UniMatch

Tehnička dokumentacija

Verzija 1.0

Studentski tim: Karlo Bašić

Luka Alfirević

Ivan Džanija

Adam Šinjori

Rita Zonjić

Nastavnik: Mihaela Vranić

Sadržaj

1. Opis razvijenog proizvoda 3

1.1 Ideja i motivacija 3

1.2 Naše rješenje 3

1.3 Mogućnosti poboljšanja 3

1.4 Opis razmatranih problema i rješenja 3

1.4.1 Prvi upit – opis zahtjeva 4

1.4.2 Drugi upit – rastav zahtjeva 5

1.4.3 Treći upit – pregled generiranih podataka 6

1.4.4 Četvrti upit – dodatno pojednostavljenje zahtjeva 8

1.4.5 Završni koraci i analiza prikupljenih podataka 9

2. Tehničke značajke 14

2.1 Razvojni alati 14

2.2 Algoritmi 14

3. Upute za korištenje 15

4. Literatura 16

Tehnička dokumentacija

# Opis razvijenog proizvoda

## Ideja i motivacija

Ideja projekta bilo je razvijanje usluge za pretraživanje podataka o studiranju na raznim svjetskim sveučilištima.

S obzirom na sve izraženiju globalizaciju, potražnju i ponudu studiranja na sveučilištima u inozemstvu motivacija izrade ovog projekta olakšanje kompletnog pregleda mogućnosti studiranja na velikom broj svjetskih sveučilišta te učenje tehnologija s kojima možemo razviti ovakvo rješenje, ali i koristiti u budućim radovima.

## Naše rješenje

Razvili smo web aplikaciju UniMatch koja na intuitivan i jednostavan način prikazuje podatke koje bi korisniku mogle biti korisne. Usluga omogućuje razna filtriranja, spremanja prijašnjih pretraživanja i obilježavanje interesantnih sveučilišta.

## Mogućnosti poboljšanja

Mogućnosti poboljšanja našeg rješenja su razna, ali najočitija su:

* proširenje inicijalnog skupa podataka
* automatizacija ažuriranja informacija u skupu podataka
* poboljšanje arhitekture sustava.

## Opis razmatranih problema i rješenja

Pri izradi aplikacije susreli smo se s problemom teškoće pronalaska podataka te manjkom podataka. Uspjeli smo na različitima stranicama sa sveučilišnim informacijama i općenitim informacijama o državama i gradovima gdje se sveučilišta nalaze prikupiti, urediti i spojiti podatke 120 sveučilišta. Pri korisničkom filtriranju podataka često bi ostalo minimalno mogućih sveučilišta koji zadovoljavaju te filtere te smo odlučili pokušati na alternativni način prikupiti podatke kako bi prošili naš skup podataka. Ideja je bila provjeriti mogućnosti besplatnih i javno dostupnih velikih jezičnih modela pri generiranju i skupljanju podataka. Modele koje smo isprobali su: Claude i ChatGPT. Model Claude nam je odmah odgovorio kako ne može odraditi takav zadatak, ali model ChatGPT nam je odgovorio kako nam može pomoći. Ako bismo se uvjerili da nam dostavlja zadovoljavajuće točne podatke proširili bismo naš skup podataka s novim podatcima.

*Kako ćemo odrediti jesu li podatci koje nam je dostavio točni ili zadovoljavajuće približni?*

Ideja je bila prvo testirati kakve podatke će nam dostaviti na sveučilišta koja imamo u našem skupu podataka. Naravno ne znamo ni da su naši podatci potpuno točni, ali poprilično smo uvjereni u približnu točnost o obzirom na izvore na kojima smo prikupili podatke.

*Kako ukloniti mogućnost generiranja zastarjelih podataka?*

Dodatan problem o kojem smo razmatrali je mogućnost da je ChatGPT „treniran“ na skupu podataka koji bi mogao biti nekoliko godina star te možda trebamo uračunati vanjske faktore kao npr. inflacija, političke događaje i sl.

Kako bi izbjegli „narativnu ekonomsku analizu“ izbjegavali smo koristi ChatGPT model 3.5 i koristiti model 4.0 koji ima mogućnost trenutne pretrage interneta. Time smo djelomično smanjili mogućnost zastarjelosti podataka.

### Prvi upit – opis zahtjeva

Na početku smo pokušali iz modela ChatGPT odmah generirati cijelu tablicu tako što smo u upitu opisali sve potrebne parametre i države koje nas zanimaju.

