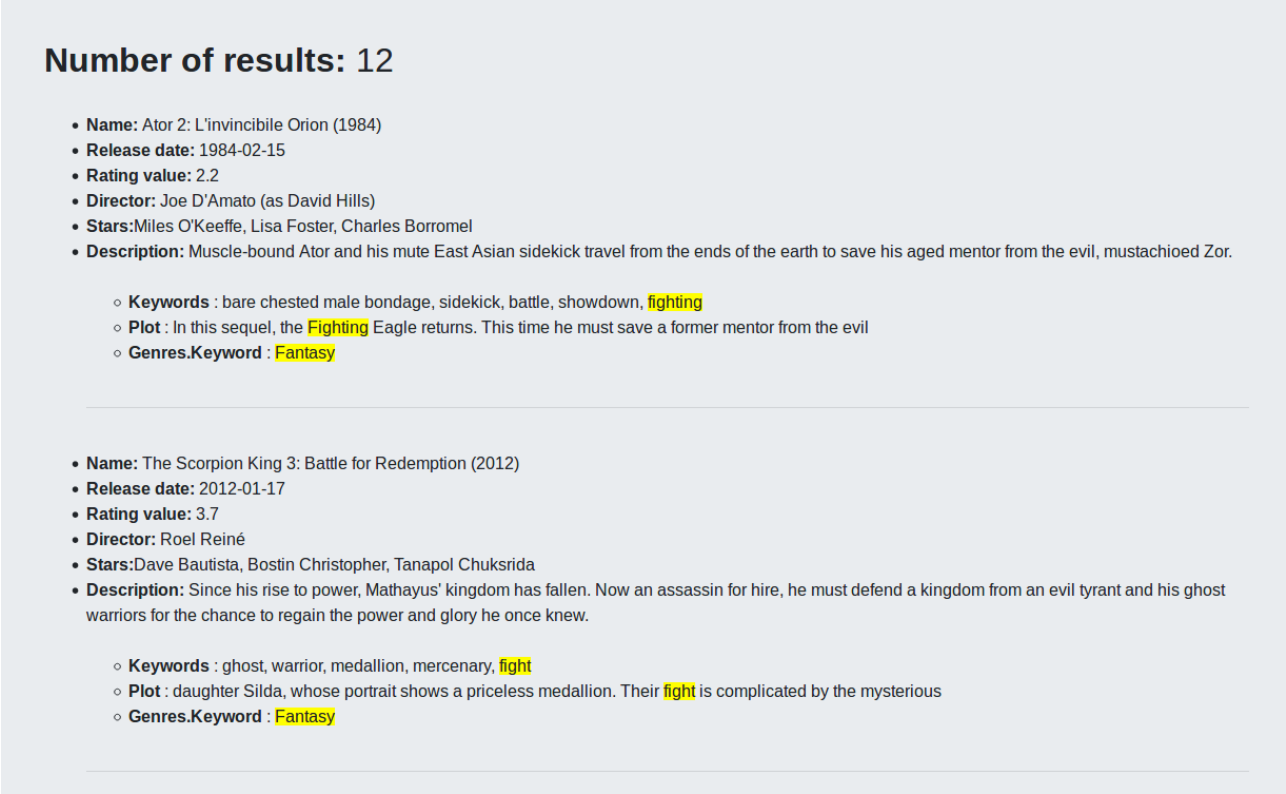
Obrázok 3: Zoradenie výsledkov podľa hodnotenia filmu

Implementácia v dopyte:

```
"sort": [
   {
      "ratingValue": {
         "order": o
      }
   },
   "_score"
]
```

Výsledky vyhľadávania

Ak dopyt nájde výsledky, tie sú zobrazené používateľovi uprostred stránky. Záznamy sú oddelené čiarou. Pre každý záznam je zobrazené meno, dátum vydania, hodnotenie, režisér, hviezdy a opis. Následne sú zobrazené všetky atribúty, v ktorých nastala zhoda a tá je zvýraznená žltou farbou.



Obrázok 4: Výsledky vyhľadávania

Filtrovanie výsledkov na základe agregácií žánrov a hviezd

Používateľ má možnosť si nájdené výsledky ďalej vyfiltrovať podľa žánrov alebo hviezd, ktoré vo filme vystupujú. Naviac je pri každom výsledky agregácie zobrazená početnosť v zátvorke.

