Fakulta informatiky a informačných technológií
Slovenská technická univerzita

Fast in-memory search

Zadanie 1

Meno: Bc. Ľudovít Popelka
Študijný program: Inteligentné softvérové systémy
Ročník: 1
Predmet: Vyhľadávanie informácií
Krúžok: štvrtok 11:00
Cvičiaci: Ing. Peter Gašpar
Ak. rok: 2018/2019

HLAVNÁ MYŠLIENKA

Poskytnúť rýchle, zmysluplné vyhľadávanie nad dátami (10.000+ položiek, niekoľko atribútov) vybranej webstránky vo forme niekoľkých vyhľadávacích scenárov. Výsledky doplnkovo vizualizovať.

NÁVRH RIEŠENIA

1. Crawlovanie webu

Použitá knižnica scrapy, kde sa vlastné tzv. vyhľadávacie pavúky použijú na listovanie zoznamu produktov na stránke, preklik na samotný detail produktu a zálohovanie html kódu stránok. Nasleduje skript na extrakciu atribútov produktu z html do formátu špecifikovaného pre ElasticSearch Bulk API. curl requestom sa dokumenty vložia do ElasticSearch.

2. Vyhľadávanie a vizualizácia

Demonštrované budú v konzole v Kibana. Taktiež bude založený Kibana dashboard a navrhnuté grafické vizualizácie.

DÁTA

Zdrojom dát je internetový obchod ebay. Jedná sa o ženské topy. Položky obsahujú atribúty:

- price

Cena produktu.

- sold

Počet predaných kusov.

- last_updated

Dátum poslednej úpravy.

- seller.name

Nickname predávajúceho.

- seller.feedback

Sptätná väzba na predávajúceho v percentách.

- condition

Stav obnosenia výrobku.

- material

Materiál, z ktorého je výrobok vyrobený.

- sleeve_style

Štýl rukávov.

- country

Krajina/región výroby.

- title

Názov v internetovom obchode.

- description

Popis výrobku.

IMPLEMENTÁCIA

Súčasťou odovzdania je aj zip súbor s kompletným kódom, bez vygenerovaných dát.

/doc	poznámky k implementácii
/doc/bulk.txt	curl request na vloženie dát do Elasticsearch
/doc/product.txt	základná schéma indexu
/doc/productadvanced.txt	rozšírená schéma indexu (preferovaná)
/doc/queries.txt	vyhľadávacie scenáre
/vinf	implementácia (scrapy projekt)
/vinf/spiders	crawlovacie pavúky
/vinf/eshop_crawler	dáta (html, json) a kód pre extrakciu atribútov

SCHÉMA

Pre schému indexu v ElasticSearch viď */doc/productadvanced.txt*. Pre pochopenie voľby použitých dátových typov poslúži dokumentácia ElasticSearch.

Poznámky k vybraným poliam schémy:

- seller

Obsahuje aj *nested* atribúty.

- raw

Aplikované na atribúty, ktoré má zmysel uložiť aj ako dátový typ *text*, aj ako *keyword*.

- completion

Atribút, ktorý používa autocomplete.

- ngram

Atribút, ktorý má vlastný analyzátor.

- settings

Rozšírené nastavenia indexu, kde je špecifikovaný vlastný analyzátor.

VYHĽADÁVANIE

1. bavlnená sexy vesta, nie čínska, max. $5, predaných aspoň 10 ks (2 možnosti ako vyhľadávanie riešiť)

2. pre všetky produkty s cenou $3 a viac, nájdi najvyššiu priemernú cenu predávajúcich, ktorí ponúkajú aspoň 3 produkty

3. produkty aktualizované za posledný rok, červené alebo modré, nie žlté, nie fialové, dlhorukávne (v názve alebo popise, ale v názve s vyššou prioritou)

4. automatické dopĺňanie písania názvu/krajiny (rôzne vyhľadávania demonštrujú rôzne parametre autocomplete)

5. 3-až-5-gram tokenizátor názvu

6. počet produktov vo zvolených cenových kategóriách pre každý materiál

VIZUALIZÁCIE

1. mapa

teplotný gradient podľa počtu produktov v krajine

2. oblak tagov

štýly rukávov, kde veľkosť značí početnosť

3. histogram

medián predaných produktov pre každú cenovú kategóriu

Viď obr. v prílohách V*.

ZÁVER

Navrhnutý bol postup pre crawlovanie webu a manuálnu extrakciu ľubovoľných html atribútov na základe tried a id z CSS. Riešenie používa Elasticsearch, aj jeho pokročilejšie funkcie. Ukázaných bolo zopár zmyslupných vyhľadávaní nad dátami, aj vizualizácie v Kibane. Do budúcna je dôležité uvedomiť si, že kvalitná analýza predpokladá kvalitne predspracované dáta, čo je úloha zrejme ešte náročnejšia.

