|  |  |  |
| --- | --- | --- |
|  | УНИВЕРЗИТЕТ У НОВОМ САДУ  **ФАКУЛТЕТ ТЕХНИЧКИХ НАУКА У НОВОМ САДУ** |  |

Невена Глигоров

**DebugIt - студентска QA платформа која интегрише *Elasticsearch* механизам претраге**

ДИПЛОМСКИ РАД

- Основне академске студије -

Нови Сад, 2024.

Образац **Q2.НА.15**-**04** - Издање 1

**КЉУЧНА ДОКУМЕНТАЦИЈСКА ИНФОРМАЦИЈА**

|  |  |  |  |
| --- | --- | --- | --- |
| Редни број, **РБР**: | |  | |
| Идентификациони број, **ИБР**: | |  | |
| Тип документације, **ТД**: | | Монографска публикација | |
| Тип записа, **ТЗ**: | | Текстуални штампани документ | |
| Врста рада, **ВР**: | | Завршни-bachelor рад | |
| Аутор, **АУ**: | | Невена Глигоров | |
| Ментор, **МН**: | | доц. др Дуња Врбашки | |
| Наслов рада, **НР**: | |  | |
| Језик публикације, **ЈП**: | | Српски (ћирилица) | |
| Језик извода, **ЈИ**: | | Српски / Енглески | |
| Земља публиковања, **ЗП**: | | Србија | |
| Уже географско подручје, **УГП**: | | Војводина | |
| Година, **ГО**: | | 2024. | |
| Издавач, **ИЗ**: | | Ауторски репринт | |
| Место и адреса, **МА**: | | Факултет Техничких Наука (ФТН), Д. Обрадовића 6, 21000 Нови Сад | |
| Физички опис рада, **ФО**: (поглавља/страна/ цитата/табела/слика/графика/прилога) | | n/n/n/n/n/n/n | |
| Научна област, **НО**: | | Електротехничко и рачунарско инжењерство | |
| Научна дисциплина, **НД**: | | Примењене рачунарске науке и информатика | |
| Предметна одредница/Кqучне речи, **ПО**: | | Неколико кључних термина из рада, односно предмета изучавања на који се рад односи. | |
| **УДК** | |  | |
| Чува се, **ЧУ**: | | Библиотека ФТН, Д. Обрадовића 6, 21000 Нови Сад | |
| Важна напомена, **ВН**: | |  | |
| Извод, **ИЗ**: | | Кратак опис рада који стаје у три до шест реченица. Треба навести проблем који се решавао, како се проблем решавао и који су резултати. | |
| Датум прихватања теме, **ДП**: | |  | |
| Датум одбране, **ДО**: | |  | |
| Чланови комисије, **КО**: | Председник: |  |
|  | Члан: |  | Потпис ментора |
|  | Члан, ментор: | др Дуња Врбашки, доцент |  |

**KEY WORDS DOCUMENTATION**

|  |  |  |  |
| --- | --- | --- | --- |
| Accession number, **ANO**: | |  | |
| Identification number, **INO**: | |  | |
| Document type, **DT**: | | Monographic publication | |
| Type of record, **TR**: | | Textual material, printed | |
| Contents code, **CC**: | | Bachelor thesis | |
| Author, **AU**: | | Nevena Gligorov | |
| Mentor, **MN**: | | Dunja Vrbaški, PhD, assist. prof. | |
| Title, **TI**: | |  | |
| Language of text, **LT**: | | Serbian (cyrillic script)/Serbian (latin script) | |
| Language of abstract, **LA**: | | Serbian/English | |
| Country of publication, **CP**: | | Serbia | |
| Locality of publication, **LP**: | | Vojvodina | |
| Publication year, **PY**: | | 2024 | |
| Publisher, **PB**: | | Author reprint | |
| Publication place, **PP**: | | Faculty of Technical Sciences, D. Obradovića 6, 21000 Novi Sad | |
| Physical description, **PD**: (chapters/pages/ref./tables/pictures/graphs/appendixes) | | n/n/n/n/n/n/n | |
| Scientific field, **SF**: | | Electrical and computer engineering | |
| Scientific discipline, **SD**: | | Applied computer science and informatics | |
| Subject/Key words, **S**/**KW**: | |  | |
| **UC** | |  | |
| Holding data, **HD**: | | Library of the Faculty of Technical Sciences, D. Obradovića 6, 21000 Novi Sad | |
| Note, **N**: | |  | |
| Abstract, **AB**: | |  | |
| Accepted by the Scientific Board on, **ASB**: | |  | |
| Defended on, **DE**: | |  | |
| Defended Board, **DB**: | President: |  |
|  | Member: |  | Menthor's sign |
|  | Member, Mentor: | Dunja Vrbaški, PhD, assist. prof. |  |

|  |  |  |
| --- | --- | --- |
|  | УНИВЕРЗИТЕТ У НОВОМ САДУ ⚫ **ФАКУЛТЕТ ТЕХНИЧКИХ НАУКА**  21000 НОВИ САД, Трг Доситеја Обрадовића 6 | Број: |
|  |
| **ЗАДАТАК ЗА ИЗРАДУ ДИПЛОМСКОГ (BACHELOR) РАДА** | Датум: |
|  |

Za sada ostaviti ovako, prazno, ali me podsetite pre finalne verzije da pošaljem ovu stranicu.

Spisak skraćenica

FTP - *File transfer protocol*

QA - *questions and answers*

GB - *gigabytes*

CRUD - *create, read, update, delete*

TF - *term frequency*

IDF - *inverse document frequnecy*

Sadržaj

[1. Uvod 7](#_Toc176103403)

[1.1 Šta pisati u uvodu? 7](#_Toc176103404)

[2. Istorijat 8](#_Toc176103405)

[3. Značaj i primena 11](#_Toc176103406)

[3.1 Postojeća rešenja 12](#_Toc176103407)

[*Stack Exchange* i *Stack Overflow* 12](#_Toc176103408)

[*Discuss The Elastic Stack* 12](#_Toc176103409)

[*GitHub community* 12](#_Toc176103410)

[Ostalo 12](#_Toc176103411)

[4. *Elasticsearch* mehanizam pretrage 13](#_Toc176103412)

[4.1 Šta je *Elasticsearch* mehanizam pretrage i kako je nastao? 13](#_Toc176103413)

[4.2 Struktura *Elasticsearch* mehanizma pretrage 13](#_Toc176103414)

[*Index* 13](#_Toc176103415)

[*Shard* 14](#_Toc176103416)

[*Cluster* i *node* 15](#_Toc176103417)

[*Index template* i *data mapping* 16](#_Toc176103418)

[*Alias* 17](#_Toc176103419)

[4.3 Način rada *Elasticsearch* mehanizma pretrage 17](#_Toc176103420)

[*Text analysis* 19](#_Toc176103421)

[*Stemming* 19](#_Toc176103422)

[5. Zaključak 20](#_Toc176103423)

[Literatura 21](#_Toc176103424)

[Dodatak A 23](#_Toc176103425)

[Dodatak B 24](#_Toc176103426)

[Podaci o kandidatu 25](#_Toc176103427)

# Uvod

U ovom dokumentu su navedene neke: sadržajne, tehničke i stilske smernice za pisanje završnog rada. Ujedno predstavlja i šablon za izradu i pisanje.

