

Poročilo tretjega projekta pri IEPS

Jan Novak Domijan
Urša Kumelj
Manca Drašček
UL FRI

Povzetek

Že narejen sistem za podobnostno iskanje smo povezali z jezikovnim modelom in ustvarili preprost spletni vmesnik za postavljanje vprašanj. Za jezikovni model smo izbrali model Gemma 3 s kontekstom treh najpodobnejših receptov. Našo implementacijo smo testirali na vnaprej določeni množici vprašanj in povzeli prednosti in slabosti modela vidnih iz danih odgovorov.

1 Uvod

Zadnji korak v projektu je integracija vektorskega iskanja in jezikovnega modela v en sam enoten RAG sistem. Ključni odločitvi v implementaciji sta bili število receptov vključenih v kontekst poizvedbe in izbira jezikovnega modela. Preizkusili smo izbran nabor jezikovnih modelov skupaj z različnimi velikostmi konteksta. V tem poročilu bomo predstavili rezultate naše analize skupaj s kratkim povzetkom dosedanje implementacije.

2 Povzetek implementacije

Ekstrakcija in predelava podatkov ostajata povsem enaka kot pri preteklih nalogah. Iz spletnih strani so izluščeni ključni podatki, ki so shranjeni v en sam segment za vsako stran. Nato na podlagi segmentov za vsak recept izračunamo vektorsko vložitev z modelom LaBSE.

Za interakcijo med sistemom in uporabnikom smo uporabili preprosto spletni vmesnik, ki omogoča uporabniku, da vpiše vprašanje in prejme odgovor z ali brez pomoči podobnostnega iskanja. Ko uporabnik naredi poizvedbo se ta pretvori v vektorsko obliko. Izračunan vektor nato uporabimo, da v podatkovni bazi najdemo k najpodobnejših receptov, kjer je k prost parameter implementacije. Poizvedbo nato podamo jezikovnemu modelu skupaj s podobnimi recepti za kontekst. V primeru, da uporabnik označi, da ne želi uporabiti RAG sistema, se poizvedba poda jezikovnemu modelu brez konteksta.

3 Izbira modela

Izbirali smo med šestimi različnimi jezikovnimi modeli: llama3.2 (1.2B), gemma3 (4.3B), deepseek-r1 (7.6B), qwen3 (8.2B), phi4 (14.7B) in GaMS (9.2B). Za vsak model smo najprej analizirali strukturo odgovorov: če so odgovori smiselni, če so slovnično pravilni in če so v pravem jeziku. Naštejemo par kratkih primerov vprašanj in odgovorov, ki prikazujejo pomankljivosti v strukturi odgovorov za različne modele:

- **phi4**

Vprašanje: Katere sestavine potrebuješ za arašidovčke?

Odgovor brez konteksta:

Osnovne sestavine za arašidovče vključujejo arašidi, kislo mleko ali kislo jogurt, soda (količica), sol in kruh ali semena

za posipanje. Lahko dodatno uporabiš tudi kislo mleko, mast ali olje ter slatkilo.

Odgovor je slovnično nesmislen.

- qwen3

Vprašanje: *Mi lahko svetuješ kaj pripraviti za praznike?*

Odgovor brez konteksta:

Prpriavi tradicionalno praznično hrano (npr. pizzico, štru-
klji, čokoladne zapečene zapečene zadebljane zapečene za-
debljane zapečene zadebljane zapečene zadebljane zape-
čene zadebljane zapečene zadebljane zapečene zadebljane
zapečene zadebljane zapečene zadebljane zapečene zade-
bljane zapečene zadebljane zapečene zadebljane zapečene
zadebljane ...

Vprašanje: *Mi lahko svetuješ kaj pripraviti za praznike?*

Odgovor brez konteksta:

[illegible]

Vprašanje: *Kako bi naredili bavarsko kremo po korakih?*

Odgovor brez konteksta:

[illegible]

Ta model je imel pogoste probleme z ponavljanjem žetonov.

