PDT - part 5

- repo link: github class
- autor: Kateřina Mušková

1.

Rozbehajte si 3 inštancie Elasticsearch-u

2.

Vytvorte index pre Tweety, ktorý bude mať "optimálny" počet shardov a replík pre 3 nody (aby tam bola distribúcia dotazov vo vyhľadávaní, aj distribúcia uložených dát)

```
{
  "settings": {
    "number_of_shards": 3,
    "number_of_replicas": 2,
```

Tři shardy rozdělí index na tři nody. Nebude zátěž jen na jeden z nich. Dvě repliky umožní, že i jen s jedním nodem by měla být zachována plná funkcionalita a přístup k datům.

3.

Vytvorte mapping pre normalizované dáta z Postgresu - Tweet musí obsahovať údaje rovnaké ako máte už uložené v PostgreSQL. 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. Mapovanie musí byť striktné.

Viz mapping.json

4.

Pre index tweets vytvorte 3 vlastné analyzéry (v settings) nasledovne:

1. Analyzér "englando".

Tento analyzér bude obsahovať nasledovné: 1.2. fitre: english_possessive_stemmer, lowercase, english_stop, english_stemmer, 1.3. char_filter: html_strip 1.4. tokenizer: štandardný - ukážku nájdete na stránke elastic.co pre anglický analyzér

2. Analyzér custom_ngram:

2.2. Filtre: lowercase, asciifolding, filter_ngrams (definujte si ho sami na rozmedzie 1- 10) 2.3. char_filter: html_strip 2.4. tokenizer: štandardný

3. Analyzér custom_shingles:

3.2. Filtre: lowercase, asciifolding, filter_shingles (definujte si ho sami a dajte token_separator: "") 3.3. char_filter: html_strip 3.4. tokenizer: štandardný

Viz analyser.json

Do mapovania pridajte:

1. každý anglický text (rátajme že každý tweet a description u autora je primárne v angličtine) nech je analyzovaný novým analyzérom "englando"
2. Priraďte analýzery

a. author.name nech má aj mapovania pre custom_ngram, a custom_shingles,
b. author. screen_name nech má aj custom_ngram,
c. author.description nech má aj custom_shingles. Toto platí aj pre mentions, ak tam tie záznamy máte.

3. Hashtagy indexujte ako lowercase

Viz mapping.json

5.

Vytvorte bulk import pre vaše normalizované Tweety.

Viz skript migrate.sh

6.

Importujete dáta do Elasticsearchu prvych 5000 tweetov

./migrate.sh -s 5000 -b 5000

7.

A 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?

Může běžet jen jeden, pokud si nastavím 2 repliky. Tím pádem budou data na všech nodech.

B Dá sa nastaviť Elastic tak, aby mu stačil jeden nód?

První node lze ještě ostranit normální cestou. Před odebráním dalšího je ale potřeba dopředu vyčlenit jeden node z hlasování. Hlasovat může každý node,

kterému není odepřeno se stát masterem. Hlasuje se například o výběru nového mastera, nebo stavu clusteru.

Vyčleněním jednoho nodu z hlasování přecházejí jeho práva na další. Vyčleněný node se pak může bezpečně smazat.

```
"_seq_no": 17,
"_primary_term": 2,
```

```
epoch      timestamp cluster          status node.total node.data shards pri relo init unassign pending_tasks max_task_wait_time active_shards_percent
1639144990 14:03:10  es-docker-cluster yellow          1         1      4   4    0    0        6             0                  -                 40.0%
```

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.

Výchozí stav:

```
"_seq_no": 25,
"_primary_term": 2,
```

Update se všemi nody:

```
"_seq_no": 26,
"_primary_term": 2
```

Druhý update se všemi nody:

```
"_seq_no": 27,
"_primary_term": 2
```

Změna po vypnutí node03, jako primary byl označen jiný shard

```
"_seq_no": 28,
"_primary_term": 3,
```

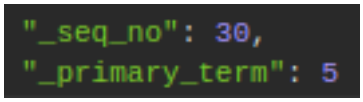
Po nastartování node es03 se nemění ani jedno z čísel. Zůstává nejnovější verze.

```
"_seq_no": 28,
"_primary_term": 3,
```

Změna při vypnutém node es02, kde byli informace primárně uložené, opět ovlivní obě čísla

```
"_seq_no": 29,
"_primary_term": 4
```

Pokud zkusíme es02 znovu nastartovat a opět provedeme změnu při vypnutém es02, změní se jen seq number, jelikož primární shard není na es02.