Ostale napomene se nalaze na adresi: [Zavrsni radovi - napomene](https://docs.google.com/document/d/e/2PACX-1vQ9_rgcnXK0Lw5_l6W16CbQ2MQudjqy9lDFMzwJmakopojXnTAzVDxl6xc-M0xCeUS3zr9VKOxlliRm/pub). Obavezno pogledati i ovaj dokument koji se s vremena na vreme dopunjava.

Možete korisiti ovaj dokument kao šablon, možete napraviti novi, a možete pisati i u LaTex-u. Rad se, na kraju, komisiji i biblioteci, predaje u pdf-u.

Prva strana, KDI tabele i zadatak rada treba da ostanu u datom formatu. Ostalo može biti formatirano drugačije.

Sam rad može biti napisan na ćirilici ili latinici s tim da naslovna strana i KDI moraju biti npisane kao što je navedeno.

## Šta pisati u uvodu?

Uvod sadrži kratak opis problema koji se rešava ili oblast koja se istražuje.

Može se navesti i motivacija za izbor određene teme, kontekst, značaj i povezanost sa drugim oblastima ili temama.

Na kraju se obično jednom rečenicom navodi šta koje poglavlje, u nastavku rada, opisuje.

Savet: Ne morate krenuti sa pisanjem odmah od Uvoda. To ume da blokira nekad. Slobodno krenite od glavne stvari pa ćete usput ili na kraju sami shvatiti šta je prirodno da piše u uvodu.

# Istorijat

Alan Emtage, administrator sistema za departman informacionih tehnologija i student kanadskog univerziteta McGill, čije je sedište u Montrealu, u septembru 1990. godine objavljuje prvi Internet mehanizam pretrage (*eng. search engine*) pod nazivom *Archie*. Ime *Archie* potiče od reči „*Archive*“, po uzoru na dugačak i mukotrpan proces pretrage kroz koji je Alan prolazio kako bi razvio svoj softver [1].

*Archie* korisnički interfejs je poprilično jednostavan, kao što je prikazano na slici 1, uzimajući u obzir da je njegova glavna namena pretraživanje fajlova na FTP sajtovima koji se nalaze na Internetu.



Slika 1. Archie interfejs [2]

Univerzitet je pred Alana postavio težak zadatak u to veme, a to je da kreira softver uz pomoć kog je moguće povezati se na Internet. Novokreirani *Archie* je predstavljao indeks FTP fajlova i njegovi korisnici su uz pomoć jednostavnih upita mogli da pretražuju fajlove. Ovi upiti su morali da budu dosta precizni, kako bi korisnici dobili ono što su želeli, pošto *Archie* nije imao podršku za upite napisane prirodnim jezikom (*eng. natural* language), niti je indeksirao sadržaj fajlova. Kako nije imao mogućnost da dâ korisnicima informaciju o tome šta se nalazi u fajlovima koji su im ponuđeni kao rezultat pretrage, korisnici su te fajlove morali da skinu (*eng. download*) na svoj računar i pogledaju njihov sadržaj.

Do 1993, *Archie* je postao veoma popularan i korišćen od strane drugih univerziteta širom sveta [3].

Nekoliko godina i mehanizama pretrage kasnije, apsolutnu dominaciju u Internet svetu ima Yahoo. Bilo da je ime *backronym* za „*Yet Another Hierarchically Organized Oracle*“ ili „*Yet Another Hierarchical Officius Oracle*“, njegova struktura je ista, takva da predstavlja direktorijum koji se sastoji od Internet sajtova, prilikom čega su sajtovi organizovani hijerarhijski. Primer interfejsa videti na slici 3 [4].



Slika 4, Yahoo! interfejs

*Yahoo* se temelji na ideji dvojice studenata sa Stanford Univerziteta, Jerry Yang i David Filo, čija je zamisao bila da naprave direktorijum pod nazivom „*Jerry and David's Guide to the World Wide Web*“, uz pomoć kojeg bi se navigirali do svojih omiljenih sajtova.

Zahvaljujući Internet pregledaču *Navigator* koji je postavio link do njihovog projekta na vrh svoje stranice, direktorijum koji su napravili Jerry i David polako postaje sve popularniji, tačnije, postaje toliko popularan da odlučuju da svoju ideju pretoče u biznis.

Iako je postojalo dosta drugih mehanizama pretrage, poput: *Lycos*, *AskJeeves*, *Excite*, *AOL*, *AltaVista*, *Infoseek* i drugih, *Yahoo* je stajao na samom vrhu i niko mu nije bio ni blizu. 1998. godine dvojica studenata odlaze u *Yahoo* sa ponudom da im prodaju svoj biznis za milion dolara, međutim *Yahoo* ih je odbio. Ono što tada *Yahoo* nije znao jeste da je odbi0 budućeg giganta koji i dan danas vlada Internetom, čiji primer interfejsa je moguće videti na slici 4 [6].



Slika 4. Google interfejs [8]

*Google* je polako postajao glavni protivnik *Yahoo* kompanije i *Yahoo*, imajući to u vidu, ponudio je *Google* da integriše njihov mehanizam pretrage u okviru svoje Internet stranice. Ovim potezom, *Google* je dobio promociju od strane kompanije *Yahoo*, kojoj to nije bila zamisao, i time se polako peo ka samom vrhu.

Larry Page i Sergey Brin su bili studenti Stanford Univerziteta kada su napravili prvi mehanizam pretrage koji je radio po principu analiziranja linkova tipa *backlink* (linkovi u okviru stranica koji upućuju na druge stranice) i na osnovu njih je određivao koliko su bitne pojedine stranice na Internetu. Projekat je dobio naziv *Backrub*, upravo zbog načina rada ovog mehanizma.

Želeli su da „organizuju sve informacije sveta i da ih učine univerzalno pristupačnim i korisnim“ i u tom duhu su *Backrub* preimenovali u *Google*, što predsatvlja matematički izraz za broj 1 koji je praćen sa 100 nula (1 googol).