ZDROJE

elastic: Docs [online] 2018. Dostupné na internete: https://www.elastic.co/guide/en/elasticsearch/reference/current/index.html

ebay.com: ebay [online] 2018. Dostupné na internete: https://www.ebay.com

scrapy: Docs [online] 2018. Dostupné na internete: https://docs.scrapy.org/en/latest/

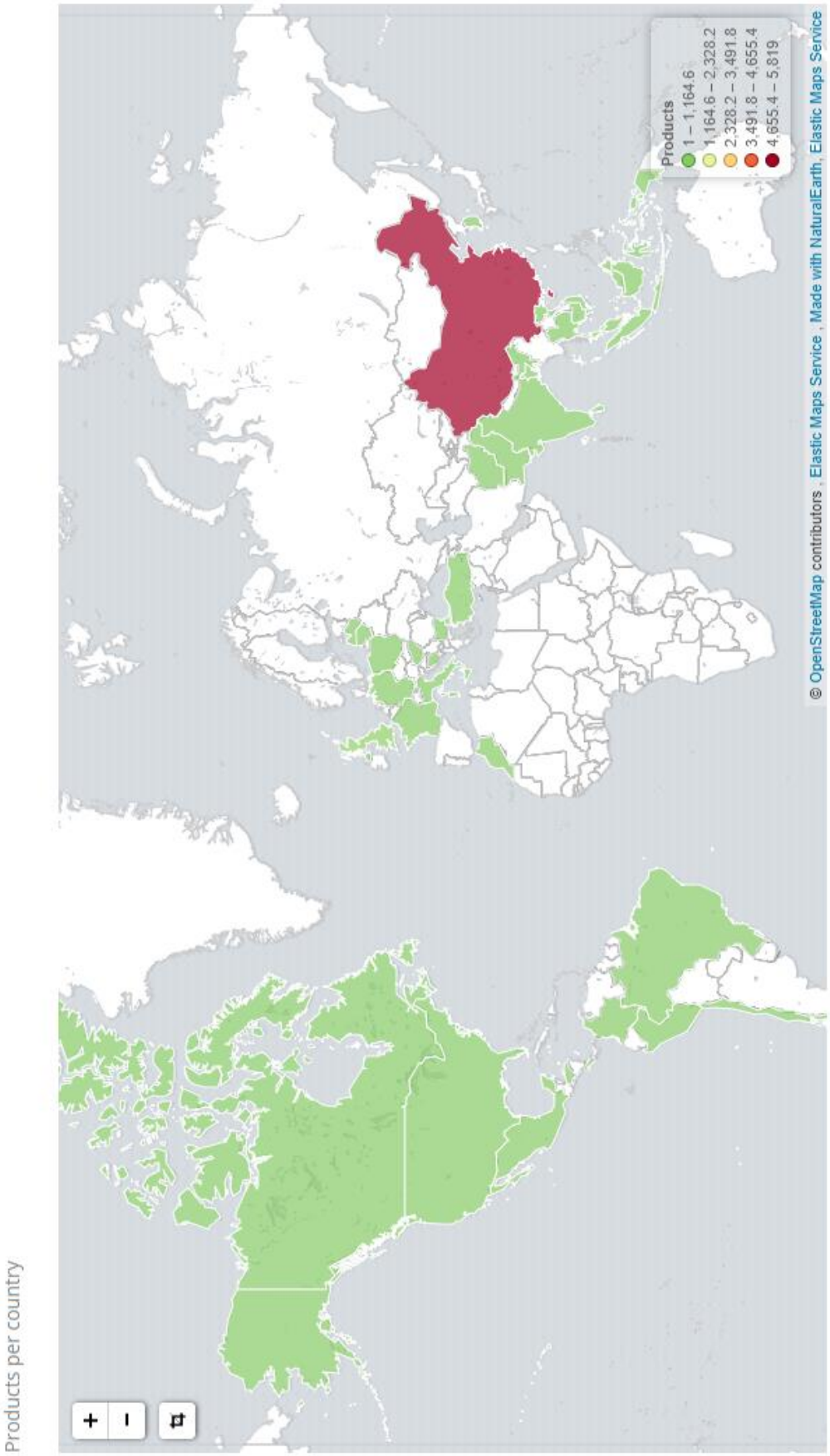
Dev Ticks: How to improve your full-text search in ElasticSearch with NGram Tokenizer [online] 2017. Dostupné na internete: https://devticks.com/how-to-improve-your-full-text-search-in-elasticsearch-with-ngram-tokenizer-e346f29f8ddb

HACKERNOON: Elasticsearch: Building AutoComplete functionality [online] 2017. Dostupné na internete: https://hackernoon.com/elasticsearch-building-autocomplete-functionality-494fcf81a7cf

KOMPAN, M: Information retrieval, Lectures [online] 2018. Dostupné na internete: http://www2.fiit.stuba.sk/~kompan/vi.html

SlidesLive od Jan Prokeš: Praktické představení možností ElasticSearch [online] 2018. Dostupné na internete: https://slideslive.com/38903659/prakticke-predstaveni-moznosti-elasticsearch