9.

Zrušte repliky a importujete všetky tweety

./migrate.sh -b 55000

10.

Vyhľadajte vo vašich tweetoch spojenie "gates s0ros vaccine micr0chip". V query použite function_score, kde jednotlivé medzikroky sú nasledovné: Query: 1. Must - vyhľadajte vo viacerých poliach (konkrétne: author.name (pomocou shingle), content (cez analyzovaný anglický text), author.description (pomocou shingles), author.screen_name (pomocou ngram)) spojenie "gates s0ros vaccine micr0chip", zapojte podporu pre preklepy, operátor je OR. 2.1 tieto polia vo vyhľadávaní boost-nite nasledovne - author.name * 6, content * 8, author.description * 6, author.screen_name * 10. 3. Filter - vyfiltrujte len tie, ktoré majú author.statuses_count > 1000 a tie, ktoré majú hashtag „qanon" 4. Should – boost-nite 10 krat tie, ktoré obsahujú v mentions.name (tento objekt je typu nested) cez ngram string "real". 5. Nastavte podmienené váhy cez functions nasledovne: 5.1. retweet_count, ktorý je väčší rovný ako 100 a menší rovný ako 500 na 6, 5.2. author.followers_count väčší ako 100 na 3 Zobrazte agregácie pre výsledky na konci. Vytvorte bucket hashtags podľa hashtagov a spočítajte hodnoty výskytov (na webe by to mohli byť facety).

```
"aggregations": {
  "hash_agg": {
    "doc_count_error_upper_bound": 0,
    "sum_other_doc_count": 0,
    "buckets": [
      {
        "key": "qanon",
        "doc_count": 744
      },
      {
        "key": "WWG1WGA",
        "doc_count": 154
      },
      {
        "key": "q",
        "doc_count": 113
      },
      {
        "key": "wwg1wga",
        "doc_count": 77
      },
      {
        "key": "WWG1WGAWORLDWIDE",
        "doc_count": 63
      },
      {
        "key": "QAnon2020",
        "doc_count": 53
      },
```


Viz query.json

11.

Konšpiračné teórie podľa Elasticu. Pracujte zo všetkými tweetami, ktoré máte. Následne pre všetky týždne zistite pomocou vnorených agregácii, koľko retweet_count sumárne majú tweety ktoré majú hashtagy z prvého zadania. Teda na základe hashtagov znova rozdeľte tweety do konšpiračných teórii ale pomocou agregácii.

```
  },
  "GLobal_Warming": {
    "doc_count": 1,
    "week_agg": {
      "buckets": [
        {
          "key_as_string": "2020-04-01T00:00:00.000Z",
          "key": 1585699200000,
          "doc_count": 1,
          "sum_retweet_count": {
            "value": 1.0
          }
        }
      ]
    }
  },
  "Illuminati": {
    "doc_count": 8,
    "week_agg": {
      "buckets": [
        {
          "key_as_string": "2020-01-01T00:00:00.000Z",
          "key": 1577836800000,
          "doc_count": 2,
          "sum_retweet_count": {
            "value": 1.0
          }
        },
        {
          "key_as_string": "2020-02-01T00:00:00.000Z",
          "key": 1580515200000,
          "doc_count": 1,
          "sum_retweet_count": {
            "value": 0.0
          }
        },
        {
          "key_as_string": "2020-03-01T00:00:00.000Z",
          "key": 1583020800000,
          "doc_count": 0,
          "sum_retweet_count": {
            "value": 0.0
          }
        },
        {
          "key_as_string": "2020-04-01T00:00:00.000Z",
          "key": 1585699200000,
          "doc_count": 5,
          "sum_retweet_count": {
            "value": 4.0
          }
        }
      ]
    }
  },
```

Viz query.json