Osnova *Google* mehanizma pretrage se zasniva na *PageRank* algortimu, koji koristi opisani koncpet *Backrub* mehanizma. Drugim rečima, ovaj algoritam istražuje koliko drugih sajtova referencira neku stranicu da bi odredio koliko je ta stranica zaista relevantna.

Naročitu pažnju velikog broja kompanija je privukao *Google* algoritam za reklame. Algoritam koristi reči koje korisnik unese tokom pretrage kako bi oforimo skup reklama koje mu treba prikazati [5][7][9].

Kada je potrebno nešto pretražiti, ljudi će za to obično reći da je to potrebno „izguglati“, što već dovoljno govori o tome gde se danas nalazi Google i njihov mehanizam pretrage. Pored ovog mehanizma, među popularnim mehanizmima su još i: *Microsoft Bing*, *Yahoo*, *Baidu* i drugi.

# Značaj i primena

Na fakultetima širom sveta prolazi i smenjuje se bezbroj studenata. Svaki od njih je tokom svojih studentskih dana uradio barem nekoliko projekata, što na zahtev predmetnog profesora, što iz radoznalosti i želje za unapređivanjem. Iako različiti, jedna stvar je zajednička svim studentima, a to je da uvek pitaju za savete i iskustva nekog od svojih starijih kolega. Iz tih razloga su tokom godina nastajali različiti *Github* repozitorijumi, *DropBox* linkovi, kompresovani folder i slično, puni materijala uz pomoć kojih je lakše položiti ispite.

Ideja podele materijala novim generacijama jeste vrlo dobra i odlično se pokazala kada je u pitanju teorijski deo ispita. Imajući to u vidu, nameće se jedno vrlo prosto pitanje, šta se dešava sa projektima koje su starije kolege radile? Projekti ne treba da se recikliraju godinama, zbog čega profesori i asistenti svake godine smišljaju nove za nove generacije studenata i time već urađeni projekti postaju u dovoljnoj meri neupotrebijivi.

Međutim, svaki od tih projekata je prilikom izrade imao po neki problem (*eng. bug*), a dosta njih je imalo slične ili iste probleme. Studenti koji su radili na njima su možda potrošili malo više vremena dok su pronašli adekvatno rešenje u moru informacija koje se danas nalaze na Internetu. Iz generacije u generaciju, studenti imaju slične probleme prilikom izrade svojih projekata bez uvida u to da li je neko od starijih kolega imao sličan problem, ali i bez načina kako ga je rešio.

Jedan od potencijalnih rešenja koje se koristi trenutno jeste da se uvede po 1 dokument sa pitanjima i odgovorima (*eng. QA sheet*) za svaku generaciju i svaki predmet. U okviru ovog dokumenta, studenti bi opisivali problem na koji su naišli i onda bi čekali odgovor nekog od asistenata. Ovo rešenje je dobro na generacijskom nivou, jer su onda svi studenti iste generacije u mogućnosti da vide odgovor na problem sa kojim se susreo neki od njihovih kolega. Glavni nedostatak ovakvog pristupa jeste što se taj dokument briše i pravi novi, za svaki predmet i za svaku generaciju ponovo.

Umesto dokumenata sa pitanjima i odgovorima, mnogo bolje rešenje je da postoji sajt koji bi obuhvatio sva implementaciona i diskusiona pitanja studenata, nezavisno od predmeta i tehnologija. Na ovakav način se postiže:

* Svaki student može da pogleda bilo koje pitanje i njegove odgovore, bez obzira na to kada je postavljeno.
* Studenti, asistenti i profesori mogu da odgovore na postavljena pitanja. Na takav način student koji je postavio pitanje brže dolazi do odgovora koji mu je potreban, a asistenti i profesori postaju rasterećeniji od gomile pitanja koje bi inače dobili.
* Najčešći problemi koji se javljaju tokom izrade projekata će vrlo brzo imati odgovore i neće biti potrebe za dupliranjem istih pitanja.
* Studenti se navikavaju da rešenje koje su pronašlo prilagode svom kodu.

## Postojeća rešenja

### *Stack Exchange* i *Stack Overflow*

*Stack Overflow* je sajt, zamišljen da predstavlja biblioteku koja bi umesto knjiga na svojim policama imala sva moguća pitanja vezana za programiranje. Detaljni i najkvalitetniji odgovori bi se nalazili na njihovim stranicama. Programeri širom sveta su upravo oni koji omogućuju da se ovaj cilj postigne, pomažući potpunim strancima na Internetu, nudeći im odgovore za opisane probleme [10].

*Stack Overflow* je samo jedan od 173 sajta koja se nalaze u okviru *Stack Exchange* mreže, koju na mesečnom nivou poseti preko 100 miliona ljudi. Trenutno predstavlja najmoćniju mrežu za pronalaženje odgovora, sa *Stack Overflow* kao vodećim sajtom. [11].

### *Discuss The Elastic Stack*

*Discuss The Elastic Stack* predstavlja skup „diskusionih foruma za *Elasticsearch*, *Beats*, *Logstash*, *Kibana*, *Elastic Cloud* i ostale proizvode u okviru *Elastic* ekosistema“ [12].

Slično kao i *Stack Overflow*, ovaj sajt je namenjen za rešavanje problema nastalih prilikom implementacije rešenja, koja su usko vezana za *Elastic* domen.

### *GitHub community*

*GitHub community* je sajt namenjen *GitHub* korisnima, kao mesto na kome mogu da dobiju odgovor na postavljeno pitanje i da pritom nauče nešto novo ili čak dobiju inspiraciju za neki novi projekat [13].

### Ostalo

Osim navedenih sajtova, koriste se i Dev.to, Experts Exchange, Code Project i slični.

# *Elasticsearch* mehanizam pretrage

## Šta je *Elasticsearch* mehanizam pretrage i kako je nastao?

*Elasticsearch* je distribuirani, *open source* mehanizam za pretragu i analitiku, čiji temelj predstavlja *Apache Lucene* bibilioteke. Omogućava skladištenje, pretragu i analiziranje velike količine podataka u realnom vremenu.

*Elasticsearch*, danas jedan od najpopularnijih mehanizama pretrage, nastao je kada je Shay Banon, suosnivač *Elastic* kompanije, želeo da kreira mehanizam pretrage koji bi njegova žena mogla da koristi za svoje recepte za kuvanje [16].

## Struktura *Elasticsearch* mehanizma pretrage

### *Index*

*Elasticsearch* koristi svoju ugrađenu NoSQL bazu koja podatke čuva kao JSON dokumente. Dokumenti predstavljaju osnovnu jedinicu podatak koju je moguće grupisati u indeks (*eng. index*), koncept sličan tabeli kod relacionih baza podataka. Da bi se dokumenti grupisali u indeks, moraju imati slične karakteristike [14].

