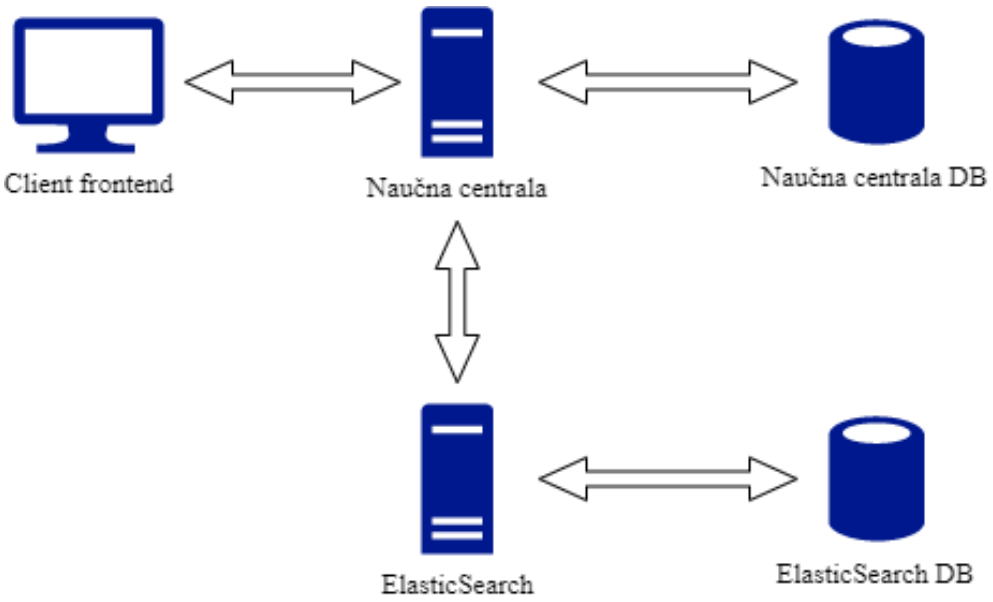
Upravljanje digitalnim dokumentima

Za projekat iz predmeta Upravljenje digitalnim dokumentima je potrebno napraviti aplikaciju koja će omogućiti pretraživanje radova u okviru Naučne centrale. Naučna centrala je sistem koji sadrži veliku količinu radova iz različitih naučnih oblasti. Kako bi korisniku bilo što lakše da pronađe odgovarajući rad, potrebno mu je omogućiti pretragu radova po različitim kriterijumima. Da bi se to postiglo, u okviru ovog projekta će se koristiti ElasticSearch platforma. Ova platforma omogućava pretraživanje kolekcije naučnih radova na osnovu postavljenih upita.

Arhitektura aplikacije

Aplikacija se sastoji iz narednih modula:

1. Client frontend – Angular aplikacija sa korisničkim interfejsom koji omogućava zadavanje različitih upita za pretragu
2. Naučna centrala – SpringBoot aplikacija u kojoj se obavlja celokupna logika i koja prosleđuje zahteve ka ElasticSearch-u
3. ElasticSearch – Platforma koja omogućava pretraživanje digitalnih radova koji se nalaze u njenoj bazi na osnovu postavljenog upita



Komunikacija između Client frontend-a i Naučne centrale se obavlja preko REST API-a.

Naučna centrala će komunicirati sa ElasticSearch-om putem REST API-a. Da bi ovo bilo moguće, potrebno je ubaciti odgovarajuće dependency-e unutar SpringBoot aplikacije za Naučnu centralu. Takođe je potrebno u konfiguracionom fajlu u okviru Naučne centrale napraviti klijenta koji omogućava slanje REST zahteva. Zahtevi će se slati na http://localhost:9200, obzirom da ElasticSearch sluša na tom portu ukoliko mu se prisupa putem REST-a.

Skladištenje podataka

U okviru baze podataka Naučne centrale će se čuvati sledeći podaci u vezi sa naučnim radom:

- Nalov rada
- Imena koautora, kao i njihove email i fizičke adrese
- Ključni pojmovi
- Apstrakt
- Naučna oblast rada
- PDF dokument

Digitalna biblioteka prihvaćenih radova će se skladištiti u okviru ElasticSearch-a, kako bi bilo omogućeno pretraživanje podataka. Da bi bila omogućena geoprostorna pretraga, neohodno je na ElasticSearch platformi čuvati *latitude* i *longitude* od svih recenzenata u sistemu.

SerbianAnalyzer

Da bi se obavilo pretprocesiranje tekstova na srpskom jeziku, najpre je instaliran plugin koji to omogućava (Serbian Analyzer). Analyzer koji će se koristiti u okviru ovog projekta je kompatibilan sa Java verzijom 13.0.1, Gradle verzijom 6.0 i ElasticSearch verzijom 7.4.0. i dostupan je na sledećem linku: https://github.com/markomartonosi/udd06/tree/plugin-update.

Prvi korak jeste preuzimanje projekta sa navedenog linka koji sadrži analyzer za srpski jezik. Kako bi se analyzer mogao ubaciti u ElasticSearch, neophodno je izbildovati odgovarajući fajl pomoću Gradle-a. Potrebno je otvoriti terminal u root folderu preuzetog projekta i izvršiti naredbu *gradlew clead build*. Na taj način je nastala arhivirana distirbucija u folderu *build/distributions* unutar root foldera sa nazivom *serbian-analyzer-1.0-SNAPSHOT.zip*. Sledeći korak je instaliranje dobijenog fajla kao plugin-a unutar ElasticSearch-a. To se ostvaruje pozicioniranjem unutar *bin* foldera u okviru root foldera u koji je preuzet ElasticSearch. Pomoću naredbe *elasticsearch-plugin install file:<absolute path of distribution archive>* se instalira plugin.

