Zadanie 4 – Elasticsearch

Odovzdanie do 11.12.2022 23:59 – dostanete za to 15 bodov. Otázky 1 až 8 sú za 7,5 boda, Otázka 9 je za 3 body a 10 za 4,5.

V rámci dokumentu napíšte každý dotaz, ukážte screenshot výsledku a opíšte v skratke čo sa stalo a prečo je to tak ako to je.

1. Rozbehajte si 3 inštancie Elasticsearch-u
2. Vytvorte index pre Tweety, ktorý bude mať “optimálny“ počet shardov a replík pre 3 nódy (aby tam bola distribúcia dotazov vo vyhľadávaní, aj distribúcia uložených dát)
3. Vytvorte mapping pre normalizované dáta z Postgresu (denormalizujte ich) – Každý Tweet teda musí obsahovať údaje rovnaké ako máte už uložené v PostgreSQL (všetky tabuľky). Dbajte na to, aby ste vytvorili polia v správnom dátovom type (polia ktoré má zmysel analyzovať analyzujte správne, tie ktoré nemá, aby neboli zbytočne analyzované (keyword analyzer)) tak aby index nebol zbytočne veľký, pozor na nested – treba ho použiť správne. Mapovanie musí byť striktné. Čo sa týka väzieb cez referencies – pre ne zaindexujte type vsťahu, id, autor (id, name, username), content a hashtags.
4. Pre index tweets vytvorte 3 vlastné analyzéry (v settings) nasledovne:
    a. Analyzér "englando". Tento analyzér bude obsahovať nasledovné:
        i. fitre: english_possessive_stemmer, lowercase, english_stop, english_stemmer,
        ii. char_filter: html_strip
        iii. tokenizer: štandardný - ukážku nájdete na stránke elastic.co pre anglický analyzér
    b. Analyzér custom_ngram:
        i. filtre: lowercase, asciifolding, filter_ngrams (definujte si ho sami na rozmedzie 1- 10)
        ii. char_filter: html_strip
        iii. tokenizer: štandardný
    c. Analyzér custom_shingles:
        i. filtre: lowercase, asciifolding, filter_shingles (definujte si ho sami a dajte token_separator: “”)
        ii. char_filter: html_strip
        iii. tokenizer: štandardný
    d. Do mapovania pridajte:
        i. každý anglický text (rátajme že každý tweet a description u autora je primárne v angličtine) nech je analyzovaný novým analyzérom "englando"
        ii. Priraďte analýzery
            1. a. author.name nech má aj mapovania pre custom_ngram, a custom_shingles
            2. b. author. screen_name nech má aj custom_ngram,
            3. c. author.description nech má aj custom_shingles. Toto platí aj pre mentions, ak tam tie záznamy máte.
        iii. Hashtagy indexujte ako lowercase

5. Vytvorte bulk import pre vaše normalizované Tweety.
6. Importujete dáta do Elasticsearchu prvych 5000 tweetov
7. Experimentujte s nódami, a zistite koľko nódov musí bežať (a ktoré) aby vám Elasticsearch vedel pridávať dokumenty, mazať dokumenty, prezerať dokumenty a vyhľadávať nad nimi? Dá sa nastaviť Elastic tak, aby mu stačil jeden nód? Čo je dôvodom toho že existuje nejaké kvórum?
8. Upravujte počet retweetov pre vami vybraný tweet pomocou vašeho jednoduchého scriptu (v rámci Elasticsearchu) a sledujte ako sa mení _seq_no a _primary_term pri tom ako zabíjate a spúšťate nódy.
9. Zrušte repliky a importujete všetky tweety
10. Vyhľadajte vo vašich tweetoch, kde použite function_score pre jednotlivé medzikroky nasledovne:
    a. Must:
        i. Vyhľadajte vo viacerých poliach naraz (konkrétne: author.description.shingles (pomocou shingle) – boost 10, content (cez analyzovaný anglický text) spojenie – boost 6 "put1n chr1stian fake jew", zapojte podporu pre preklepy, operátor je OR.
        ii. V poly referencies.content slovo “nazi”
        iii. Hashtag “ukraine”
    b. Filter:
        i. vyfiltrujte len tie, ktoré majú author.following_count > 100, tie ktoré majú author.followers_count > 100 a tie, ktoré majú nejakú linku
    c. Should:
        i. Ak sa v context_annotations.domain.name nachádza “Person” boostinte o 5
        ii. Ak sa v context_annotations.entity.name nachádza “Soros” boostnite o 10
        iii. Ak je vyhľadaný string “put1n chr1stian fake jew” aj fráza s tým ze sa môže stat jedna výmena slov boostnite o 5
    d. Agregácie:
        i. Vytvorte bucket pro-russia ktorý obsahuje hastagy používane Kremľom na propagandu: istandwithputin, racism, 1trillion, istandwithrussia, isupportrussia, blacklivesmatter, racism, racistukraine, africansinukraine, palestine, israel, freepalestine, istandwithpalestine, racisteu, putin
            1. Pre neho spravte týždňový histogram, kde pre každý týždeň zobrazte štatistiky