Postoje 2 tipa strukture podataka koje se koriste kod mehanizama pretrage u cilju čuvanja i organizovanja podataka:

* ***forward index*** - dokumenti se mapiraju na termine koje sadrže. To znači da se čuva lista svih reči koje se nalaze u svakom dokumentu. Ova struktura je dobra kada je u pitanju indeksiranje, pošto se svaka reč nadovezuje na prethodnu, ali s druge strane, nije previše efikasna kada je reč o pretrazi na osnovu termina. Na slici 5 se nalazi primer koji prikazuje kako se korišćenjem *forward index* strukture podataka smešta 5 dokumenata i njihov sadržaj. Ukoliko korisnik želi da pronađe sve dokumente u kojima se pominje reč „*Elasticsearch*“, mehanizam bi morao da prođe kroz sve dokumente kako bi korisniku vratio rezultat „dokument1, dokument4“.



Slika 5. Forward index

* ***inverted index*** - termini se mapiraju na dokumente u kojima se nalaze. To znači da za svaku reč postoji lista dokumenata u kojima se nalazi. Kod ovakve strukture je indeksiranje sporije, pošto je za svaku reč potrebno proveriti da li već postoji u indeksu, ali je pretraga veoma brza, iz razloga što se podaci već unapred čuvaju kao upiti (*eng. queries*). To omogućava da pretrage budu jako brze, čak i kada postoji veoma velika količina podataka. *Elasticsearch* mehanizam pretrage koristi upravo ovu strukturu za smeštanje svojih podataka. Na slici 6 se nalazi primer koji prikazuje kako se korišćenjem *inverted index* strukture podataka smešta 5 dokumenata i njihov sadržaj. Ukoliko korisnik želi da pronađe sve dokumente u kojima se pominje reč „Elasticsearch“, mehanizam odmah može da vrati rezultat „dokument1, dokument4“, upravo iz razloga što se podaci čuvaju kao upiti [15].



Slika 6. Inverted index

### *Shard*

„*Shard* predstavlja osnovnu gradivnu jedinicu *Elasticsearch* distribuirane arhitekture.“ Indeks može da sadrži veoma veliku količinu podataka. a bi upravljanje podacima bilo što efikasnije, ti podaci se obično podele u nekoliko *shard* komponenti [18]. Na slici 7 se nalazi primer indeksa podeljenog na 6 *shard* komponenti.



Slika 7. Shards in index

Postoje 2 tipa *shard* komponenti:

* **primarni** (*eng. primary*) - sadrže originalne podatke i obavljaju indeksiranje i pretragu. Broj primarnih *shard* komponenti se definiše prilikom kreiranja indeksa i ne može se naknadno menjati.
* **kopije** (*eng. replica*) - predstavljaju kopiju primarnih *shard* komponenti. Uz pomoć njih se uvodi redudantnost i poboljšavaju se performanse pretrage, jer omogućavaju izvršavanje paralelnih upita. Broj kopija se može dinamički menjati, prilikom čega se automatski balansira broj kopija među node komponentama u cluster komponenti. Preporuka je da svaka primarna *shard* komponenta ima svoju kopiju [17].

„Broj *shard* komponenti jedne *data node* komponente je proporcijalan *heap* memoriji te *node* komponente. Node koji ima 30GB *heap* memorije treba da ima najviše 600 shard komponenti“, a što ih je manje, time bolje. Korišćenjem previse shard komponenti se bespotrebno zauzima memorija i time usporava pretraga, indeksiranje zahteva i druge operacije [19].

Najveća efikasnost pretrage se postiže ukoliko se upiti izvršavaju na više raličitih *shard* komponenti paralelno, prilikom čega se sve *shard* komponente nalaze u okviru različitih *node* komponenti [18].

Prednosti korišćenja *shard* komponenti:

* **horizontalno skaliranje** - „*sharding*“ je isto što i horizontalno particionisanje, a to je horizontalna podela podataka u više baza. To znači da se jedan set podataka sa svim poljima prebaci u jednu bazu, a drugi set podataka sa svim njegovim poljima prebaci u drugu bazu, kao što je prikazano na slici 8. Zahvaljujući tome, *cluster* može da se skalira horizontalno prilikom priliva novih podataka.



Slika 8. Sharding [20]

* **bolje performanse pretrage** - upiti se mogu vršiti nad više *shard* komponenata u isto vreme, što je mnogo brže od izvršavanja upita nad indeksom koji sadrži samo jednu *shard* komponentu.
* **veća dostupnost** - *node* komponente mogu da prestanu da rade iz mnogo razloga, bilo zbog nedostatka resursa, problema sa konekcijom ili nekih konfiguracionih grešaka. Praksa je da se prilikom raspodele *shard* komponenti po *node* komponentama, u istoj *node* komponenti ne nađe primarna *shard* komponenta i njena kopija [20]. To znači da iako neki node otkaže, podaci idalje ostaju dostupni krajnjem korisniku, jer se sve operacije preusmeravaju na one *shard* komponente koje su dostupne [17].

### *Cluster* i *node*

*Node* predstavlja server na kome se izvršava jedna *Elasticsearch* instanca. Jedna ili više *node* komponenti povezanih zajedno čine *cluster*. Svaki *node* u okviru jedne *cluster* komponente zna za ostale *node* komponente, što omogućava preusmeravanje zahteva ka odgovarajućoj *node* komponenti.

*Node* komponente imaju svoje uloge (*eng. roles*) i na osnovu toga se razlikuje više node tipova:

* ***master-eligible node*** - ova uloga omogućuje da *node* ima mogućnost da bude izabran za *master node*, čije je zaduženje da upravlja celokupnom *cluster* komponentom. ***Master node*** je zadužen za kreiranje/brisanje indeksa, praćenje koje *node* komponente su u okviru njegove *cluster* komponente i odlučuje koje *shard* komponente će biti dodeljene kojim *node* komponentama.
* ***data node*** - sadrži podatke i izvršava CRUD operacije, pretragu i agregaciju.
* ***ingest node*** - ima mogućnost korišćenja *ingest pipeline* procesora, uz pomoć kojih se sekvencijalno vrše transformacije nad podacima pre njihovog indeksiranja [23].
* ***remote-eligible node*** - može se ponašati kao krajnji korisnik
* ***machine-learning node*** - *node* koji se mora dodati kako bi moglo da se koristi mašinsko učenje, a može ih biti i više.
* ***transform node*** - *node* koji se mora dodati kako bi mogle da se koriste transformacije nad podacima. Transformacije omogućavaju kreiranje sumarizovanih indeksa od postojećih [24].
* ***coordinating node*** - vrši koordinaciju između dve faze pretrage. Prva faza je *scatter* faza i u njoj kooridnatorski node prosleđuje zahtev onim *node* komponentama koji imaju tražene podatke. Nakon što *data node* komponente izvrše zahtev, rezultat vraćaju koordinatoru. Sledi *gather* faza u kojoj koordinator uzima sve rezultate koje je dobio i stapa ih u jedan. Svaki node je implicitno kooridnatorski, međutim ukoliko postoji neki node kome nije eksplicitno navedena nijedna uloga, taj node će onda jedini biti koordinatorski [22].