Obrázok 5: Agregácie podľa hviezd a žánrov

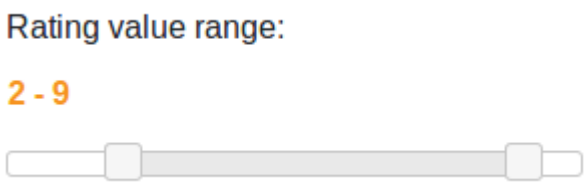
Implementácia v dopyte:

```
"aggs": {
   "by_genre": {
      "terms": {
         "field": "genres.keyword",
```

```
            "size": 10
        }
    },
     "by_stars": {
        "terms": {
            "field": "stars.keyword",
            "size": 10
        }
    }
}
```

Vyhľadávanie od-do podľa hodnotenia

Používateľ má možnosť si výsledky vyfiltrovať podľa ich hodnotenia zadaním hodnoty od-do na **jquery-ui** prvku **slider**.



Obrázok 6: Filter od-do podľa hodnotenia

Implementácia v dopyte:

```
"range": {
    "ratingValue": {
        "gte": int(rf),
        "lte": int(rt)
    }
}
```

Histogram rokov vydania

Na pravej strane v aplikácii má používateľ možnosť vidieť histogram ukazujúci početnosť filmov vydaných pre jednotlivé roky.

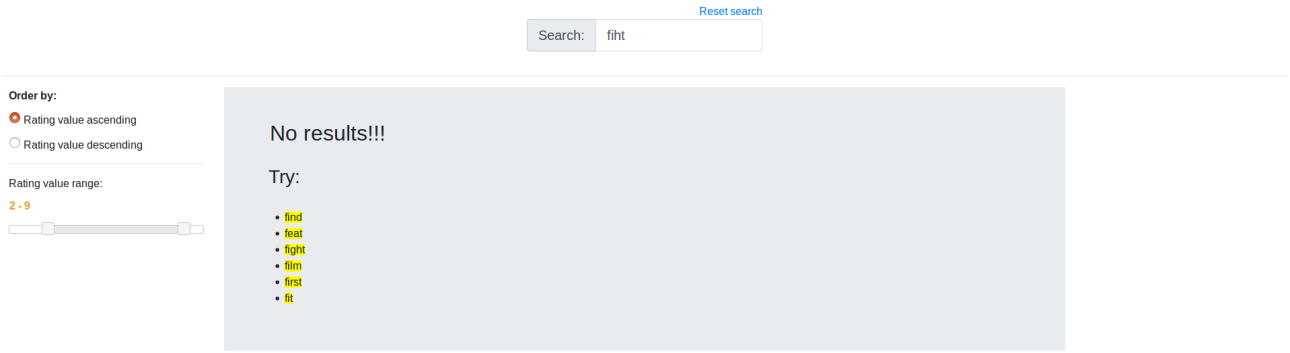
Obrázok 7: Histogram vydaných filmov vzhľadom na rok

Implementácia v dopyte:

```
"by_releaseYears": {
   "date_histogram": {
      "field": "releaseDate",
      "interval": "year"
   }
}
```

Návrhy vyhľadávania

Ak používateľ zadá frázu, ktorá vráti nulový počet výsledkov, aplikácia zavolá dopyt, ktorý navrhne používateľovi akú frázu podobnú tej, ktorú zadal má zadať aby získal výsledky. Na tútoo funkcionalitu bol využitý **phrase suggester**, ktorý bol mapovaním nastavený aby vyhľadával návrhy v atribútoch názov, popis a dej filmu.