I am working on an application in which I need accurate dataset containing universities with parameters that consist of:

* university (name of the university)
* collegeRank (global ranking of university)
* tuition (EUR/year) (how much is paid each year to study at the university, price needs to be in euros)
* percOfIntStud (percentage of international students based on all students)
* acceptance rate (how many people manage to get into the university in percentage)
* avgSafetyIndex (how secure is the country, how many crimes they have, and based on that index in the range from 0 to 100)
* cost of living (this consists of how much monthly persons pays on groceries, restaurants, transportation, and utilities excluding rent, price needs to be in euros)
* rent (monthly average for rent in euros)
* groceries (monthly average for groceries in euros)
* recreationCost (EUR/month) (monthly average for recreation and culture in euros)
* healthcare price (average out of the pocket spend on healthcare per year in euros) –
* avgMntTransportCost (average transportaion price in euros per month)
* link (link for the university)
* and each of the columns: Computer Science, Business, Economics, Psychology, Biology, Law, Medicine, Mathematics, Art, Physics they have values 1 or 0 based on the fact if university has these majors

The parameters avgSafetyIndex, cost of living, rent, groceries, recreationCost, healthcare price, avgMntTransportCost are based on the city in which university is located.

I want you to find me data for these countries: … (popis svih država)

For each country find the best university based on ranking and based on chosen university for each country write down other parameters for it.

Make me an excel file for this dataset.

1. **ODGOVOR**

Slika na kojoj se prikazuje tekst, Font, snimka zaslona, pismo

Opis je automatski generiran

U prvom odgovoru nije generirao popunjenu tablicu, već je umjesto toga naveo izvore na kojima bismo mogli pronaći tražene podatke. Ovaj pristup ukazuje na ograničenja u generiranju specifičnih i točnih podataka unutar samog modela, te nas je usmjerio na dodatne korake u procesu prikupljanja podataka s vanjskih izvora. Najviše se usredotočio na traženje izvora vezanih uz troškove života.

### Drugi upit – rastav zahtjeva

**Slika na kojoj se prikazuje tekst, snimka zaslona, Font, dokument

Opis je automatski generiran**

Slika 1.4.2‑1. Odgovor na zahtjev za kreiranjem skupa podataka

U drugom odgovoru nije popunio Excel tablicu s potrebnim podacima, već je poslao praznu tablicu koja sadrži samo parametre koje smo prethodno naveli. Ovo je jedan od standardnih problema modela ChatGPT pri generiranju datoteka.

**Slika na kojoj se prikazuje tekst, snimka zaslona, Font, dokument

Opis je automatski generiran**

Slika 1.4.2‑2. Odgovor na zahtjev za popunjavanjem skupa podataka

Ni nakon dodatnog upita ne šalje popunjenu Excel tablicu te ovaj put ne generira uopće Excel tablicu već samo govori da će započeti proces prikupljanja podataka i kako će pristupiti tom procesu. Pristup problemu je neočekivano detaljan i pravilan. Razdvojio je proces na više malih koraka i manjih iteracija.

Slika na kojoj se prikazuje tekst, snimka zaslona, Font

Opis je automatski generiran

Slika 1.4.2‑3. Odgovor na zahtjev za početak prikupljanja podataka

U odgovoru koji generira govori kako je počeo prikupljati podatke te predlaže slanje napretka u stalnim intervalima.

### Treći upit – pregled generiranih podataka

Slika na kojoj se prikazuje tekst, snimka zaslona, Font

Opis je automatski generiran

Slika 1.4.3‑1. Odgovor na upit trenutnog napretka

Nakon niza interakcija ChatGPT je poslao Excel tablicu pod nazivom *University\_Dataset\_Partial*, koja sadrži podatke za pet država: Australiju, Austriju, Kanadu, Njemačku i Japan. Iako su prethodno bili navedeni i linkovi za države poput Kostarike, Španjolske, Italije i Švicarske, na kraju se nije odlučio obraditi te zemlje, niti je slijedio abecedni redoslijed pri odabiru početnih država za analizu. Na odluku o izboru tih država nismo imali utjecaj.