- **deepseek-r1**

Vprašanje: *Potrebujem recept za arašidovčke.*

Odgovor brez konteksta:

Ako tražite receptore za arašidovke, možete pretražiti na Hrvatskoj e-biznis-u na stranici [Strana sa primierom](#).

Ovaj primjer će vam pomoći pronaći odgovarajuću znu jedinku.

Vprašanje: *Kako se naredi jabolčni kompot?*

Odgovor brez konteksta:

Strukturna formula je $C_6H_{10}O_7$.

Model je imel največje probleme z razlikovanjem med jeziki. Ne samo, da ni podal nobenega odgovora v slovenščini, ampak je celo menjaval abecedo znotraj besede. Veliko odgovorov je bilo popolnoma nepovezanih z vprašanjem.

• llama3.2

Vprašanje: *Kako se naredi jabolčni kompot?*

Odgovor brez konteksta:

Jabolčni kompot je zbirica jahovine, kremi, sirve in

Odgovor s kontekstom:

{answer}

Vprašanje: *Potrebujem recept, ki traja največ eno uro.*

Odgovor brez konteksta:

Recept, ki traja največ eno uro, je: [Recept za 1-2 urošnje].

Odgovor s kontekstom:

{answer}

Model ponekod sploh ni vrnil odgovora. V primeru, da je vrnil odgovor je ta bil večinoma v hrvaščini ali angleščini.

• GaMS

Odgovor brez konteksta:

Seveda, tukaj je recept:

Sestavine: - 200 g mletih arašidov - 150 ml vode - 1 žlica medu ali javorjevega sirupa (po želji) - Ščepec soli

Postopek: 1. V mešalniku zmešajte mlete arašide, vodo in med ali javorjev sirup. Mešajte na visoki hitrosti približno 2 minuti, dokler ne dobite gladkega in kremastega nadeva. 2. Dodajte ščepec soli za poudarjanje okusa. 3. Po potrebi dodajte več vode ali medu, da dosežete željeno konsistenco in sladkost.

Uživajte v zdravih arašidovčkih!

Odgovor s kontekstom:

[1]

Model večkrat ni vrnil odgovora. Veliko odgovorov je tudi bilo vsebinsko napačnih. Kljub pomankljivostim ta model ni imel večjih slovničnih napak za razliko od prej naštetih modelov.

Nato smo se osredotočili na modela GaMS in gemma3 saj nista imeli večjih pomankljivosti v gramatični pravilnosti slovenskih odgovorov. Za oba modela smo prešteli število vsebinsko pravilnih odgovorov: model GaMS je vrnil 33 pravilnih odgovorov od 80 in

model gemma3 42 pravilnih odgovorov od 80. Na podlagi rezultatov smo izbrali model gemma3.

4 Evalvacija

V ekstrakciji in obdelavi podatkov nismo opazili nobenih možnih izboljšav. Recepti so večinoma kratka besedila s podobno strukturo kar pomeni, da je težko utemeljiti kakršenkoli pristop, ki bi še nadaljnje razčlenil besedilo v več segmentov. Posledično tudi nismo imeli dodatnih orodij za zmanjšanje pomankljivosti podobnostnega iskanja. Vendar smo lahko pri tej nalogi pomankljivosti iskanja zmanjšali s povečanjem števila iskanih dokumentov. Idealno bi želeli jezikovnemu modelu podati čimveč receptov dokler ostanemo znotraj dovoljene velikosti podanega konteksta. Naša edina skrb je bila, da bi se jezikovni model preveč posvetil receptom, ki niso bili relevantni glede na poizvebo. Preizkusili smo 4 različne vrednosti zajetih podobnih dokumentov: 3, 10, 100 in 1000. Tukaj podamo en primer poizvedbe skupaj z odgovori z uporabo različnih velikosti konteksta. V vsakem primeru je bil iskan recept vsebovan v kontekstu.

