Ondrej Harnúšek
Po 8.00-9.50 1.40

Zadanie 1 - Vyhľadávanie informácií

1 Získanie datasetu

- Dataset bol získaný crawlovaním stránky https://www.kickstarter.com pomocou skriptu `crawler.py`.
- Crawler našiel na hlavnej stránke sekciu s kategóriami a následne prehľadával každú z nich.
- V rámci jednej kategórie prehľadával prvých 200 stránok, na ktorých našiel finálne adresy stránok projektov.
- Na stránke projektu našiel vybrané črty projektu a uložil ich do samostatného súboru (napríklad súbor `data_sample.json`).
- Celkovo sme získali: 200MB+ dát a 50k+ záznamov

2 Vytvorenie indexu

- Po úspešnom získaní všetkých dát sme vytvorili index a pridali doňho jednotlivé projekty pomocou skriptu `data_assembler.py`
- Index má polia rôznych typov a aj jedno vnorené pole
- Keďže dataset je v anglickom jazyku, bol použitý english analyzer
- Definícia indexu sa nachádza v `appendix.md`

3 ElasticSearch dopyty

- Okrem základných bool query sme použili: nested, boost_mode, multi_match, range a agregácie
- Všetky dopyty sa nachádzajú v `appendix.md`

3.1 Chcem nájsť všetky projekty z kategórie keramiky, ktoré nie sú z USA

- Použitie nested – kategória je vnorený objekt

3.2 Chcem nájsť prebiehajúce projekty. Ich skóre vypočítať podľa relatívneho prekročenia cieľa. Minimum je 1 násobné prekročenie

- Použitie boost_mode a skriptu
- Hodnota skóre je pomer hodnoty prekročenia cieľa ku cieľu projektu.
- Prekročenie musí byť nezáporné číslo (rovnako ako skóre samotné)
- Skóre = $\frac{(\text{vyzbieraných} - \text{cieľ}) + |\text{vyzbieraných} - \text{cieľ}|}{2}$ / $cieľ$

3.3 Chcem nájsť početnosť ukončených projektov z kategórie comics pre každý stav projektu.

- Použitie nested – kategória je vnorený objekt
- Použitie agregácie podľa stavov projektu

3.4 Chcem nájsť prebiehajúce projekty z V. Británie, ktoré majú v texte slová "queen Victoria". Nech sú zoradené podľa najsneskôr vytvorených.

- Použitie multi_match pri vyhľadávaní v rôznych textových poliach. Slová sa nemusia nachádzať za sebou, ale obe sa musia v texte vyskytovať.

3.5 Chcem nájsť priemernú vyzbieranú sumu prebiehajúcich projektov, ktoré budú končiť za 1 deň.

- Použitie range s relatívnym dátumom
- Vypočítanie priemeru pomocou agregácie

3.6 Chcem nájsť počet úspešných projektov z V. Británie a roka 2019. Počet nech je aspoň 10.

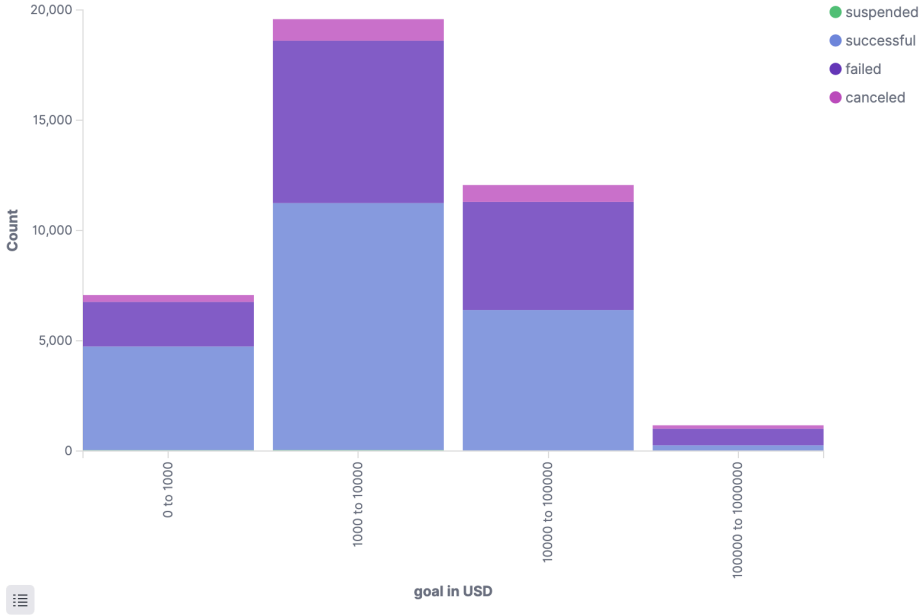
- Použitie range s absolútnym dátumom a filtrami
- Vypočítanie početnosti pomocou agregácie