Na slici 9 se nalazi primer jednog distriburanog sistema, koji se sastoji od tri servera, odnosno, *node* komponente. *Shard* komponente su raspoređene tako da ne postoje primarna i kopija *shard* komponente u istoj *node* komponenti. *Shard* komponente sa slovo P na početku označavaju primarne, a sa slovom R kopije *shard* komponenti.



Slika 9. Cluster nodes with shards [25]

### *Index template* i *data mapping*

*Index template* je skup podešavanja koja se primenjuju nad indeksom prilikom njegovog kreiranja. U okviru njega se definišu brojevi primarnih *shard* komponenti i njihovih kopija, *data mapping*, prioritet i ostalo [26].

*Data mapping* je šema indeksa, odnosno, sadrži informacije o svim poljima i njihovima tipovima koje dokumenti u okviru tog indeksa imaju. Postoje dva tipa za *data* mapping:

* **statički** - unapred su definisana polja i tipovi,
* **dinamički** - prilikom indeksiranja dokumenata, *Elasticsearch* automatski osvežava (*eng. update*) šemu, kofigurišići imena polja i njihove tipove.

Korišćenjem statičkog *data mapping* tipa, može se definisati indeks koji bi imao polja i tipove redom: ime: *string*, prezime: *string* i brojGodina: *int*, kao što je prikazano na Listingu 1 sa leve strane. Kada bi stigao dokument koji se nalazi sa desne strane na istom listingu, bio bi odbijen, zato što se tip vrednosti za polje „brojGodina“ ne poklapa sa definisanim u šemi indeksa.

Ukoliko bi se umesto statičkog *data mapping* tipa koristio dinamički, onda bi dokument sa desne strane bio prihvaćen, pri čemu bi se napravila izmena u šemi indeksa, gde bi se za tip vrednosti za polje „brojGodina“ prihvatao i *int* i *string* [27].

{ {

"ime":"Marko", "ime":"Marko",

"prezime":"Markovic", "prezime":"Marko",

"brojGodina":35 "brojGodina":"35"

} }

*Listing 1: Dva dokumenta sa različitim tipovima za polje „brojGodina“*

### *Alias*

Aliasi se koriste za grupisanje indeksa. Na listingu 2 se nalazi primer kreiranja aliasa, gde se nalaze svi logovi vezani za indekse: korisnik, pitanje i komentar [28].

POST \_aliases

{

"actions": [

{

"add": {

"indices": ["user-logs","question-logs", "comment-logs"],

"alias": "logs"

}

}

]

}

*Listing 2: Alias, kao skup više indeksa*

## Način rada *Elasticsearch* mehanizma pretrage

Osnovna namena bilo kog mehanizma pretrage, pa tako i *Elasticsearch*, jeste da korisniku ponudi rezultat sa svim dokumentima koji odgovaraju otkucanom upitu. Postavlja se pitanje kako korisnci znaju od svih ponuđenih dokumenata koji je najbolji, odnosno, koji najviše odgovara? *Elasticsearch* je rešenje našao u tome da svakom dokumentu dodeli *score*. Što je *score* veći, to dokument ima veću relevantnost za korisnika. Dokumenti kojima je dodeljen *score* 0 nemaju nikakvu relevantnost za korisnika i oni se ne ubrajaju u krajnji rezultat. U zbirnom rezultatu se dokumenti ređaju u opadajućem redosledu na osnovu *score* vrednosti.

Moguće je koristiti različite tipove upita, odnosno prertrage, među kojima se najčešće koriste:

* **struktuirana pretraga** (*eng. structured search*) - koristi se za upite nad podacima koji su inherentni, odnosno koji predstavljaju celinu, kao što su: datum, vreme i brojevi. Ovi upiti se koriste za traženje dokumenata kod kojih odgovarajuća polja imaju: stopostotno poklapanje, nalaze se u granicama od do i slično. Razlog tome je što *term level queries* koji se koriste za pretragu nemaju fazu analize (o tome više u narednim poglavljima). Postoji mogućnost vršenja pretrage i nad tekstualnim podacima, ali samo u slučaju podataka čije su vrednosti kratke i precizne, poput oznaka (*eng. tags*) objava. Svakom dokumentu *Elasticsearch* ne dodeljuje *score*, već „*da*“ ili „*ne*“, što označava da dokument odgovara ili ne odgovara upitu. Na listingu 3 se nalazi primer dobavljanja dokumenata postavljenih pre najviše 7 dana iz *Question* indeksa.

var response = \_elasticClient.Search<Question>(s => s

.Query(q => q

.Bool(b => b

.Filter(f => f

.DateRange(r => r

.Field(ff => ff.PostedOn)

.GreaterThanOrEquals(DateTime.UtcNow.AddDays(-7))

)

)

)

)

);

*Listing 3: Struktuirana pretraga*

* **nestruktuirana pretraga** (*eng. unstructured search*) - cilj ovog tipa pretrage je da na osnovu otkucanog upita pronađe dokumente u kojima postoji najviše poklapanja, tako što se dokumentima dodeljuje *score*. Koristi *full text queries*, koji za razliku *term level queries* ima fazu analize u kojoj se sve reči u upitu porede sa onim u inverted index. Najčešće se koristi za upite nad tekstualnim poljima. Na listingu 4 se nalazi primer upita gde je potrebno pronaći sva dokumenta iz *Question* indeksa, kod kojih polje *Title* sadrži reč „*implementacija*“.

var response = \_elasticClient.Search<Question>(s => s

.Query(q => q

.Match(m => m

.Field(f => f.Title)

.Query("implementacija")

)

)

);

*Listing 4: Nestruktuirana pretraga*

* **kombinovni upiti** (*eng. combining queries*) - kombinacija upita koja se naziva *compound query*. Na listingu 5 se nalazi primer kombinovanog upita koji kombinuje upite sa listinga 3 i listiga 4. Iz *Question* indeksa se dobavljaju svi dokumenti kod kojih polje *Title* sadrži reč „*implementacija*“. Nakon toga se vrši filtracija dokumenata, tako da krajnji rezultat sadrži samo ona pitanja koja su postavljena u poslednjih 7 dana [30].