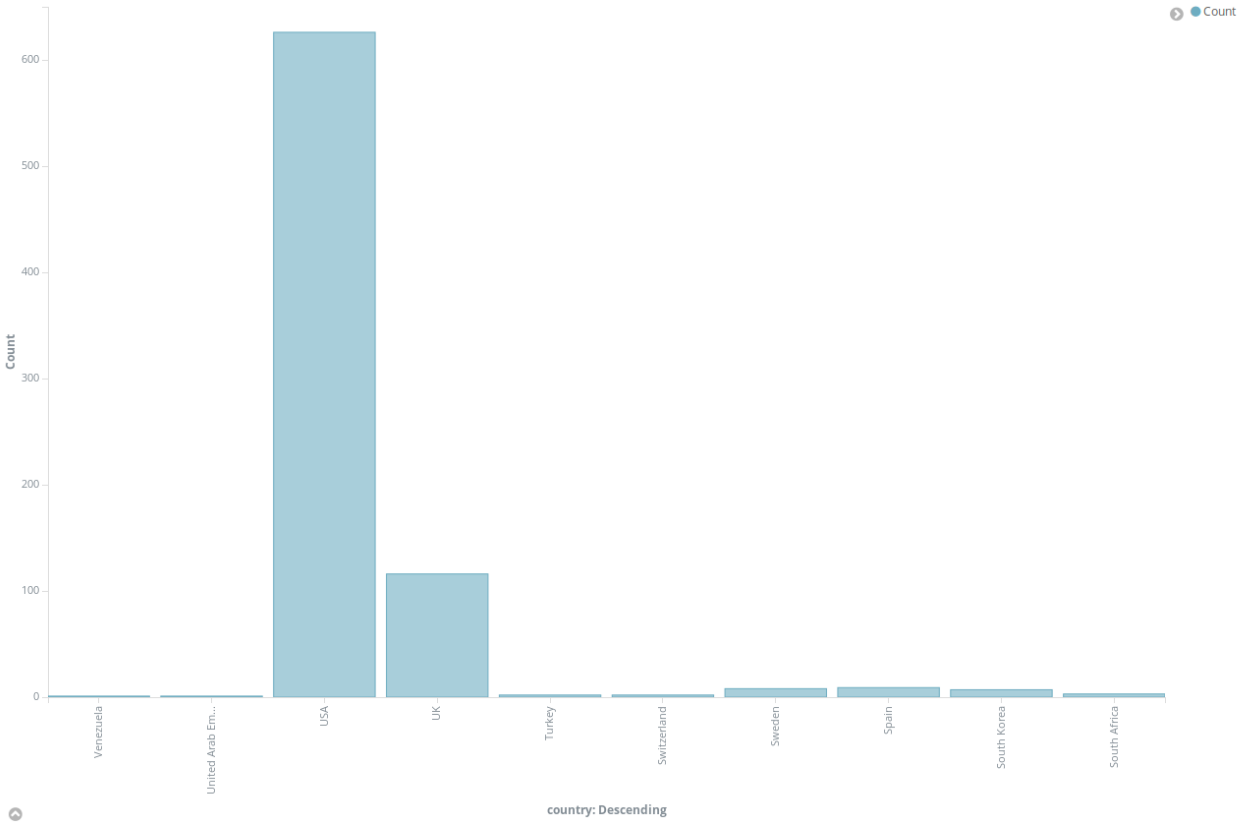
Obrázok 8: Návrhy pri nulovom počte výsledkov

Vizualizácie

V nasledujúcej kapitole sú zobrazené tri vizualizácie uskutočnené nad vyextrahovanou databázou.

