Zbornik radova Fakulteta tehničkih nauka, Novi Sad

UDK: 004.738.1:766
DOI: https://doi.org/10.24867/28BE32Sukovic

SOFTVERSKI SISTEM ZA SEMANTIČKU PRETRAGU TEKSTA PRIMENOM VEKTORSKIH BAZA PODATAKA

SOFTWARE SYSTEM FOR SEMANTIC TEXT SEARCH USING VECTOR DATABASE TECHNOLOGY

Srđan Šuković, *Fakultet tehničkih nauka, Novi Sad*

Oblast – PRIMENJENE RAČUNARSKE NAUKE I INFORMATIKA

Kratak sadržaj – *Ovaj rad predlaže alternativni pristup pretrage teksta upotrebom vektorskih baza podataka i pretreniranih modela veštačke inteligencije. Rad obrađuje proces kreiranja vektorske reprezentacije tekstualnih podataka, indeksiranje istih u različite vektorske baze i postupak pretrage sličnosti sa upitima korisnika.*

Ključne reči: *Vektorska baza, vektorska reprezentacija, semantička pretraga.*

Abstract – *This paper proposes an alternative approach to text search using vector databases and pretrained artificial intelligence models. The paper deals with the process of creating a vector representation of textual data, indexing them into different vector databases and the procedure of searching for similarities with user queries.*

Keywords: *Vector database, vector embedding, semantic search.*

1. UVOD

Semantička pretraga teksta predstavlja metod pretrage koji se ne oslanja isključivo na tačne vrednosti ključnih reči, nego na razumevanje celokupnog konteksta i namere osobe koja pretražuje. Ova ideja se realizuje uz pomoć pretreniranih modela veštačke inteligencije koji imaju sposobnost razumevanja teksta u prirodnom jeziku, tako što na osnovu teksta kreiraju visoko-dimenzionalne vektorske reprezentacije koje uspešno u sebi nose semantičko značenje ulaznog teksta [1]. Vektorske baze, kao što su *Pinecone, Weaviate* i *Qdrant*, efikasno rešavaju problem čitanja, pisanja i upravljanja vektorima.

Ovaj rad objedinjuje mogućnosti koje nude modeli veštačke inteligencije i vektorske baze podataka - uz pomoć modela generišu se vektorske reprezentacije tekstualnog sadržaja, koje se dalje upisuju u vektorske baze. Odatle, na osnovu korisničkog upita, koji se takođe prevodi u vektorsku reprezentaciju, radi se pretraga sličnosti u odnosu na prethodno upisane vektore i korisnici dobijaju najbliskije rezultate njihovim upitima. Ovaj algoritam nije ograničen ni na jedan domen problema, i može poboljšati bilo koju funkcionalnost pretrage u već postojećim aplikacijama. Konkretno, u ovom radu, algoritam je implementiran na primeru pretrage filmova.

1.1 Skup podataka

Za implementaciju rešenja korišćen je skup podataka [2] koji sadrži preko 34 hiljade različitih prepričanih filmova, sa svojim opisnim podacima (naslov, godina, žanr, poreklo, režiseri, glumci, fabula, link do *Vikipedije*). S obzirom da skup podataka nije skroz prilagođen problemu koji ovaj rad rešava, neophodno je da indeksiranju prethodi pretprocesiranje podataka. Pretprocesiranje uključuje otklanjanje nepotrebnih informacija, nepotpunih sadržaja, duplikata i slično.

1.2 Pinecone

Pinecone [3] je specijalizovana baza podataka za upravljanje vektorima, dizajnirana da olakša rad sa velikim skupovima podataka i složenim zadacima pretraživanja u domenu veštačke inteligencije i mašinskog učenja. Kao usluga koja se temelji na konceptu vektorskog pretraživanja, *Pinecone* omogućava korisnicima da efikasno organizuju, indeksiraju i pretražuju podatke visoke dimenzionalnosti, što je posebno relevantno za aplikacije poput preporuka sadržaja, semantičkog pretraživanja, detekcije anomalija i personalizovanih iskustava korisnika.