Podaci u tablici bili su u velikoj mjeri slični našim, s izuzetkom parametra *safety index*, gdje su vrijednosti koje je ChatGPT generirao bile značajno veće od naših. Također, primijetili smo da su brojke koje je ChatGPT unosio bile zaokružene, na primjer, školarina za Sveučilište u Beču bila je 1500 eura, dok smo mi pronašli točnu vrijednost od 1453 eura. Unatoč tim odstupanjima, poveznice prema stranicama sveučilišta u tablici ispravno vode na odgovarajuće izvore, što je bio pozitivan rezultat.

Slika na kojoj se prikazuje tekst, snimka zaslona, Font, broj

Opis je automatski generiran

Slika 1.4.3‑2. Odgovori na upite za daljnji nastavak generiranja podataka

Kroz daljnje interakcije postalo je jasno da ChatGPT ima poteškoća u generiranju cijele tablice s podacima, jer izričito je naveo „I'll notify you“, što uopće nije njegova mogućnost.

Odlučili smo smanjiti veličinu zahtjeva. Nakon što smo smanjili zahtjev na 20 država. ChatGPT i dalje nije bio u mogućnosti odmah pružiti podatke za sve njih, već je poslao tablicu pod nazivom *University\_Dataset\_20\_Countries*. Tablica je u stvarnosti sadržavala podatke samo za 7 država: početnih 5 (Australija, Austrija, Kanada, Njemačka i Japan) te dodatne dvije (Ujedinjeno Kraljevstvo i Indija).

Ovaj nesklad između naziva tablice i stvarnog broja država ukazuje na ograničenja u kapacitetima generiranja podataka unutar trenutnih mogućnosti ChatGPT-a te zadane kompleksnosti zadatka.

### Četvrti upit – dodatno pojednostavljenje zahtjeva

Slika na kojoj se prikazuje tekst, snimka zaslona, Font

Opis je automatski generiran

Slika 1.4.4‑1. Odgovor na upit sa smanjenom kompleksnošću zadatka

Nakon što smo prepoznali da ChatGPT nije u mogućnosti generirati tablicu s velikim brojem sveučilišta, odlučili smo pokušati s tekstualnim odgovorima umjesto tablica. U ovom slučaju, model je odmah pružio traženih 5 sveučilišta s odgovarajućim parametrima, što je bio pozitivan pomak. Sljedeći korak bio je pokušaj proširenja broja sveučilišta po odgovoru na 20, no ChatGPT je i dalje vraćao samo 10 sveučilišta po odgovoru.

Ovaj pristup omogućio nam je da dođemo do svih traženih podataka, iako kroz dosta iteracija i ponovnih pokušaja, ali sigurno brže nego da smo sami tražili svaki podatak. Time smo uspješno prikupili i obradili podatke za sve države koje su nas interesirale.

Slika na kojoj se prikazuje tekst, snimka zaslona, Font, dokument

Opis je automatski generiran

Kada smo provjerili odakle ChatGPT izvlači podatke, ponovio je iste izvore kao i na početku konverzacije. Iako s određenom sumnjom, možemo pretpostaviti da su upravo ti izvori korišteni tijekom cijelog procesa prikupljanja podataka.

### Završni koraci i analiza prikupljenih podataka

U prikupljenoj tablici nedostaju države: Albania, Algeria, Argentina, Armenia, Azerbaijan, Bahamas te smo za njih morali poslati još jedan upit. Također u prikupljenoj tablici se nalaze i države koje mu nismo spomenuli: Yemen, Zambia. Ovaj podatak je teško objasniti s obzirom da ne znamo proces razmišljanja i sinteze informacija.

**Grafički prikazi i podatci mjere kvalitete modela linearne regresije:**

A graph with a red line and blue dots

AI-generated content may be incorrect.

Slika 1.4.5‑1. Grafički prikaz i mjere kvalitete modela linearne regresije za cijene školarina

A graph with a red line and blue dots

AI-generated content may be incorrect.

Slika 1.4.5‑2. Grafički prikaz i mjere kvalitete modela linearne regresije za postotak internacionalnih studenata

Slika na kojoj se prikazuje snimka zaslona, crta, dijagram, radnja

Opis je automatski generiran

Slika 1.4.5‑3. Grafički prikaz modela linearne regresije za postotak prihvaćenih prijava

A graph with blue and orange dots

AI-generated content may be incorrect.

Slika 1.4.5‑4. Grafički prikaz i mjere kvalitete modela linearne regresije za životne troškove

A graph with a red line and blue dots

AI-generated content may be incorrect.