var response = \_elasticClient.Search<Question>(s => s

.Query(q => q

.Bool(b => b

.Must(m => m

.Match(mm => mm

.Field(f => f.Title)

.Query("implementacija")

)

)

.Filter(f => f

.DateRange(r => r

.Field(ff => ff.PostedOn)

.GreaterThanOrEquals(DateTime.UtcNow.AddDays(-7))

)

)

)

)

);

*Listing 5: Kombinovani upit*

### *Text analysis*

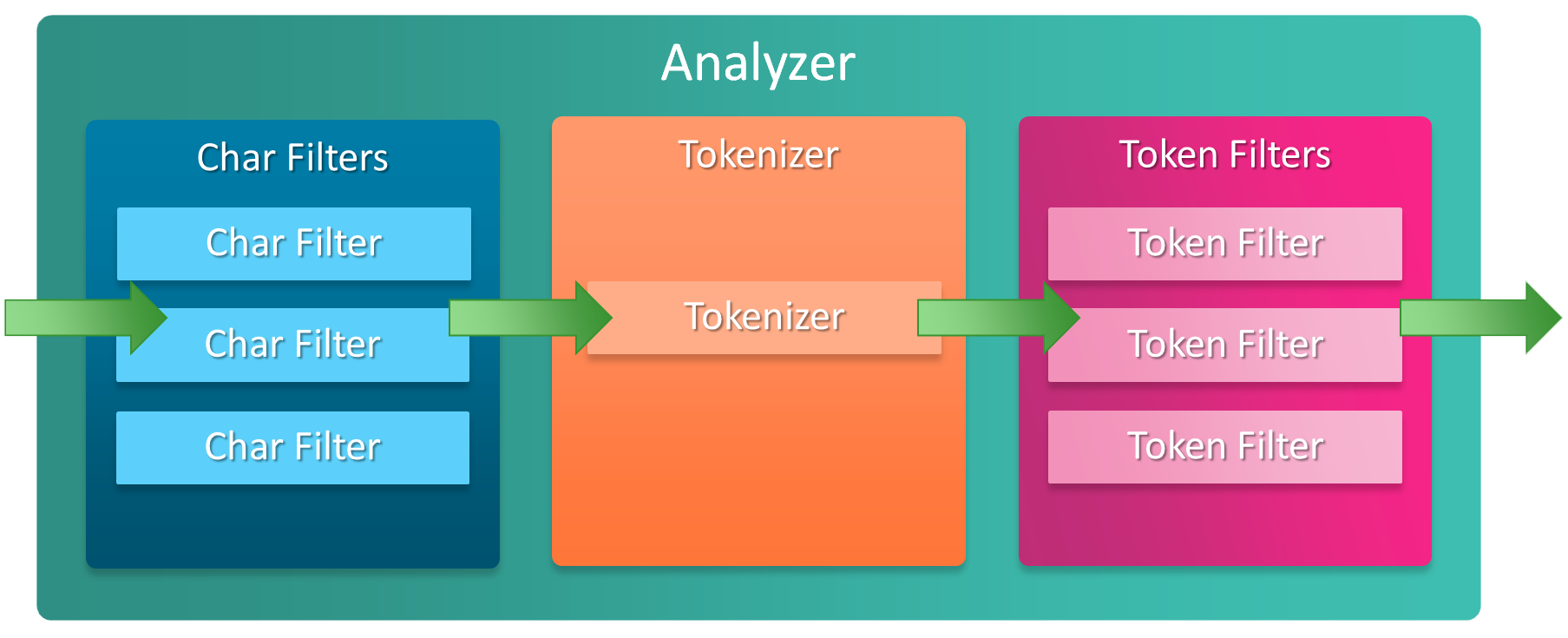
Prilikom čuvanja i pretraživanja podataka, sva tekstualna polja prolaze kroz *Text analysis* proces, koji ima dve osnovne funkcionalnosti:

* **tokenizacija** (*eng. tokenization*) - proces u kome *tokenizer* komponenta vrši deljenja teksta na tokene, najčešće reči, na osnovu nekog definisanog pravila.
* **normalizacija** (*eng. normalization*) - proces u kome se vrše transformacije tokena. Transformacije mogu biti različite, od *stemming* procesa do posmatranju sinonima i *stop words* i tako dalje [31].

*Analyzer* modul izvršava *text analysis* proces. Elasticsearch ima nekoliko ugrađenih tipova *analyzer* modula i oni imaju podršku za rad za različite jezike. Osim njih, moguće je kreirati i sopstveni *analyzer* modul, kombinujući komponente.

*Analyzer* modul je sastavljen od tri tipa komponenti kroz koje tekst prolazi u toku *text analysis* procesa, a primer njegove strukture je moguće videti na slici 10:

* ***character filters*** - koriste se za dodavanje, menjanje i brisanje karaktera iz niza karaktera kojim je predstavljen tekst na ulazu. Izvršavaju se redom, jedan za drugim. Primer primene bi bio uklanjanje *HTML* tagova iz teksta. *Analyzer* modul može imati nijedan ili više *character filter* komponenti.
* ***tokenizer*** - na ulazu dobija tekst predstavljen kao niz karakter koji deli na tokene, najčešće to budu reči, na osnovu definisanog pravila. Tekst se može podeliti na tokene posmatranjem: razmaka, određenog slova, šablona i slično. *Analyzer* modul ima tačno jedan *tokenizer*.
* ***token filters*** - na ulazu dobijaju niz tokena. U taj niz je moguće dodati novi, izmeniti ili obrisati postojeći token, u cilju što bolje normalizacije teksta. Izvršavaju se redom, jedan za drugim. *Analyzer* modul može imati nijedan ili više *token filters* komponenti [33].



Slika 10. Struktura analyzer modula [34]

U okviru *text analysis* procesa je moguće koristiti i *normalizer* modul koji za razliku od *analyzer* modula na izlazu daje samo jedan, a ne niz tokena, što znači da nema *tokenizer* komponentu. Još jedna razlika između ova dva modula je što *normalizer* modul ima manji skup *character* i *token* filtera. Koristi se za polja koja su tipa *keywords*.

### *Stemming*

Stemming je proces u kome se reč redukuje na svoj koren. Koren na koji je reč redukovana ne mora biti prava reč, odnosno, ne mora ni zaista biti koren te reči. Bitno je da se postigne da iste reči napisane u različitom obliku budu prepoznate kao iste.

Koliko dobre rezultate će *stemming* proces dati uveliko zavisi od jezika nad kojim se primenjuje, ali i od tipa *stemmer token filter* komponente koja se koristi.

