SVEUČILIŠTE JOSIPA JURJA STROSSMAYERA U OSIJEKU
FAKULTET PRIMIJENJENE MATEMATIKE I INFORMATIKE
Sveučilišni prijediplomski studij Matematika i računarstvo

Dodavanje eksternog znanja u LLM: MathosGPT

Dokumentacija

Matej Kolak, Domagoj Žiroš

Osijek, 2023.

Sadržaj

Uvod

ChatGPT i konverzacijski agenti

Konverzacijski agenti, poput **ChatGPT**, predstavljaju izvanrednu tehnološku inovaciju koja omogućuje interakciju između ljudi i računala putem prirodnog jezika. ChatGPT je napredan model temeljen na dubokom učenju koji je treniran na ogromnom skupu podataka. Ovaj konverzacijski agent je sposoban za razumijevanje i generiranje ljudskog jezika, omogućujući korisnicima postavljanje pitanja, dobivanje informacija i komunikaciju na prirodan način. Međutim, iako je ChatGPT impresivan, postoje određene limitacije koje mogu otežati zadovoljavanje specifičnih potreba korisnika.

Limitacije konverzacijskih agenata

1. ***Ograničeno znanje svijeta***
   ChatGPT ima ograničeno znanje koje je stečeno tijekom treninga i nije ažurirano sa svježim informacijama o stvarnom svijetu.
2. ***Potreba za specifičnim informacijama***
   Korisnici često traže specifične i kontekstualne informacije koje nisu dostupne standardnim konverzacijskim agentima.

MathosGPT

MathosGPT nastaje kao rješenje ovih izazova tako što omogućava razgovor o temama vezanim uz **Fakultet primijenjene matematike i informatike u Osijeku**.

Ovaj dokument će vas voditi kroz MathosGPT aplikaciju, objasniti kako funkcionira, kako je koristiti i kako rješava izazove s kojima se suočavaju konvencionalni konverzacijski agenti.

Pregled aplikacije

Aplikacija MathosGPT proširuje osnovni ChatGPT i obogaćuje ga dodatnim informacijama o Fakultetu primijenjene matematike i informatike u Osijeku. MathosGPT posjeduje znanje u **šest kategorija**, omogućujući korisnicima brz pristup specifičnim informacijama o fakultetu. Navedene kategorije su:

1. **Fakultet** - opće informacije o fakultetu te radu i misiji fakulteta
2. **Kadrovi** - informacije o zaposlenicima fakulteta što uključuje nastavnike, suradnike te administrativno i pomoćno osoblje
3. **Studenti** - opće informacije vezane uz studiranje, studentske aktivnosti te podršku za studente
4. **Upisi** - informacije za buduće studente
5. **Studiji** - informacije o studijima koji se izvode na fakultetu
6. **Kolegiji** - informacije o pojedinim kolegijima koji se izvode na fakultetu

Ciljni korisnici ove aplikaciji su studenti, budući studenti, osobe koje su na bilo koji način povezane s fakultetom te opća javnost.

Arhitektura aplikacije

Aplikacija se sastoji od **klijentskog** i **poslužiteljskog** dijela.

Klijentski dio aplikacije izrađen je u **JavaScriptu** uz upotrebu **Angular frameworka**. Ovaj dio sadrži sučelje koje omogućuje korisnicima postavljanje pitanja i interakciju s konverzacijskim agentom.

Poslužiteljski dio aplikacije izrađen je u **Pythonu** uz pomoć **Flask frameworka**. Poslužitelj obavlja zadatke stvaranja novih razgovora s MathosGPT asistentom i nastavka postojećih razgovora.

Aplikacija sadrži vlastitu bazu znanja o Fakultetu primijenjene matematike i informatike. Baza je sastavljena od tekstualnih datoteka prikupljenih tehnikama web scrapinga uz pomoć Python modula **bs4** i **requests**. Na

osnovu tih znanja korisnici su u mogućnosti razgovarati o temama vezanim uz fakultet.

Važno je napomenuti da iako aplikacija korisnicima pruža relevantne informacije, sam ChatGPT nije “naučio” ove podatke, već mu je prilikom svakog upita proslijeđen relevantan skup znanja iz baze podataka na osnovu kojih ChatGPT može odgovoriti na pitanja korisnika.