Slika 1.4.5‑5. Grafički prikaz i mjere kvalitete modela linearne regresije za cijene mjesečnog najma

A graph with blue and orange dots

AI-generated content may be incorrect.

Slika 1.4.5‑6. Grafički prikaz i mjere kvalitete modela linearne regresije za cijene namirnica

A graph with blue and orange dots

AI-generated content may be incorrect.

Slika 1.4.5‑7. Grafički prikaz i mjere kvalitete modela linearne regresije za cijene rekreacije

A graph with a red line

AI-generated content may be incorrect.

Slika 1.4.5‑8. Grafički prikaz i mjere kvalitete modela linearne regresije za cijene zdravstvenog osiguranja i skrbi

A graph with a red line

AI-generated content may be incorrect.

Slika 1.4.5‑9. Grafički prikaz i mjere kvalitete modela linearne regresije za cijene prijevoza

Kod grafičkog prikaza linearne regresije i vrijednosti regresora te R2 mjere ciljamo na vrijednost što bliže 1. Vrijednost R2 mjere nam govori koliko varijance u podatcima naš model može objasniti i to u ovom kontekstu nije intuitivno interpretirati, ali vidimo kako imamo jedan koeficijent jednostavne linearne regresije. Za taj koeficijent bi očekivali da je što bliže 1 jer nam označava da kada se vrijednost podatka u našem skupu podataka promijeni za neku vrijednost za istu toliku vrijednost bi se trebao promijeniti i podataka u skupu podataka koje je generirao ChatGPT.

S obzirom na kompleksnost prikupljanja podataka bili smo zadovoljni ovim rezultatima i nastavili s generiranjem dodatnih podataka kako bi napunili naš skup podataka. Također uz prikupljanje, obradu i vizualizaciju podataka htjeli smo naučiti i razviti aplikaciju koja bi mogla te podatke na smislen i intuitivan način prikazati tako da nam 100%-tna točnost nije bila toliko presudna za daljnji rad, već sama količina podataka.

**VIZUALIZACIJE**

A map of the world

AI-generated content may be incorrect.

Slika 1.4.5‑10. "HEATMAP" po indeksu sigurnosti države

A map of the world

AI-generated content may be incorrect.

Slika 1.4.5‑11. "HEATMAP" po prosječnom ranku sveučilišta u državi

A map of the world

AI-generated content may be incorrect.

Slika 1.4.5‑12. "HEATMAP" po indeksu životnih troškova države

# Tehničke značajke

## Razvojni alati

Za razvoj web aplikacije sa korisničke strane korišten je okvir besplatne programske podrške [Angular](https://angular.dev/) temeljen na [TypeScript](https://www.typescriptlang.org/) platformi.

Za razvoj poslužiteljske strane korišten je besplatni okvir besplatne programske podrške [Django](https://www.djangoproject.com/) temeljen na platformi [Python](https://www.python.org/).

Za razvoj i spajanje podataka koji pružaju informacije na web aplikaciji korišten je [Jupyter](https://jupyter.org/) i ostala potrebna proširenja platforme [Python](https://www.python.org/).

Za kolaboraciju i verzioniranje korišteni su alati: [Git](https://git-scm.com/) – osnovni alat za verzioniranje i praćenje napretka i promjena u kodu i [GitHub](https://github.com/) – proširenje i platforma za javno postavljenje i kolaboriranje na repozitoriju.

Za izradu sustava primarno je korišteno razvojno okruženje [VSCode](https://code.visualstudio.com/) koje pruža velike mogućnosti uređivanja teksta te brzog testiranja programa.

## Algoritmi

Algoritam filtriranja je izrazito jednostavan prolazak po cijelom skupu podataka i spremanja samo redaka koji u svim stupcima imaju vrijednost veću od donje granice i manju od gornje granice. Pošto imamo konstantan broj stupaca možemo reći da je kompleksnost algoritma linearna - O(N).

Algoritam rangiranja imaju složenost – O(N \* log N) zbog algoritma sortiranja. Svaki podatak prvo pomnožimo sa težinom koju je korisnik odredio time što je odabrao koliko mu je taj faktor bitan u rangiranju, a zatim sortiramo te umnoške kako bi dobili rangiranje.

# Upute za korištenje

# Literatura

1. Walpole, Ronald E., Myers, Raymond H., Myers, Sharon L., Ye, Keying. „*Probability & statistics for engineers & scientists“*. 9th ed. Global ed. Tokyo: Pearson, 2016.