Jedna od ključnih prednosti *Pinecone*-a je njegova sposobnost da precizno i brzo vrši pretraživanje sličnosti u velikim skupovima podataka. To se postiže korišćenjem sofisticiranih algoritama za vektorsko pretraživanje, koji omogućavaju brzo pronalaženje najrelevantnijih rezultata na osnovu sličnosti ugrađenih vektora. Ova sposobnost je posebno važna u primenama gde je potrebno obraditi velike količine tekstualnih, audio ili vizuelnih podataka i pronaći visoko relevantne informacije ili obrasce.

Platforma *Pinecone* omogućava skalabilnost i visoku dostupnost, što je čini pogodnom za upotrebu u produkciji. Njani alati i API-ji su dizajnirani da budu intuitivni i laki za korišćenje, što omogućava brzu integraciju u postojeće sisteme.

Takođe, *Pinecone* podržava integraciju sa popularnim okvirima za mašinsko učenje i obradu podataka, kao što su *TensorFlow* i *PyTorch*, olakšavajući razvoj i implementaciju složenih modela veštačke inteligencije.

NAPOMENA:
Ovaj rad proistekao je iz master rada čiji mentor je bio dr Dragan Ivanović, red. prof.

1.3 Weaviate

Weaviate [4] je napredna, platforma otvorenog koda za upravljanje bazom podataka koja se fokusira na kombinovanje modela grafova sa vektorskim pretraživanjem, pružajući tako jedinstvenu infrastrukturu za obradu i analizu podataka. Osnovana s idejom da unapredi pristup obradi složenih podataka, Weaviate se posebno ističe u domenima kao što su *NLP* (engl. *Natural Language Processing*), semantičko pretraživanje i *AI* (engl. *Artificial Intelligence*) aplikacije.

Jedinstvenost *Weaviate*-a leži u njegovoj hibridnoj strukturi koja spaja tradicionalno grafovsko skladištenje podataka sa naprednim metodama vektorskog pretraživanja. Ova kombinacija omogućava korisnicima da izvode kompleksne upite, analiziraju velike količine podataka i pronalaze relevantne informacije kroz sličnosti u sadržaju. *Weaviate* automatski transformiše podatke u vektorske reprezentacije, što olakšava precizno i kontekstualno pretraživanje.

Weaviate je dizajniran da bude visoko skalabilan i prilagodljiv, omogućavajući efikasnu upotrebu u različitim okruženjima i aplikacijama. Podržava raznolike tipove podataka, uključujući tekst, slike, i složene strukture podataka, što ga čini pogodnim za širok spektar upotreba. Integracija sa popularnim alatima za mašinsko učenje poput *TensorFlow* i *PyTorch* dalje proširuje njegovu primenljivost, čineći ga pogodnim za razvoj *AI* modela.

1.4 Qdrant

Qdrant [5] je moderna, otvorena platforma za upravljanje podacima, specijalizovana za efikasno vektorsko pretraživanje i analizu podataka. Razvijena kako bi zadovoljila rastuće potrebe za obradom složenih i velikih skupova podataka, posebno u domenima veštačke inteligencije i mašinskog učenja, *Qdrant* se ističe svojom sposobnošću da brzo i precizno obradi i indeksira podatke visoke dimenzionalnosti.

Ključna karakteristika *Qdrant*-a je njegova napredna podrška za vektorsko pretraživanje, omogućavajući korisnicima da izvrše kompleksne upite i dođu do relevantnih informacija u velikim bazama podataka. Ova funkcionalnost je posebno korisna u aplikacijama poput pretrage sličnih slika, personalizacije sadržaja, semantičkog pretraživanja i sistema preporuka, gde je važno brzo pronalaženje najrelevantnijih rezultata.