Taj postupak odvija se uz pomoć Python modula **llama_index**. Llama_index sadrži **klase za pohranjivanje dokumenata** iz baze nad kojima se kasnije vrše upiti. **Dokumenti** se ovisno o veličini razdvajaju u manje **čvorove** koji se zatim pohranjuju u jednu od mogućih klasa, ovisno o načinu na koji ih želimo pohraniti. Moguće ih je pohraniti u polje, vezanu listu, stablo, graf ili neku složeniju strukturu. Ova aplikacija koristi pohranu dokumenata u polje. Jednom kad imamo klasu za pohranu dokumenata generiramo njezinu odgovarajuću **engine klasu**. Engine klasa moze biti **query_engine**, **chat_engine** ili **agent**. Jednom kada korisnik pošalje upit, engine klasa dohvaća dokumente relevantne za taj upit i zatim ih prosljeđuje ChatGPT-ju zajedno sa upitom. ChatGPT zatim vraća odgovor na upit koji na kraju naša aplikacija prikazuje korisniku. S obzirom da je aplikacija strukturirana kao razgovor a ChatGPT asistentom, koristi se chat_engine.

Jedna od trenutnih limitacija aplikacije je da su i njezini podaci fiksni, odnosno njezino znanje seže do određenog trenutka u vremenu. Svaka daljna nadopuna znanja mora se izvršiti ručno.

Instalacija i pokretanje

Tehnički zahtjevi

Prije same instalacije potrebno je osigurati da operativni sustav korisnika zadovoljava tehničke zahtjeve koji uključuju:

Instaliran Python
OpenAI API ključ

Ostali tehnički zahtjevi se ostvaruju u toku instalacije:

Instalirani Python moduli OpenAI i Flask

Instalacija

Instalacija se provodi prema sljedećim koracima:

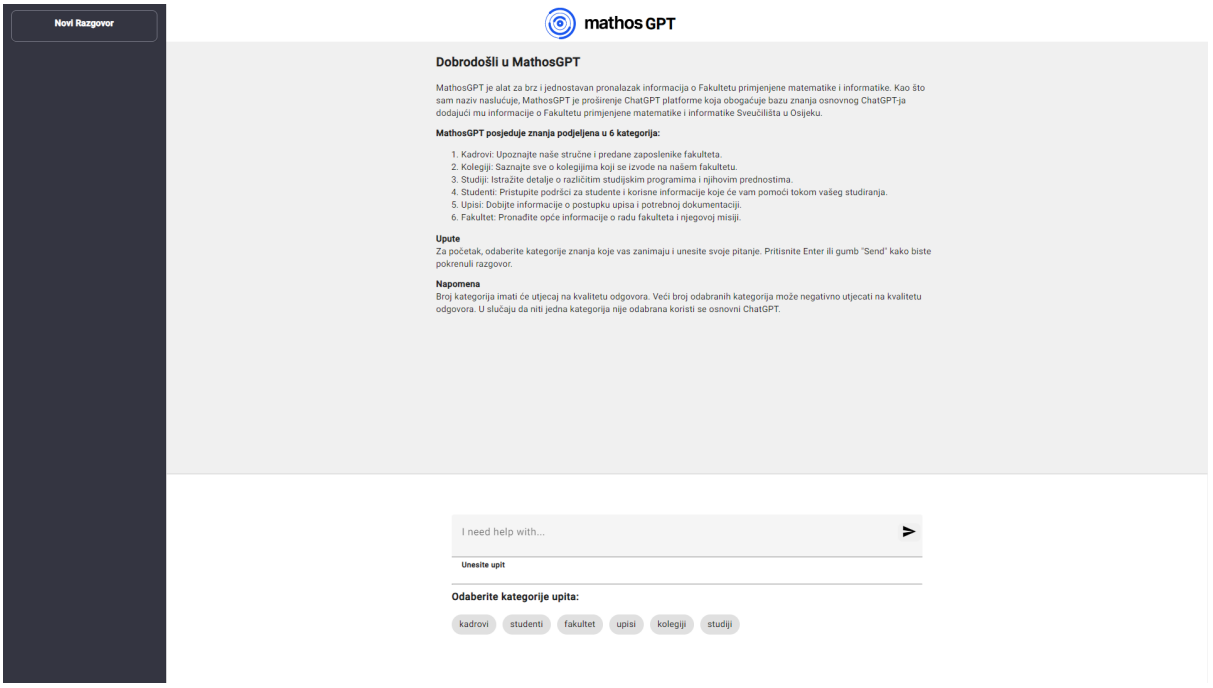
1. Preuzeti aplikaciju sa linka MathosGPT kao zip ili koristeći naredbu `git clone`
2. Otvoriti naredbeni redak u mapi gdje je spremljen kod aplikacije
3. Pokrenuti naredbu `pip install -r requirements.txt` kako bi se instalirali potrebni moduli
4. Unijeti svoj OpenAI API ključ na odgovarajuće mjesto u datoteci `./indices/indices.py`