Vprašanje: *Podaj mi recept za arašidovčke.*

Odgovori:

- $k = 3$:

Recept za arašidovčke:

Sestavine: 350 gramov arašidovega masla (moder pokrovček iz interspara) 100 gramov mletih mandljev 200 gramov rjavega sladkorja 300 gramov pirine moke 2 jajci 120 gramov puste skute 1 navaden jogurt lca

Postopek: Sestavine dobro prengnetemo v testo, nato jih poljubno oblikujemo ter pečemo 11 minut na 200 stopinj.

Jezikovnemu modelu so bili podani trije recepti, dva za arašidove piškote in en za jabolčno pito, in je uspešno razbral recept za arašidove piškote.

- $k = 10$:

I couldn't find a specific recipe for peanut butter cookies within the provided text snippets. However, I found recipes for Oreo cookies, Oreo rice, and basic muffins that could be adapted. Here's a combined approach based on the Oreo rice recipe, which seems the closest match:

Adapted Oreo Rice Recipe for Peanut Butter Cookies

This recipe uses the core principles of the Oreo rice, adapting it for a cookie texture.

Ingredients:

- 1.5 cups (190g) all-purpose flour
- 1/2 cup (65g) unsweetened cocoa powder (for a richer flavor)
- 1 1/4 cups (250g) granulated sugar
- 3/4 teaspoon baking soda
- 1/2 teaspoon salt
- 2 large eggs
- 1/2 cup (110g) melted butter
- 3/4 cup (185g) hot water
- Approximately 15-20 Oreo cookies (whole or coarsely crushed - adjust to your preference)

Instructions:

1. Preheat oven to 180°C (350°F). Line a baking sheet with parchment paper.
2. In a large bowl, whisk together the flour, cocoa powder, sugar, baking soda, and salt.
3. Add the melted butter and hot water to the dry ingredients and mix until just combined. Don't overmix.
4. Fold in the whole or coarsely crushed Oreo cookies.
5. Drop by rounded tablespoons onto the prepared baking sheet.
6. Bake for 16-18 minutes, or until the edges are golden brown.
7. Let cool on the baking sheet for a few minutes before transferring to a wire rack to cool completely.

Notes:

- Adjust the amount of Oreo cookies to your desired level of cookie flavor and texture.
 - For a smoother cookie, crush the Oreos finely. For a chunkier cookie, use larger pieces.
- I apologize that I couldn't provide a dedicated peanut butter cookie recipe from the given text. This adaptation combines elements from the available snippets to create a workable recipe.

Jezikovni model ni uspel razbrati recepta za arašidove piškote in je podal drug recept. Ta recept bi naj bil prilagojen iz recepta "Mlečni riž z Oreo piškoti" vendar besedilo recepta tega ne podpira. Model je vzel sestavine iz recepta "Oreo muffini" z razliko, da je zamenjal mleko za maslo. Postopek pa ne sovпада z nobenim od podanih receptov in je najverjetneje izmišljen.

- $k = 100$:

I am sorry, but none of the provided recipes contain a recipe for "arašidovčki" (hazelnut cookies). The recipes listed are for various other baked goods, such as palacinke (palatinke), rulade (rulade), and mandljevi piškoti (hazelnut cookies).

Jezikovni model ni uspel najti iskanega recepta in je našel par receptov, ki so bili vsebovani v kontekstu.

- $k = 1000$:

Okay, here's a classic and delicious recipe for Peanut Butter Cookies: {recept}.

Jezikovni model ni uspel najti iskanega recepta in je podal svoj recept, ki ni prisoten v kontekstu.

Opazimo, da je bila poizvedba uspešna le za relativno majhen kontekst. Za vrednosti nad 10 model sploh ni uspel razbrati iskanega recepta. Opazimo tudi, da večji kot je kontekst manj se je jezikovni model zanašal nanj. Za vrednosti $k = 10$ in $k = 100$ so prisotne vsaj delne informacije iz podanega konteksta kar pa ni bilo res za $k = 1000$, kjer je podal svoj recept.