Qdrant je dizajniran s fokusom na visoku performansu i skalabilnost, što ga čini idealnim za upotrebu u zahtevnim poslovnim okruženjima i aplikacijama koje zahtevaju obradu velikih količina podataka. Njagova arhitektura omogućava efikasno upravljanje resursima i optimizaciju za brzo pretraživanje, što značajno smanjuje vreme potrebno za obradu upita.

1.5 Kreiranje vektorske reprezentacije od tekstualnog sadržaja

Za kreiranje vektora korišćeni su *HuggingFace* pretrenirani modeli veštačke inteligencije [6], specijalizovani za kreiranje vektorskog sadržaja. Model koji je pokazao najbolje performanse je *all-mpnet-base-v2*. Navedeni modeli su otvorenog koda i besplatni su za korišćenje.

1.6 Upis podataka u baze

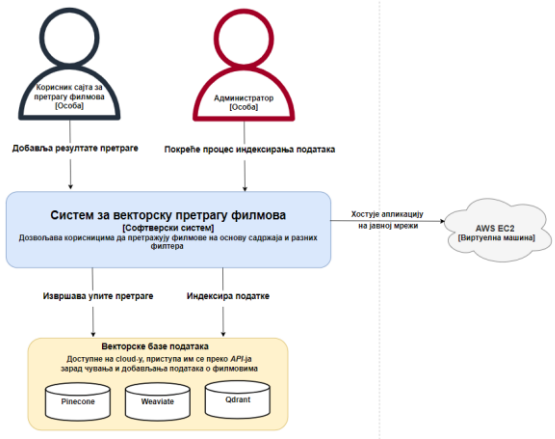
Za upis podataka u svaku od baza, potrebno je prilagoditi se API-ju koji one nude. Ono što je zajedničko za sve, jeste da se u jednom pozivu biblioteke vrši upis prethodno generisanog vektora i metapodataka koji se odnose na njega. U ovom slučaju to je vektor generisan od teksta prepričanog filma i samog naslova, dok su metapodaci godina, žanr, glumci, režiseri i poreklo. Baze koje su korišćene u radu su *Pinecone*, *Weaviate* i *Qdrant*.

1.7 Dobavljanje podataka

Proces dobavljanja podataka se sastoji iz 4 celine: vektorizacija upita, adaptacija korisničkog upita i filtera *SDK*-ju odgovarajuće baze, sam poziv klijenta baze i prilagođavanje odgovora korisničkom interfejsu aplikacije. Vektorizacija upita vrši se na isti način kao i prilikom indeksiranja, uz pomoć specijalizovanih modela veštačke inteligencije i zajednička je za sve baze.

2. SPECIFIKACIJA SISTEMA SEMANTIČKE PRETRAGE TEKSTA

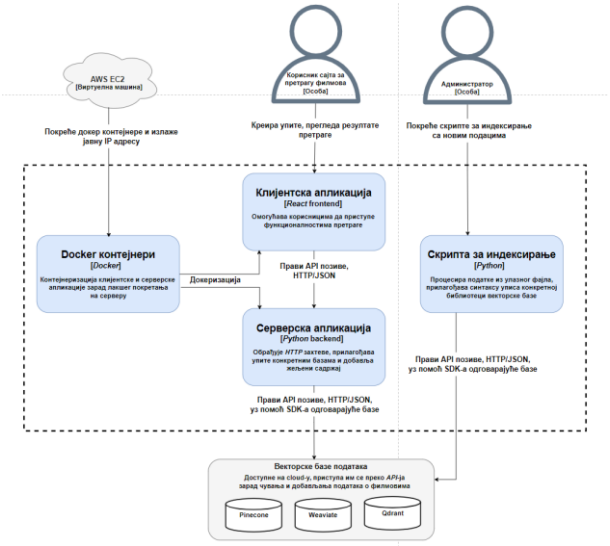
U ovom poglavlju prikazana je arhitektura sistema sa aspekta *C4* modela dijagramiranja, zasnovanom na apstrakcijama koje odražavaju kako arhitekte i programeri softvera razmišljaju i grade softver. Na slici 2.1 prikazan je konteksni dijagram sistema, koji prikazuje kako sistem interaguje sa eksternim entitetima, kao što su korisnici, administratori, vektorske baze i klaud provajderi (engl. *cloud providers*).