Pokretanje

Jednom kada su ispunjeni svi tehnički zahtjevi te dovršena instalacija u naredbeni redak upisujemo naredbu `python app.py`. Tom naredbom pokreće se Flask server na adresi **`http://127.0.0.1:5000`** koju zatim otvaramo u pregledniku te tada možemo započeti sa radom.

Korištenje

Jednom kada je aplikacija pokrenuta, korisnik se susreće sa početnom stranicom. Korisničko sučelje osmišljeno je tako da bude jednostavno za sve vrste korisnika.

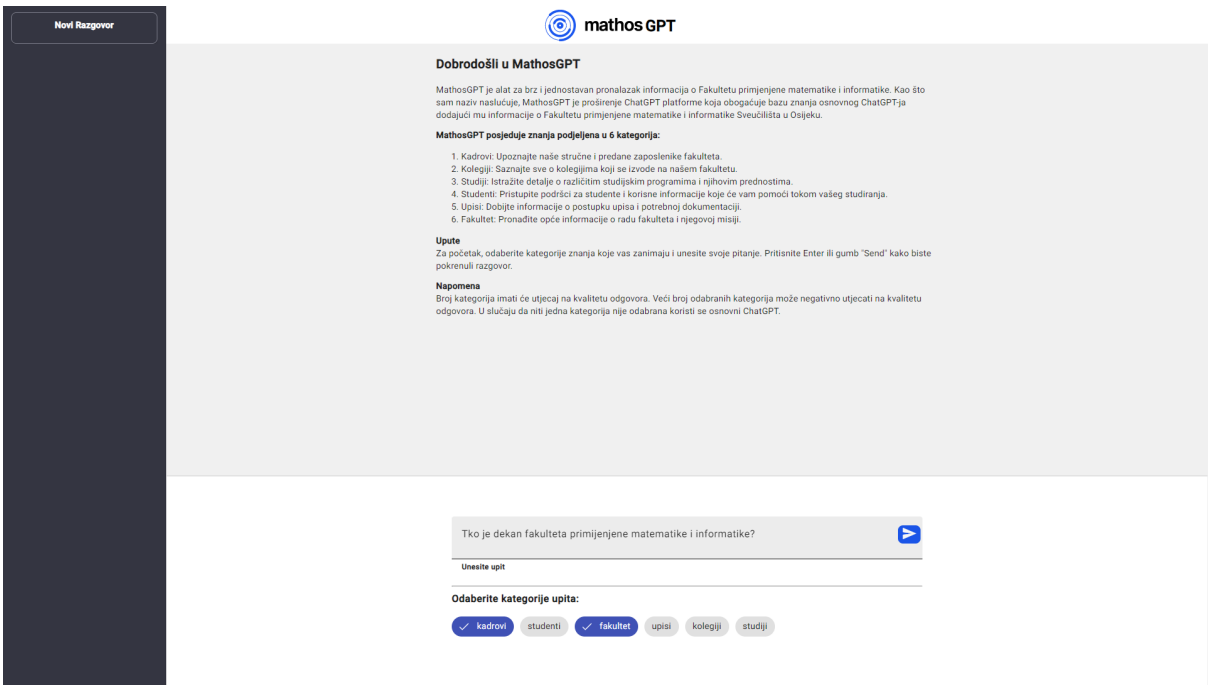


Slika 1. *Početna stranica aplikacije*

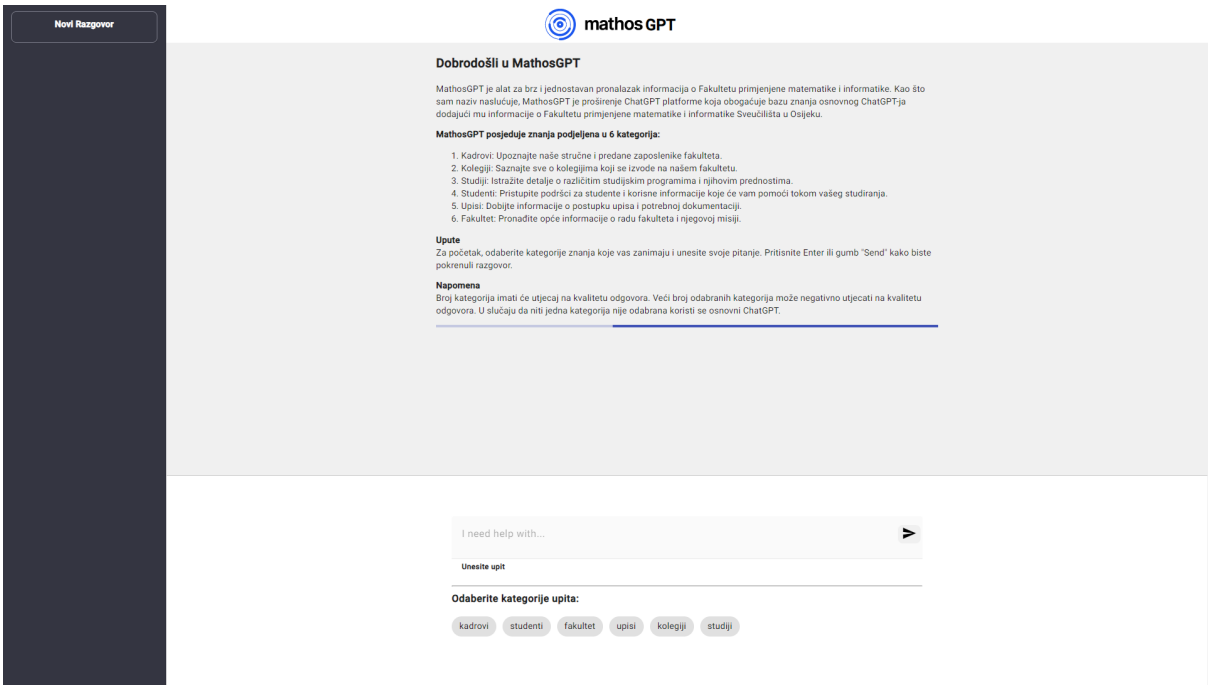
Početna stranica sadrži kraću verziju uputa za korisnike te naputak za uspješnu pretragu informacija. Za početak rada korisnik treba:

1. **Unijeti upit** u odgovarajuću traku za unos teksta koja se nalazi u donjem dijelu stranice.
2. **Odabrati jednu ili više kategorija** znanja na osnovu kojih bi trebao dobiti zadovoljavajući odgovor na svoje pitanje.
   Napomena: Jednom kad su kategorije odabrane i upit poslan u daljni razgovor neće biti moguće uključiti dodatne kategorije.
3. Jednom kada su ta dva koraka odrađena korisnik treba **pritisnuti tipku enter** unutar trake za unos ili **gumb send** koji se nalazi u traci za unos kako bi se upit poslao.

Nakon što je upit poslan pojaviti će se traka za povratnu informaciju koja će prikazivati animaciju učitavanja sve dok poslužitelj ne vrati odgovor na upit. Traka će se nalaziti odmah ispod teksta uputa na početnoj stranici.