4 Kibana vizualizácie

- Nad indexom sme vytvorili a Horizontal Bar, Line, Vertical Bar a Tag Cloud vizualizácie

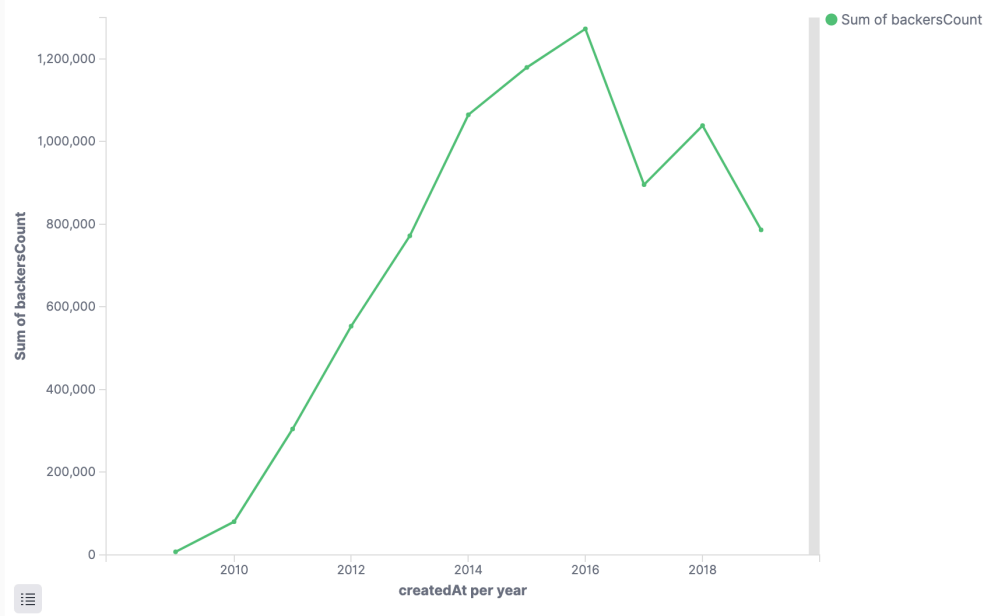
4.1 Početnosť stavov projektov podľa rôznej veľkosti cieľa

- Vidíme, že aspoň polovica projektov s cieľmi do 100,000$ býva úspešných



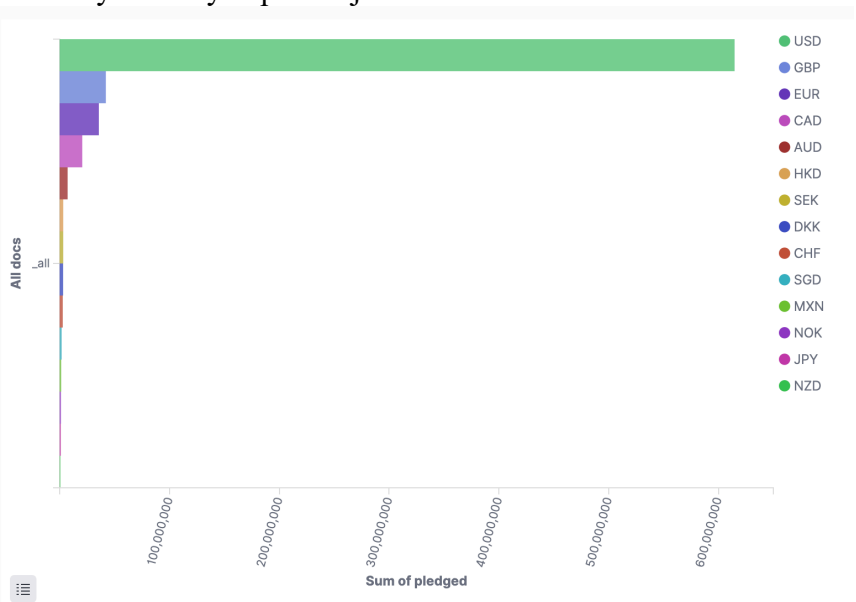
4.2 Počet podporovateľov za každý rok

- Vidíme, že v roku 2016 nastal prvý krát pokles počtu vytvorených projektov



4.3 Suma vyzbieraných peňazí (v $) pre každú menu

- Vidíme, že viac vyzbieraných peňazí je z libier ako eur.



4.4 Tagcloud najčastejších miest bežiacich projektov