ElasticSearch

Kako bi se koristio ElasticSearch, prvo je potrebno pokrenuti server. To se postiže pozicioniranjem unutar *bin* foldera u okviru root foldera u koji je preuzet ElasticSearch i pokretanjem *elasticsearch.bat* fajla. Da bi ElasticSearch sa instaliranim plugin-om imao mogućnost analiziranja teksta na srpskom jeziku, potrebno je još putem PUT zahteva naglasiti da se, pored podrazumevanog analyzera za engleski jezik, koristi i analyzer za srpski jezik.

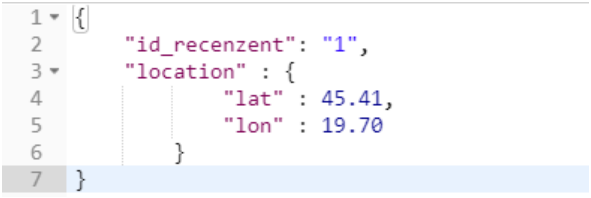
Na slici ispod je prikazano kako je moguće „obavestiti“ ElasticSearch da prilikom dodavanja novog dokumenta u index *casopis* koristi i analyzer za srpski i analyzer za engleski jezik.

Kao što je već rečeno, komunikacija sa ElasticSearch-om će se obavljati putem REST API-a. Takođe, koristiće se i ElasticSearchTemplate kao i ElasticSearchRepository koji omogućavaju indeksiranje i pretraživanje. U okviru ovog projekta će se koristiti ElasticSearch verzija 7.4.0.

Indexing unit

Prilikom dodavanja novog rada u izabrani časopis, indexing unit će imati ovakav oblik.

Obzirom da je za geoprostornu pretragu potrebno pronaći recenznte čija je adresa na određenoj udaljenosti od datih adresa, u ElasticSearch će biti potrebno dodavati i recenzente sistema. Za njihovo dodavanje će biti korišćen indexing unit sledećeg oblika:



More like this

U okviru ovog projekta potrebno je omogućti *More like this* pretragu. Ovo je omogućeno ElasticSearch platformom putem MLT query-a. Ovaj upit pronalazi dokumente koji su poput datog skupa dokumenata. Da bi to učinio, MLT bira skup reprezentativnih termina ovih ulaznih dokumenata, formira upit koristeći te termine, izvršava ga i vraća rezultate. Korisnik je taj koji kontroliše ulazne dokumente, kako se biraju termini i kako se upit formira.

Za projekat Naučne centrale potrebno je pronaći sve recenzente koji su recenzirali radove slične podnetom tekstu rada. To znači da će se u upitu za MLT u okviru parametra *fields* (spisak polja za preuzimanje i analiziranje teksta) nalaziti polje koje sadrži tekst naučnog rada. Time smo ograničili ElasticSearch da pretragu po sličnosti bazira samo na ovom polju. Jedini obavezan parametar *More like this* upita jeste parametar *like*. U ovom parametru korisnik može navesti tekst u slobodnoj formi i/ili jedan ili više dokumenata sa kojima je potrebno naći sličnost. U ovom projektu će parametar *like* imati vrednost teksta podnetog dokumenta.

```
{
    "query": {
        "more_like_this" : {
            "fields" : ["sadrzaj"],
            "like" : "tekst poddnetog dokumenta"
        }
    }
}
```

Geoprostorna pretraga

Za projekat Naučne centrale je takođe potrebno podržati geoprostornu pretragu. ElasticSearch ima podršku za takve upite. Ovakva pretraga u okviru ElasticSearch-a omogućava filtriranje dokumenata, tako da se kao rezultat dobiju samo oni dokumenti koji postoje na određenoj udaljenosti od zadate.

Za potrebe ovog projekta će se koristiti lokacija kao *geo_point* tačka. To znači da podatke o lokaciji možemo indeksirati u različitim formatima, konkretno u ovom slučaju će se koristiti *latitude* i *logitude*. Za svakog recenzenta u sistemu će se na ElasticSearch platformi čuvati podatak o njegovom ID-u, kao i *latitude* i *longitude* njegove adrese. Da bi se izvršila geoprostorna pretraga, iz baze podataka Naučne centrale će se dobaviti autor i svi koautori podnetog rada, odnosno njihovi parametri *latitude* i *longitude*. Potom će se vršiti geoprostorna pretraga na osnovu tih parametara, kako bismo dobili recenzente koji se nalaze na traženoj udaljenosti od autora i svih koautora.

```
{
    "query": {
        "bool": {
            "must": {
                "match_all": {}
            },
            "must_not": {
                "geo_distance": {
                    "distance": "100km",
                    "autor.lokacija": {
                        "lat": 43.320904,
                        "lon": 21.895760
                    }
                }
            }
        }
    }
}
```

Na slici iznad je prikazan primer geoprostorne pretrage, gde se kao rezultat vraćaju svi dokumenti čija je udaljenost veća od 100km od date lokacije.

Milica Makarić, E2 55/2019