Vizualizácia 1

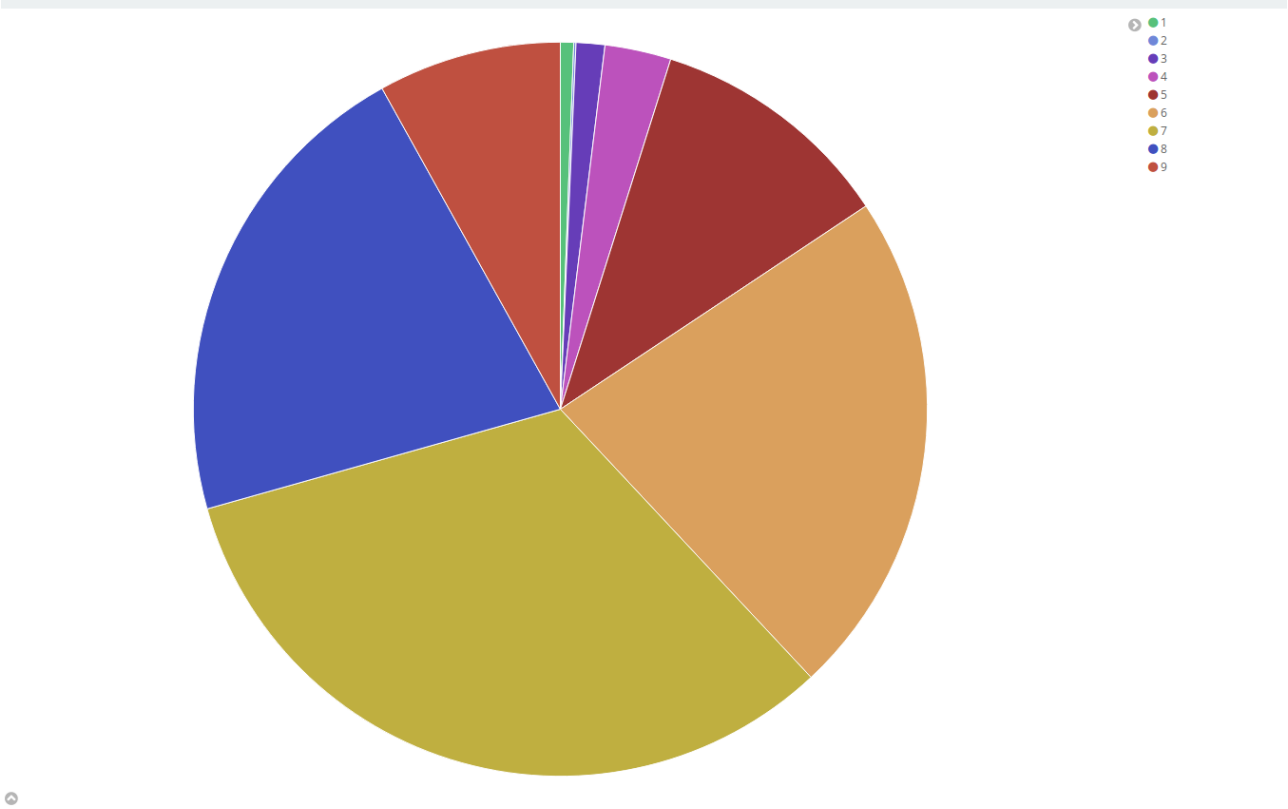
Nasledujúca vizualizácia zobrazuje počet filmov vzhľadom na krajinu za obdobie od roku 2012 do roku 2017.



Obrázok 9: Vizualizácia 1

Vizualizácia 2

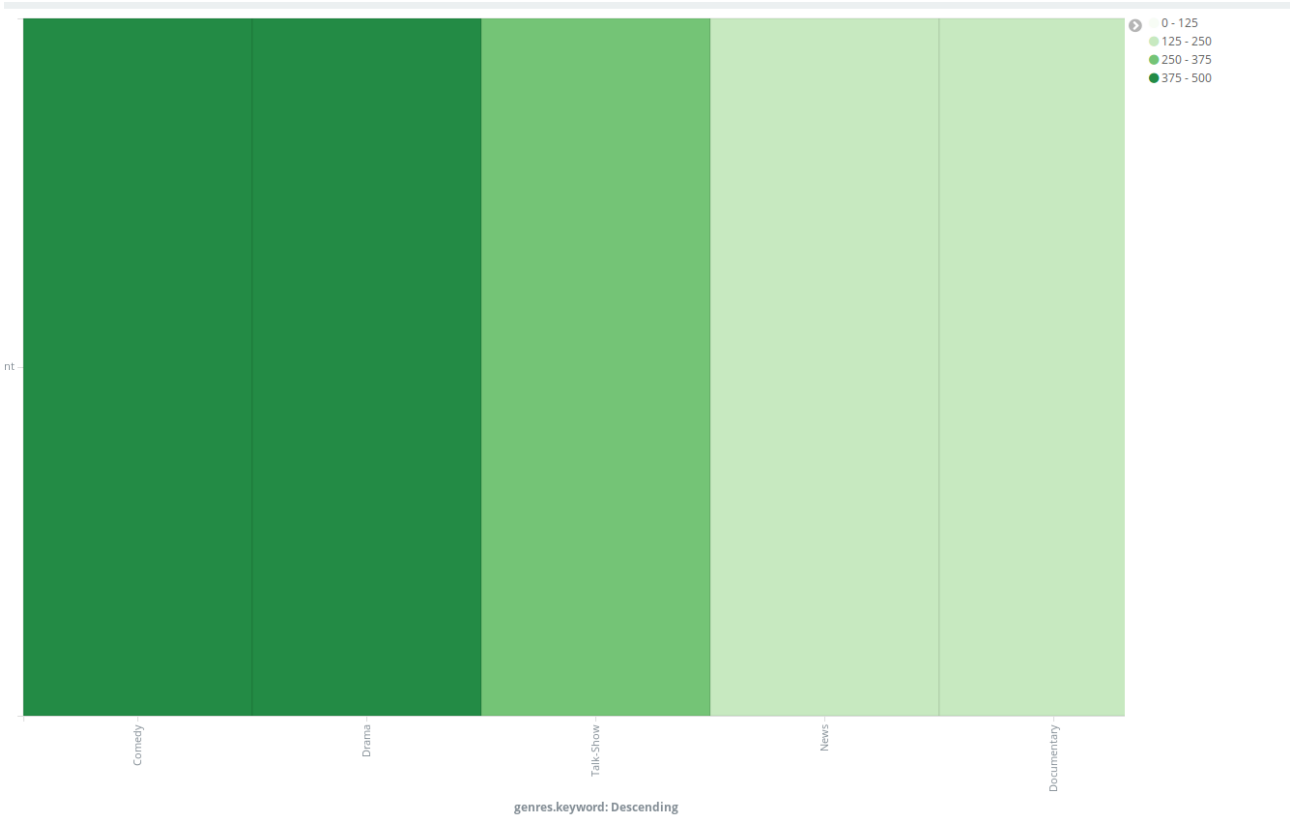
Nasledujúca vizualizácia zobrazuje počet filmov vzhľadom na ich hodnotenie od 1 do 10 po intervaly 1 za obdobie od roku 2012 do roku 2017.