Nadaljevali smo z analizo kvalitete odgovorov. Vprašanja za testiranje podobnostnega iskanja smo prilagodili, da so bolj podobna vprašanjem, ki bi jih uporabnik podal jezikovnemu modelu. Preizkusili smo več različnih modelov, vendar smo se na koncu odločili za

model Gemma 3, saj je deloval boljše v kontekstu iskanja receptov kot drugi. Kljub temu je še vedno imel določene pomankljivosti. Tukaj naštevam te pomankljivosti skupaj s primeri poizvedb, ki jih prikažejo.

(1) Jezikovni model ni razlikoval dobro med jeziki. Vprašanja so bila podana izključno v slovenščini, odgovori so pa večinoma v angleščini neodvisno od uporabe dodatnega konteksta. Tudi če je odgovor v sloveščini so lahko prisotne besede, ki niso del slovenskega jezika. Dobili smo tudi dva odgovora, ki sta bila v drugih jezikih, en v hrvaščini in drugi v slovaščini.

Primeri:

- **Vprašanje:** Katere sestavine potrebuješ za bananino torto?

Odgovor brez konteksta:

Potrebuješ:

- Banane
- Mokry šulka
- Cukor
- Škorica
- Konzervovana slivky
- Odstrek vanilky
- Sol'

Vprašanje je bilo podano v slovenščini odgovor je pa v slovaščini.

- **Vprašanje:** Mi lahko svetuješ kaj pripraviti za praznike?

Odgovor brez konteksta:

Here are some ideas for preparing for holidays:

- Food: Plan your holiday meals and start gathering ingredients. Consider traditional dishes or new recipes you'd like to try.
- Decorations: Start decorating your home with festive lights, ornaments, and other holiday decorations.
- Gifts: Begin shopping for gifts for loved ones.
- Activities: Plan holiday-themed activities, such as decorating cookies, attending holiday events, or volunteering.
- Travel: If you're traveling, book flights and accommodations.

Vprašanje je bilo podano v slovenščini odgovor je pa v angleščini. Za večino vprašanj je bil dan odgovor v angleščini.

- **Vprašanje:** Mi lahko svetuješ kaj pripraviti za piknik?

Odgovor brez konteksta:

Seveda! Tukaj je nekaj idej za piknik:

Kosmci: Sandviči ali sendviči z različnimi nadevmi (meso, piščanec, tofu, zelenjava), krauti in sendviči, tortice, sadje (jabolka, banane, jagode, melone), polnitev (sir, kruh, meduzi), keksi in piškotki

Pijače: Voda, sok, pivo ali vino (če je to primerno), kava ali čaj
Oprema: Pladja in posode, ščetke in krtače, papirni brisači, šunka, križanke ali igre, sončna očala in narečje, pletenjak ali odeja

Uživajte v pikniku!

Odgovor s kontekstom:

Za piknik lahko pripravite "Piknik" sladico ali "Lahko malinovo pecivo". Obava se!

Med besedilom opazimo besede, ki niso slovenske, besede, ki niso pravilno napisane, in besede, ki niso pravilno uporabljene: Kosmci, Sandviči, nadevmi, polnitev, meduzi, Pladja, narečje, Obava.

(2) Receptov, ki so prisotni le v Sloveniji, brez konteksta nismo dobili.

Primeri:

- **Vprašanje:** Katere sestavine potrebuješ za čokoladne ježke?

Odgovor brez konteksta:

Za čokoladne ježke potrebuješ čokolado, piškotne koledarje ali vafle za ogrodje in za okras poslikane čokolade, posip in druge sladice.

Odgovor z kontekstom:

Za čokoladne ježke potrebuješ maslo, sladkor, moko, čokolado in jajca.

Odgovor brez konteksta je napačen. Najverjetneje zato, ker čokoladni ježki niso znani izven Slovenije.

- **Vprašanje:** Kako se naredi jabolčni kompot?