*Stemming token filters* su zaduženi za *stemming* i postoje dve vrste:

* ***algorithmic stemmers*** - koriste set pravila koja se primenjuju na tokene. Primer bi bio skidanje prefiksa i sufiksa sa reči. Lako se podešavaju i koriste vrlo malo memorije, a pritom su brži od *dictionary stemmers*. Međutim, ovaj tip *stemming token filters* komponenti ne radi dobro sa rečima koje ne sadrže koren u svom obliku.
* ***dictionary stemmers*** - koriste rečnik koji je potrebno navesti pri implementaciji. Za svaku reč, odnosno token, traže koren u okviru datog rečnika. Ovim se rešava problem pronalaženja korena za izuzetke, ali i reči koje deluju kao da imaju zajednički koren, ali je smisao tih reči potpuno međusobno različita. Iako se ova opcija za *stemming token filters* čini kao bolja, izbegavа se, zato što je potrebno naći provereno dobar rečnik, koji ima što veći broj reči, a takvi su često retko dostupni. Osim toga, da bi rečnik mogao da se koristi, potrebno je da bude celokupno učitan u RAM memoriju, što znači da je potrebno učitati sve njegove reči, prefikse i sufikse, što poprilično usporava *stemming* poces.

Bez obzira na to koji *stemming token filters* se koristi, potrebno je iste primeniti na tekst i prilikom upisivanja u indeks i prilikom pretrage, s obzirom da *stemming* menja oblik reči [32].

### *Stop words*

*Stop words* predstavljaju skup reči koje ni na koji način ne doprinose pretrazi kako bi njen rezultat bio bolji, pa se iz tih razloga ignorišu. To su reči koje se često koriste u svakodnevno govoru, a za srpski jezik bi to bili: veznici, zamenice, predlozi, pomoćni glagoli i slično. S*top words* se koriste uz pomoć *stop token* filtera. *Elasticsearch* nudi listu predefinisanih *stop words* za više jezika koju je moguće koristiti. [35][36]

### *Synonyms*

*Synonyms* predstavljaju jednu od ključnih stvari za uspešnu pretragu. Njihovim korišćenjem se poboljšava rezultat pretrage, jer je moguće naći i one dokumente koji sadrže sinonime umesto otkucanih reči u upitu.

*Synonyms* se čuvaju u *synonyms sets* u vidu *synonym rules* celina. *Synonym rules* definišu skup reči koje su sinonimi [38]. Svaki *synonyms set* je potrebno sačuvati u okviru *Elasticsearch* mehanizma kako bi mogao da se koristi.

Postoje tri načina na koje je moguće čuvati *synonyms sets*:

* ***synonyms API*** - najfleksibilniji način. Sve izmene nad *synonyms set* su dinamičke i analyzer koji referencira taj *set* se sam automatski učitava po potrebi.
* ***synonyms file*** - kreiranje fajlova koji sadrže *synonym sets*. Potrebno je da svaki *node* u *cluster* komponenti sadrži ovaj fajl. Sa svakom izmenom fajla potrebno je okačiti ažurirani fajl za svaki *node* u *cluster* komponenti i ponovo učitati *analyzer* koji se koristi.
* ***inline*** - sinonimi se direktno dodaju u *token filter*. Nije preporučljivo.

Prilikom korišćenja *synonyms* potrebno je voditi računa o tome da postoji *synonyms set* koji neki indeks referencira. U suprotnom, indeks će biti parcijalno kreiran i neće moći da se koristi. Kod kreiranih *synonyms set* treba voditi računa da su sve *synonym rules* celine dobro napisane, u suprotnom *analyzer* neće moći da se učita.

Potrebno je definisati *token filter* za korišćenje *synonyms* i za tu svrhu postoje dva *token filter* tipa:

* ***synonym graph*** - koristi se kada jedna reč ima više različitih sinonima.
* ***synonym*** - ne preporučuje se za slučajeve kada jedna reč ima više različitih sinonima [37].

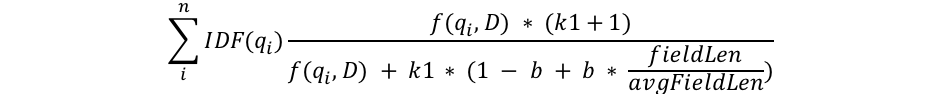
### *Scoring mehanizam*

Prilikom pretrage, *Elasticsearch* svakom dokumentu, na osnovu upita koji je korisnik uneo, dodeljuje *score* vrednost. Uz pomoć te vrednosti se određuje relevantnost dokumenta za krajnjeg korisnika i vrši se sortiranje dokumenata, koja se nalaze u krajnjem rezultatu, u opadajućem redosledu.

*Scoring* mehanizam se zasniva na *BM25* algoritmu. Ovaj algoritam posmatra tri parametra prilikom određivanja *score* vrednosti dokumenta:

* ***term frequency (TF)*** - broj puta koji se određena reč iz upita nalazi u dokumentu. Što je reč češća u ovkiru dokumenta, to je *TF* vrednost veća i dokument ima veću relevantnost.
* ***inverse document frequency (IDF)*** - određuje koliko je reč iz upita relevantna na osnovu toga koliko se često pominje u svim dokumentima u okviru indeksa. Što je reč češće prisutna u dokumentima, to je manje relevantna i dobija manju *IDF* vrednost i obrnuto, što se ređe pojavljuje smatra se više relevantnom i dobija veću *IDF* vrednost.
* ***field length normalization*** - posmatra dužinu polja u kome je pronađena reč iz upita. Što je polje duže, to je ova vrednost manja. Kraća polja dokumenata u kojima je pronađena reč dobijaju veću vrednost za ovaj parametar.

Ovi parametri se računaju za svaku reč koja se nalazi u upitu. Formula koja se koristi za računanje ukupne *score* vrednosti dokumenta je prikazana na slici 11.



Slika 11. BM25 formula [40]

Korišćenjem *Explain API* koji nudi *Elasticsearch*, moguće je bolje razumeti način dodeljivanja vrednosti paramterima. Na listingu 6 se nalazi primer za *Explain API endpoint*, čiji je deo rezultata prikazan na slici 12 [39].