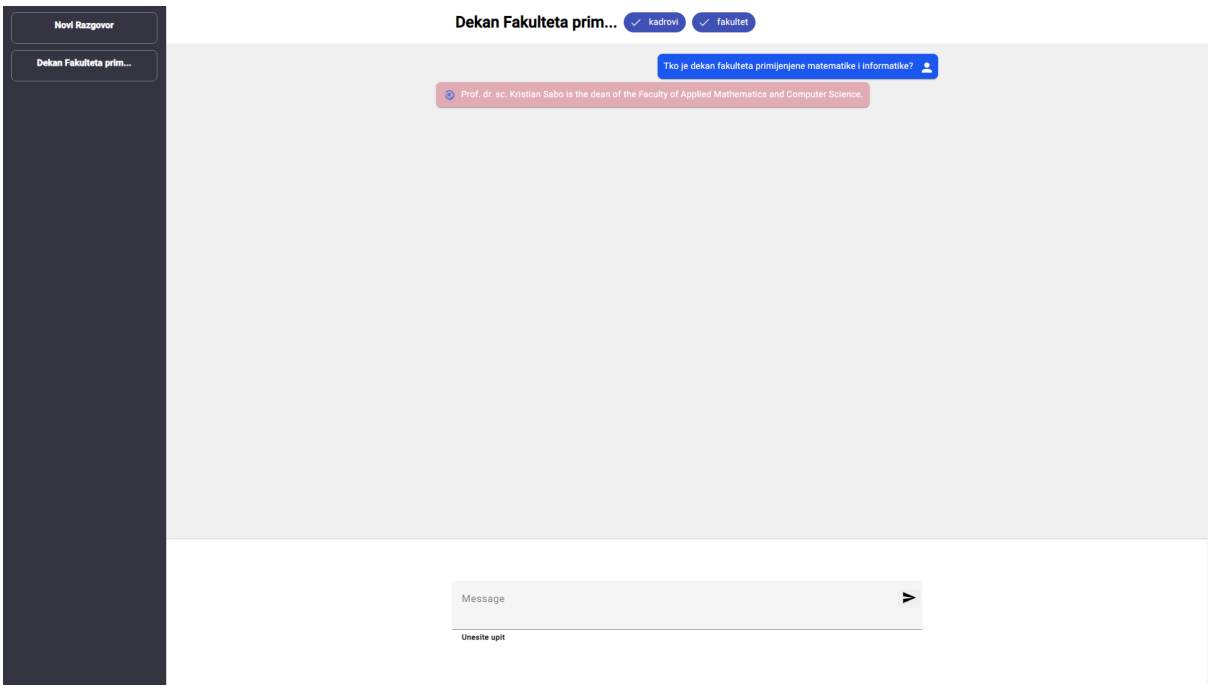


Slika 2. *Unos upita i odabir kategorija*



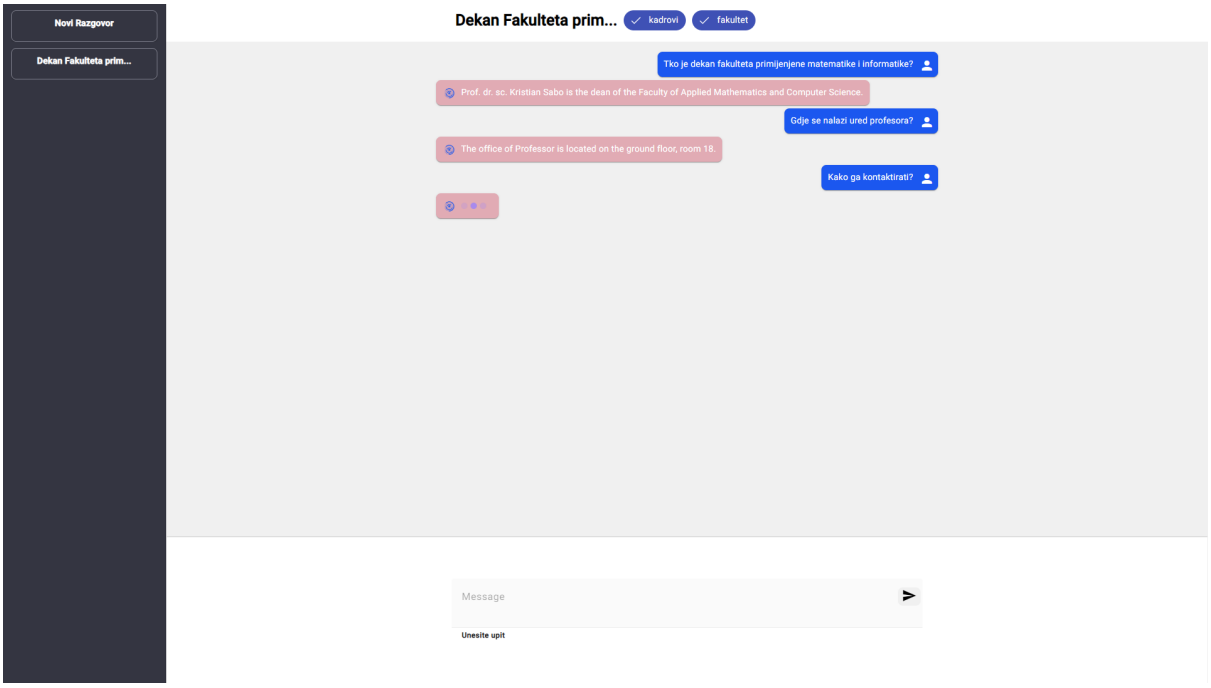
Slika 3. *Pojava trake za povratnu informaciju*

Jednom kad poslužitelj vrati odgovor, aplikacija prelazi na stranicu sučelja za razgovor u kojem će odgovor biti prikazan i biti će moguće nastaviti razgovor.



Slika 4. *Stranica sučelja za razgovor*

Prelaskom na stranicu sučelja za razgovor korisnik pri vrhu ima priliku vidjeti naziv razgovora generiran od strane ChatGPT-ja, odabrane kategorije, poruke razgovora, a na dnu sučelja nalazi se traka za sljedeće upite koje korisnik šalje. U navigacijskoj traci sa lijeve strane pojavljuje se i gumb sa nazivom trenutnog razgovora kojim se uvijek ponovno možemo vratiti u razgovor u kojem se trenutno korisnik nalazi.



Slika 5. *Nastavak razgovora*

Prilikom nastavka razgovora, na strani ChatGPT asistenta pojavljuje se animacija koja označava da odgovor na posljednje pitanje tek treba doći.

Valja napomenuti da kada korisnik vodi razgovor u koji su uključene određene kategorije, on i dalje može pitati pitanja na koja bi i osnovni ChatGPT mogao odgovoriti. Pod tim se podrazumijevaju bilo koja pitanja koja nemaju veze sa fakultetom.