Slika 2.1 *Konteksni dijagram sistema za semantičku pretragu filmova*

Detaljniji prikaz komponenata sistema prikazan je na dijagramu kontejnera (slika 2.2), koji prikazuje osnovne kontejnere sistema i njihove međusobne odnose. Ovakvo softversko rešenje sastoji se iz dve glavne celine:

1. Deo za indeksiranje podataka – implementiran kroz *Python* skripte koje pokreće administrator sistema. Oslanja se na *Python* biblioteku *sentence-transformers* koja služi kao apstrakcija korišćenja raznih modela za kreiranje vektora od teksta.
2. Deo za opsluživanje podataka – *Flask* aplikacija koja obrađuje zahteve i vrši upite ka vektorskim bazama.



Slika 2.2 *Dijagram kontejnera sistema*

Pored prethodno pomenutih celina sistema, implementiran je i korisnički interfejs uz pomoć *React* biblioteke, koji komunicira sa delom za opsluživanje podataka korisnicima.

3. EKSPERIMENT

Zarad verifikacije rezultata izdvojen je podskup celog skupa podataka, od kojeg je sačinjena mapa gde je ključ naslov filma, a vrednost tekst prepričanog filma. Zatim, upotrebom *eugenesiow/bart-paraphrase* modela sa *HuggingFace* platforme, tekst je preformulisan tako da bude donekle izmenjen, ali sadrži isto značenje kao i pre primene modela.
Ovaj pristup je odabran kako bi se dočarala fleksibilnost i moć semantičke pretrage teksta, sa upitom čiji je tekst izmenjen u odnosu na prethodno indeksirane podatke, ali je pak značenje zadržano.
Zatim, svaka od vrednosti iz mape prolazi kroz isti tok kao kada aplikacija primi zahtev za pretragu: kreiranje vektorske reprezentacije i izvršavanje upita nad bazom. Najzad, najsličniji rezultat koji baza vrati, upoređuje se sa filmom koji je parafraziran iskorišćen za upit, i ukoliko dođe do podudaranja, pretraga se smatra uspešnom. Na osnovu broja pozitivnih i negativnih rezultata računa se metrika preciznosti.

4. REZULTATI I TUMAČENJA

Kvalitet odgovora na upit zavisi gotovo isključivo od modela koji je kreirao vektore i metrike sličnosti po kojoj se oni dobavljaju. Na osnovu analize performansi modela dokumentovane na *HuggingFace* platformi, za testiranje odabrani su sledeći modeli:

1. *all-mpnet-base-v2* – koji sveobuhvatno ima najbolje performanse što se kvaliteta vektora tiče
2. *all-MiniLM-L12-v2* – koji je tri puta manji i brži od prethodnog, međutim daje slabije performanse.

Kao metrika sličnosti, za oba modela odabrana je kosinusna sličnost, kao što je preporučeno u dokumentaciji. U tabeli 4.1 prikazani su broj pogodaka, broj promašaja i preciznost dobijeni prilikom testiranja sistema sa vektorima dobijenim od navedenih modela, na podskupu podataka od 66 elemenata. Rezultati se odnose na to da li se očekivani dokument dobija kao najrelevantniji ili ne.

Model	**Broj pogodaka**	**Broj promašaja**	**Preciznost**
all-mpnet-base-v2	62	4	0,939
all-MiniLM-L12-v2	55	11	0,833

Tabela 4.1 Performanse sistema za očekivani najrelevantniji rezultat

U slučaju kada se proverava da li se očekivani rezultat nalazi u top 4 rezultata dobijenih pretragom, dobijaju se sledeći rezultati, prikazani u tabeli 4.2.

