Sistem za pomoganje pri istraživanju

Opis problema

Savremeni sistemi za istraživanje informacija suočavaju se sa izazovima efikasnog pretraživanja, analize i generisanja odgovora na korisnička pitanja iz velikih količina podataka. Ovaj projekat se fokusira na razvoj modularnog sistema koji koristi agente za automatizovano prepoznavanje pitanja, pretragu podataka i generisanje odgovora. Sistem je osmišljen da omogući korisnicima interaktivnu i smisleno vođenu pretragu informacija, uz podršku tehnologija prirodnog jezika i dubokog učenja. Krajnji cilj je obezbediti precizne, koncizne i korisne odgovore, čak i u složenim scenarijima.

Detaljna analiza problema

Osnovni cilj sistema je da pomogne korisnicima u istraživanju relevantnih informacija iz domena veštačke inteligencije, sa posebnim fokusom na oblasti mašinskog učenja i dubokog učenja. Duboko učenje, kao deo šire oblasti mašinskog učenja, obuhvata složene algoritme i modele inspirisane funkcionisanjem ljudskog mozga, što ga čini ključnim za rešavanje kompleksnih problema sa nestruktuiranim podacima poput slika, teksta i zvuka. Sistem je osmišljen kako bi se prvenstveno koristio za istraživanje osnovnih pojmova, arhitektura, i primena dubokog učenja. U okviru toga, obuhvata i teme vezane za povezane oblasti poput konvolucionih neuronskih mreža (CNN), rekurentnih neuronskih mreža (RNN), transformera, i drugih naprednih arhitektura.

Osnovni izazov sistema je obezbeđivanje relevantnih odgovora na korisnička pitanja, pri čemu se mora voditi računa o razumevanju pitanja, efikasnoj pretrazi podataka i generisanju odgovora visokog kvaliteta. Ključni aspekti analize problema uključuju:

- Raznolikost korisničkih pitanja: Pitanja mogu biti različitog tipa – definicije, objašnjenja, analize, poređenja ili liste. Sistem mora adekvatno prepoznati tip pitanja kako bi prilagodio pristup.
- Obim i složenost podataka: Dokumenti koji se pretražuju mogu biti veliki i sadržavati složene informacije, što otežava preciznu i brzu pretragu.

- Preciznost generisanih odgovora: Odgovori moraju biti koncizni, koherentni i tačni, a sistem mora izbeći pružanje nepotpunih ili nerelevantnih informacija.
- Interaktivnost i kontinuitet komunikacije: Sistem mora omogućiti korisnicima da postavljaju dodatna pitanja povezana sa prethodnim odgovorima, pri čemu se održava kontekst razgovora.
- Integracija agenata: Sve faze procesa moraju biti međusobno usklađene i efikasno integrisane.

Opisi podproblema

Problem je razložen na sledeće podprobleme:

1. **Razumevanje korisničkog pitanja**
   Prvi korak je analiza korisničkog pitanja kako bi se izdvojile ključne informacije poput ključnih reči, entiteta i tipa pitanja. Ovo omogućava sistemu da usmeri naredne faze prema relevantnim podacima.
2. **Pretraga informacija u dokumentima**
   Nakon analize pitanja, sistem koristi ključne reči i semantičku pretragu kako bi identifikovao relevantne delove teksta u PDF dokumentima. Glavni izazovi su obim podataka i potreba za preciznim rangiranjem rezultata.
3. **Generisanje odgovora**
   Na osnovu pretraženih informacija, sistem koristi modele za prirodni jezik kako bi generisao jasan, precizan i profesionalno strukturisan odgovor. Ovaj proces uključuje kombinovanje ključnih informacija i sažimanje podataka.
4. **Upravljanje kontekstom i dodatnim pitanjima**
   Sistem mora pratiti tok razgovora i povezivati dodatna pitanja sa prethodnim odgovorima kako bi se obezbedila koherentnost i kontinuitet.
5. **Integracija agenata**
   Svaki agent ima specifičan zadatak, ali njihova međusobna integracija omogućava neometan tok podataka i koordinaciju između faza sistema.

Opis agenata i njihov rad

Sistem se sastoji od četiri osnovna agenta, od kojih svaki ima specifičnu funkciju:

Prvi agent (Razumevanje pitanja)

- Uloga: Analiza korisničkog pitanja kako bi se identifikovale ključne reči, entiteti i tip pitanja. Prvi agent postavlja osnovu za dalju obradu prilagođavajući pristup na osnovu tipa pitanja.
- Tehnologije: SpaCy za jezičku analizu, DistilBERT za prepoznavanje entiteta.
- Rezultat: Strukturisani podaci o pitanju (ključne reči, entiteti, tip pitanja) koji se prosleđuju Drugom agentu.