Odgovor brez konteksta:

Da naredite jabolčni kompot, potrebujete:

- 1-2 kg jabolk
- 100-200g sladkorja (po okusu)
- Vodo
- Po želji: cimpe, klinčki, limonovo lupino

Postopek:

1. Jabolka olupite, odmrsljate in narežite na male koščke.
2. V ponvi zmešajte jabolka, vodo in sladkor. Dodajte začimbe, če želite.
3. Pustite, da zavre, nato pa zmanjšajte ognjišče in kuhati, dokler jabolka ne postanejo mehka in tekočina ne zgosti.

Odgovor je pravilen tudi brez dodatnega konteksta. Sklepamo, da zato ker je jabolčni kompot poznan v veliko državah. Kljub temu odgovor vsebuje besedo "odmrsljate", ki ni prisotna v nobenem slovarju ali korpusu.

(3) Prisotne so tudi hibe, ki so splošno značilne za jezikovne modele. Opazili smo občasne samozavestno napačne odgovore in halucinacije.

Primeri:

- **Vprašanje:** Kako bi naredili bavarsko kremo po korakih?

Odgovor brez konteksta:

Sestavine:

- 500 ml polno mleko
- 100 g sladkorja
- 4 celica jajca
- 1 vanilinovo olje ali 1 vaniljev list

- 30 g moke

Navodila:

- Obložite posodo z majčnimi soljami.
- Zmešajte mleko, sladkor in jajca.
- Dodajte moko in dobro premešajte.
- Pecite pri 180°C za 45-60 minut ali dokler se krema ne zgosti.

Recept ne vsebuje potrebnih sestavin za bavarsko kremo in nekaj sestavin je izmišljenih.

- **Vprašanje:** Katere sestavine potrebuješ za arašidovčke?

Odgovor brez konteksta:

Potrebuješ: - 115 g masla, ostanek iz zamrzovalnika je idealen
- 115 g medenjaka, zdrobljenega - 1/2 čajne žlične soli - 2 kosovi (po 125 g) moke

Podobno kot prejšnji recept nima vseh nujno potrebnih sestavin.

Kljub vsem pomankljivostim je model deloval dobro s podanim kontekstom. Če smo ga spraševali o specifičnem receptu, ki je znotraj konteksta, je znal odgovoriti, in v primerih kjer recepta ni bilo v kontekstu je to uspel razbrati. Če smo ga spraševali o ločenem delu recepta (sestavinah, postopku, trajanju) je vrnil iskan del besedila. V večini primerov je tudi brez konteksta podal na prvi pogled pravi recept kot na primer zgoraj opisan primer z jabolčnim kompotom.

5 Zaključek

Integralni smo že naredili sistem za podobnostno iskanje z jezikovnim modelom v en sam RAG sistem skupaj s preprostim spletnim vmesnikom. Jezikovni model smo izbrali tako, da smo sestavili nabor 80 vprašanj in analizirali strukturo ter pravilnost odgovorov za vsak model. Modeli llama3.2, deepseek-r1, qwen3 in phi4 so imeli večje pomankljivosti v strukturi odgovorov in jih zato nismo upoštevali. Preostala modela gemma3 in GaMS smo primerjali glede na število vsebinsko pravih odgovorov in se odločili za model gemma3 saj je ta pravilno odgovoril na 42 od 80 vprašanj medtem, ko je model GaMS pravilno odgovoril na 33 od 80 vprašanj. Na koncu smo sestavili sistem z modelom gemma3 še preizkusili podrobneje analizirati prednosti in slabosti izbranega modela in prišli do naslednjih zaključkov:

- **Prednosti:**

- sistem je sposoben pravilno odgovoriti na večino vprašanj z ali brez konteksta,
- sposoben je izluščiti pravilne informacije iz konteksta
- in v večini primerov ugotovi, če je v kontekstu prisoten ustrezen recept.

- **Slabosti:**

- jezikovni model velikokrat odgovori v napačnem jeziku,
- odgovori so zadovoljivi le z relativno majhnim kontekstom kar onemogoča bolj splošna vprašanja,
- brez konteksta ne najde receptov, ki so značilni za slovenski prostor

Poročilo tretjega projekta pri IEPS

- in včasih tudi ponudi "izmišljen" odgovor.