Ukoliko korisnik želi započeti novi razgovor, uvijek se može vratiti na početno sučelje pritiskom na gumb **novi razgovor**, a isto tako se može vratiti i na postojeći razgovor pritiskom na gumb **sa nazivom razgovora.**

Napomena

Odabir kategorija važna je stavka za dobivanje zadovoljavajućeg odgovora. Ako korisnik ne označi niti jednu kategoriju prilikom upita, on će tada razgovarati sa osnovnim ChatGPT asistentom koji nema pristup znanjima o fakultetu. Korisnik također može odabrati i sve kategorije. Naputak je da odabir kategorije bude što precizniji, jer odabir većeg broja kategorija povećava šanse da će:

1. *Chat engine dohvatiti krive i/ili suvišne dokumente*
2. *ChatGPT zbog velikog broja dohvaćenih informacija dati krivi odgovor*

Daljnja unapređenja

Aplikacija je trenutno u funkcionalnoj verziji ali je daleko od savršenog. Postoji više dijelova koje bi trebalo usavršiti i proširenja koja bi valjalo implementirati za bolje korisničko iskustvo.

U ovom trenutku zabilježena su:

- Reorganizacija dokumenata za unapređenje dobivenih odgovora
- Mogućnost automatiziranog periodičnog ažuriranja baze znanja
- Mogućnost unošenja vlastitih korisničkih dokumenata
- Unapređenje korisničkog sučelja