Model	**Broj pogodaka**	**Broj promašaja**	**Preciznost**
all-mpnet-base-v2	66	0	1.0
all-MiniLM-L12-v2	64	2	0,969

Tabela 4.2 Performanse modela za očekivani top 4 rezultat

5. ZAKLJUČAK

U radu je predstavljen sistem za vektorsku pretragu filmova na osnovu njihovog sadržaja i dostupnih filtera. Ovakvo softversko rešenje omogućuje korisnicima preciznu i brzu pretragu filmova, gde je dovoljno da osoba svojim rečima napiše šta želi da se u filmu dešava, i sistem će dobaviti najrelevantnije rezultate naspram korisničkog upita, i time proces traženja filma svesti na minimum.

Rešenje je implementirano paralelnom upotrebom tri vektorske baze podataka - *Pinecone*, *Qdrant* i *Weaviate*, i dva pretrenirana modela za generisanje vektora od teksta - *all-mpnet-base-v2* i *all-MiniLM-L12-v2* sa *Huggingface* platforme. Korisnicima je na sajtu omogućen odabir baze koju žele da koriste za svoj upit, dok su u svaku od njih indeksirani vektori modela *all-mpnet-base-v2*, koji je pokazao preciznost od 100% na podskupu podataka za top 4 rezultata.

Odlične performanse vektorskih baza, čak i u sklopu besplatnog plana na veoma ograničenoj memoriji i računskoj snazi, i pomenutih modela koji su otvorenog pristupa, i mnogo manji u odnosu na komercijalne modele koje nudi OpenAI platforma, predstavljaju izuzetno dobru osnovu za dalju nadogradnju sistema.

5.1 Dalji razvoj sistema

Da bi se dostigao nivo komercijalnog proizvoda, neophodno je unaprediti postojeći sistem baš u ovom smeru. Pretplatom na produkcioni plan vektorskih baza obezbedila bi se veća brzina i računarska moć za veliki broj konkurentnih korisničkih sesija. Adaptacijom rešenja za proizvoljne skupove podataka određenog formata omogućilo bi neograničenu primenu u svakom polju industrije. Dalje, prelaskom na komercijalne modele za generisanje vektora obezbedilo bi višestruko bolju preciznost i zadovoljstvo korisnika. Radi poređenja, trenutno najperformantniji model *OpenAI* platforme, *text-embedding-3-large*, generiše vektore sa 3072 dimenzije, što je tačno 4 puta više od modela *all-MiniLM-L12-v2*, koji se u ovom radu pokazao najbolje. Shodno tome, vektorske reprezentacije indeksiranih podataka bile bi višestruko kvalitetnije i pretraga bi pokazivala još bolje rezultate.

6. LITERATURA

[1] S. J. Russel i P. Norvig, u *Artificial intelligence: a modern approach*, New Jersey, Pearson Education, 2010, pp. 1-29.

[2] J. R, „kaggle.com,“ [Na mreži]. Available: https://www.kaggle.com/datasets/jrobischon/wikipedia-movie-plots.

[3] pinecone.io, „https://www.pinecone.io/,“ pinecone.io. [Na mreži].

[4] „weaviate.io,“ Weaviate, [Na mreži]. Available: https://weaviate.io/developers/weaviate. [Poslednji pristup 16 February 2024].

[5] „qdrant.tech,“ Qdrant, [Na mreži]. Available: https://qdrant.tech/documentation/. [Poslednji pristup 16 February 2024].

[6] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue i A. Moi, „HuggingFace's Transformers: State-of-the-art Natural Language Processing,“ Hugging Face, Brooklyn, 2020.

Kratka biografija:

Srđan Šuković rođen je 11.01.2000. godine u Zrenjaninu. Smer računarstvo i automatika na Fakultetu tehničkih nauka u Novom Sadu upisao je 2018 godine. Osnovne studije završio je u septembru 2022. godine. Od oktobra 2022. godine upisuje master studije i počinje da radi kao softverski inženjer.