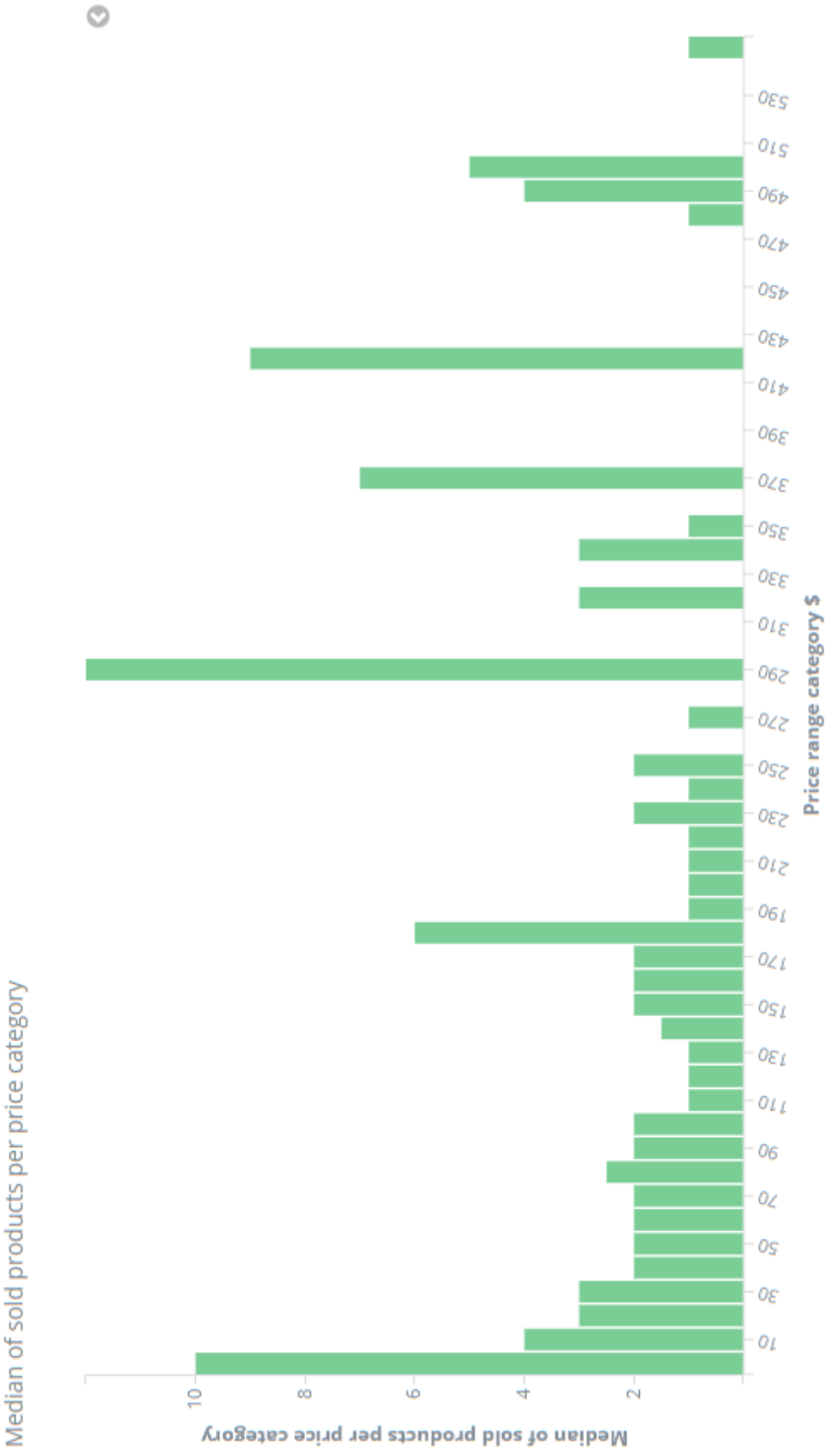
PRÍLOHA V1


Products per country
Products
1 – 1,164.6
1,164.6 – 2,328.2
2,328.2 – 3,491.8
3,491.8 – 4,655.4
4,655.4 – 5,819
© OpenStreetMap contributors , Elastic Maps Service , Made with NaturalEarth, Elastic Maps Service

PRÍLOHA V2


Sleeve style cloud
Cold Shoulder Sleeve
Off-Shoulder
Bell Sleeve
Strapless
Sleeveless
Regular
Long Sleeve
Half Sleeve
Short Sleeve
Cap Sleeve
3/4 Sleeve
Other
Batwing
Spaghetti Strap
Batwing, Dolman
Kimono Sleeve
Frequency - Sleeve style

PRÍLOHA V3


Median of sold products per price category
Median of sold products per price category
10
8
6
4
2
Price range category $
530
510
490
470
450
430
410
390
370
350
330
310
290
270
250
230
210
190
170
150
130
110
90
70
50
30
10