GET question-index/\_explain/-7

{

"query": {

"match": {

"tags": "dependency-injection"

}

}

}

*Listing 6: Explain API endpoint*



Slika 12. deo Explain API rezultata

Featuers of serch engines

Pogledati o NEST biblioteci na strani 10 u StuffAboutSearchEngine.docx

Dodati aggregations u deo o strukturi Elasticsearch mehanizma pretrage mozda

# Zaključak

# Literatura

1. Capitol Technology University, „Alan Emtage Creator of Archie, the World's First Search Engine“*, https://www.captechu.edu/blog*, Artikal o kreatoru prvog mehnaizma pretrage.
2. LinkedIn, Archie interfejs, *https://media.licdn.com/dms/image*, Archie interfejs.
3. Stackscale Group Aire, „Archie, the first Internet search engine“*, https://www.stackscale.com/blog*, Artikal o prvom mehanizamu pretrage.
4. DBPedia, „History of Yahoo!“, *https://dbpedia.org*, Kratka istorija o nastanku Yahoo! mehanizma pretrage.
5. BusinessHistory, „Yahoo!: The first king of the internet“, *https://businesshistory.domain-b.com/focus*, Nastanak i razvijanje Yahoo! mehanizma pretrage.
6. Chatri Sityodtong, „Lessons From Yahoo's Fall: $125b to $5b“, *https://chatrisityodtong.com/blog/entrepreneurship*, Artikal o tome kako je Google zamenio Yahoo!.
7. Google, „From the garage to the Googleplex“, *https://about.google*, Nastanak Google kompanije.
8. Vecteezy, *https://static.vecteezy.com*, Google interfejs.
9. Search Engine Land, „A guide to Google: Origins, history and key moments in search“, *https://searchengineland.com/guide*, Istorija Google kompanije i njihovog mehanizma pretrage.
10. Stack Overflow, „Welcome to Stack Overflow“, *https://stackoverflow.com*, Upoznavanje sa Stack Overflow sajtom.
11. Stack Exchange, „The world's largest programming community is growing“, *https://stackexchange.com*, O Stack Exchange mreži.
12. Elastic, „About Discuss the Elastic Stack“, *https://discuss.elastic.co*, O Discuss the Elastic Stack sajtu.
13. GitHub Communitz, „Overview“, *https://github.com/community*, Opis čemu je namenjen GitHub Community.
14. Jay Gopalakrishnan, „Elasticsearch: What It Is, How It Works, And What It's Used For“, *https://www.knowi.com/blog*, Informacije o tome šta je i kako radi Elasticsearch.
15. GeeksForGeeks, „Difference between Inverted Index and Forward Index“, *https://www.geeksforgeeks.org*, Opis razlike između inverted i forward index strukture podataka sa primerima.
16. Elastic, „Recipes, Elasticsearch. More recipes, morphood + Elasticsearch“, *https://www.elastic.co/elasticon*, Kratak pregled za Webinar .
17. Opster Team, „Understanding Shards in Elasticsearh“, *https://opster.com/guides/elasticsearch*, Osnove shard komponenti u Elasticsearch mehanizmu pretrage.
18. Opster Team, „Elasticsearch Shards“, *https://opster.com/guides/elasticsearch*, Primeri i dobre prakse korišćenja shard komponenti.
19. Elastic, „Size your shards“, *https://www.elastic.co/guide/en/elasticsearch/reference/7.17*, Preporuke o odabiru broja shard komponenti.
20. Nader Medhat, „Understand Database Sharding The Good and Ugly“, *https://miro.medium.com*.
21. Oskar Polak, „Do all shards (within index) have the same content?“, *https://stackoverflow.com/questions*, Pitanje vezano za sadržaj shard komponenti.
22. Elastic, „Nodes“, *https://www.elastic.co/guide/en/elasticsearch*, Objasnjenje node komponenti u Elasticsearch mehanizmu pretrage.
23. Elastic, „Ingest pipelines“, *https://www.elastic.co/guide/en/elasticsearch*, Teorija i primeri.
24. Elastic, „Transforming data“, *https://www.elastic.co/guide/en/elasticsearch*, Teorija.
25. Sematext, „How to Find and Fix Elasticsearch Unassigned Shards“, *https://sematext.com*.
26. Elastic, „Index templates“, *https://www.elastic.co/guide/en/elasticsearch*, Teorija i primeri.
27. logz.io, „Elasticsearch Mapping: The Basics, Updates & Еxamples“, *https://logz.io/blog/elasticsearch-mapping*, Objašnjenja za data mapping sa primerima.
28. Elastic, „Aliases“, *https://www.elastic.co/guide/en/elasticsearch/reference/7.17*, Teorija i primeri.
29. Stackoverflow, „What are aliases in elasticsearch for?“, *https://stackoverflow.com*, Objašnjenje o korišćenju aliasa.
30. Elastic, „Writing queries“, *https://www.elastic.co/guide/en/elasticsearch/client/net-api/7.17*, Pisanje upita korišćenjem NEST biblioteke.
31. Opster Expert Team - Madhusudhan, „Elasticsearch Text Analyzers - Tokenizers, Standard Analyzers, Stopwords and more“, *https://opster.com/guides/elasticsearch*, Način rada Elasticsearch Analzyer mehanizma
32. Elastic, „Stemming“, *https://www.elastic.co/guide/en/elasticsearch*, Objašnjen način rada stemming procesa.
33. Elastic, „Anatomy of an analyzer“, *https://www.elastic.co/guide/en/elasticsearch*, Sturktura analyzer modula.
34. Elastic, „Writing analyzers“, *https://www.elastic.co/guide/en/elasticsearch*, Slika koja prikazuje strukturu analyzer modula.
35. Opster Team, „Elasticsearch Stop Words“, *https://opster.com/guides/elasticsearch*, Objašnjenje sa implementacijom za stop words.
36. Elastic, „Stop token filter“, *https://www.elastic.co/guide/en/elasticsearch*, Stop words token filter.
37. Elastic, „Search with synonyms“, *https://www.elastic.co/guide/en/elasticsearch*, Značaj i korišćenje synonyms kod pretrage.
38. Elastic, „Synonyms APIs“, *https://www.elastic.co/guide/en/elasticsearch*, Synonyms APIs kao način čuvanja synonyms sets.
39. Opster Team, „Understanding Elastcisearch Scoring and the Explain API“, *https://opster.com/guides/elasticsearch*, Način rada scoring mehanizma u Elasticsearch mehanizmu pretrage.
40. Elastic, „Practical BM25 - Part 2: The BM25 Algorithm and its Variables“, *https://www.elastic.co/blog*, Princip rada BM25 algoritma u okviru Elasticsearch mehanizma.

# Dodatak A

Ovaj deo imaju samo neki radovi, većina radova ih nema.

To su obično elementi koji su preveliki da se nađu u samom sadržaju rada. Na primer: predugačke tabele, prevelike slike, predugačak kod.

# Dodatak B

# Podaci o kandidatu

Kandidat Ime Prezime je rođen/a xxxx. godine u Gradu. Završio/la je srednju školu u Gradu, 20xx. godine. Fakultet Tehničkih Nauka u Novom Sadu je upisao/la 20xx. godine. Ispunio/ispunila je sve obaveze i položio/la je sve ispite predviđene studijskim programom sa prosečnom ocenom od xx.xx.

Ovo je generički tekst, može se menjati i može se postaviti fotografija.