Obrázok 10: Vizualizácia 2

Vizualizácia 3

Nasledujúca vizualizácia zobrazuje početnosť filmov vzhľadom na ich žáner za obdobie od roku 2012 do roku 2017.



Obrázok 11: Vizualizácia 3

Mapovanie dokumentu movie v indexe imdb

V nasledujúcej ukážke kódu je zobrazené nastavenia mapovania pre dokument typu **movie** v indexe **imdb**.

```
PUT imdb
{
        "settings": {
        "index": {
                "number_of_shards": 1,
                "analysis": {
                "analyzer": {
                        "suggest": {
                                "type": "custom",
                                "tokenizer": "standard",
                                "filter": ["standard", "shingle"]
                        }
                },
                "filter": {
                        "shingle": {
                                "type": "shingle",
                                "min_shingle_size": 2,
                                "max_shingle_size": 5
                        }
                }
                }
        }
        },
        "mappings": {
        "movie": {
                "properties": {
                     "name": {
                        "type": "text",
                        "analyzer" : "english",
              "fields": {
                 "suggest": {
                    "type": "text",
                    "analyzer": "suggest"
                 }
              }
                     },
                "plot": {
                        "type": "text",
                        "analyzer" : "english",
```

```
                "fields": {
                    "suggest": {
                        "type": "text",
                        "analyzer": "suggest"
                    }
                }
            },
            "genres": {
                "type": "text",
                "analyzer" : "english",
            "fields": {
              "keyword": {
                "type": "keyword",
                "ignore_above": 256
              }
            }
            },
            "director": {
                "type": "text",
                "analyzer" : "english"
            },
            "keywords": {
                "type": "text",
                "analyzer" : "english"
            },
            "awards": {
                "type": "text",
                "analyzer" : "english"
            },
            "stars": {
                "type": "text",
                "analyzer" : "english",
            "fields": {
              "keyword": {
                "type": "keyword",
                "ignore_above": 256
              }
            }
            },
            "duration": {
                "type": "text",
                "analyzer" : "english"
            },
            "actors": {
                "type": "text",
                "analyzer" : "english"
            },
            "creators": {
                "type": "text",
                "analyzer" : "english"
```

```
                },
                "description": {
                        "type": "text",
                        "analyzer" : "english",
              "fields": {
                "suggest": {
                   "type": "text",
                   "analyzer": "suggest"
                }
              }
                },

                "ratingValue": {
                        "type": "float"
                },
                "ratingCount": {
                        "type": "integer"
                },

                "language": {
                        "type": "keyword"
                },
                "country": {
                        "type": "keyword"
                },

                "releaseDate": {
                        "type": "date"
                }
                }
        }
        }
}
```