Drugi agent (Pretraga informacija)

- Uloga: Pretraga relevantnih informacija u PDF dokumentima na osnovu ključnih reči i semantičke sličnosti. Drugi agent se oslanja na ključne reči i originalno korisničko pitanje kako bi pronašao najrelevantnije delove teksta.
- Tehnologije: pdfplumber za ekstrakciju teksta, SentenceTransformer za semantičku pretragu.
- Rezultat: Rangirani pasusi sa informacijama i ocenom relevantnosti, koji se prosleđuju Trećem agentu.

Treći agent (Generisanje odgovora)

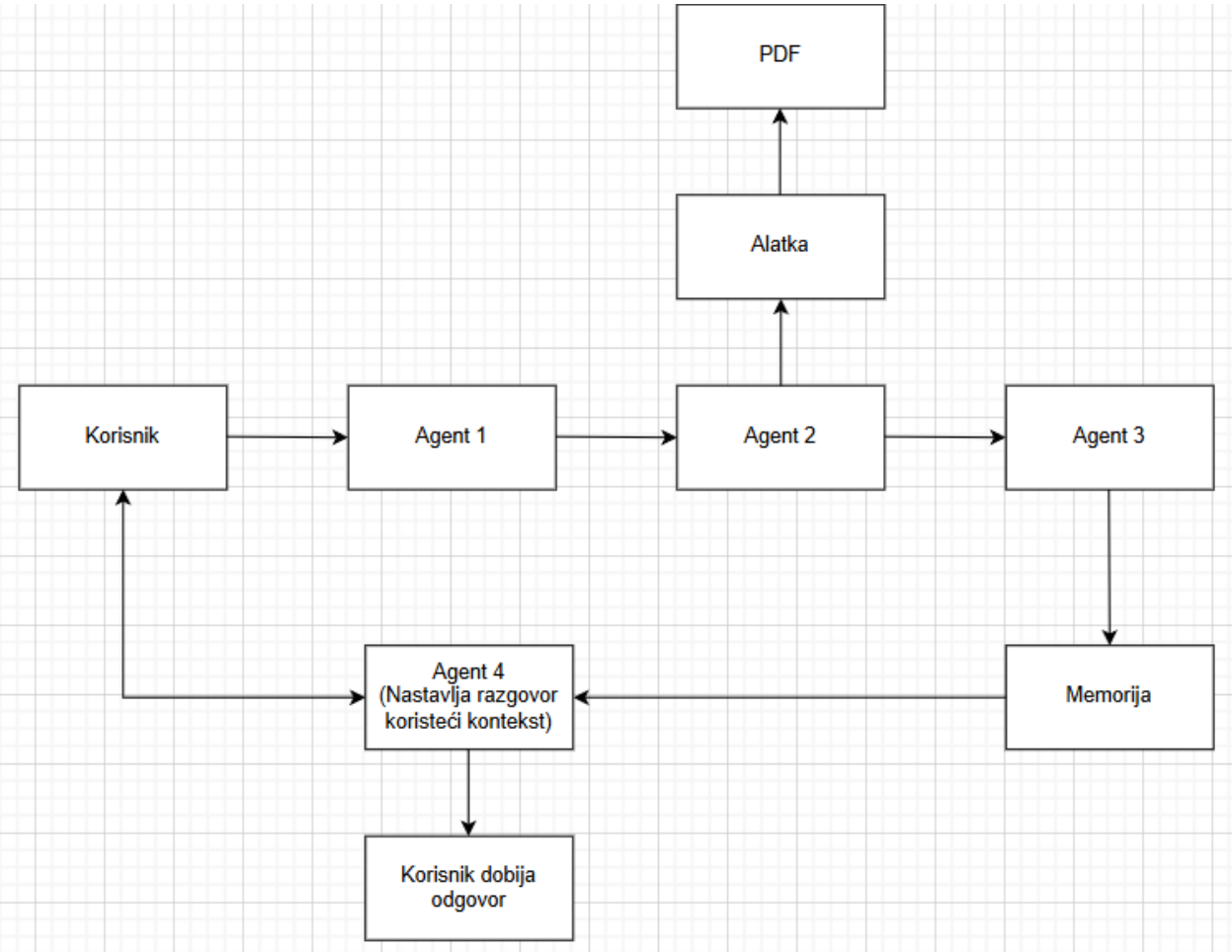
- Uloga: Generisanje smislenog odgovora na osnovu rezultata Drugog agenta. Treći agent koristi informacije iz rangiranih pasusa kako bi kreirao koherentan i precizan odgovor.
- Tehnologije: BART model za sažimanje i generisanje teksta.
- Rezultat: Jasan i precizan odgovor korisniku, prilagođen originalnom pitanju.

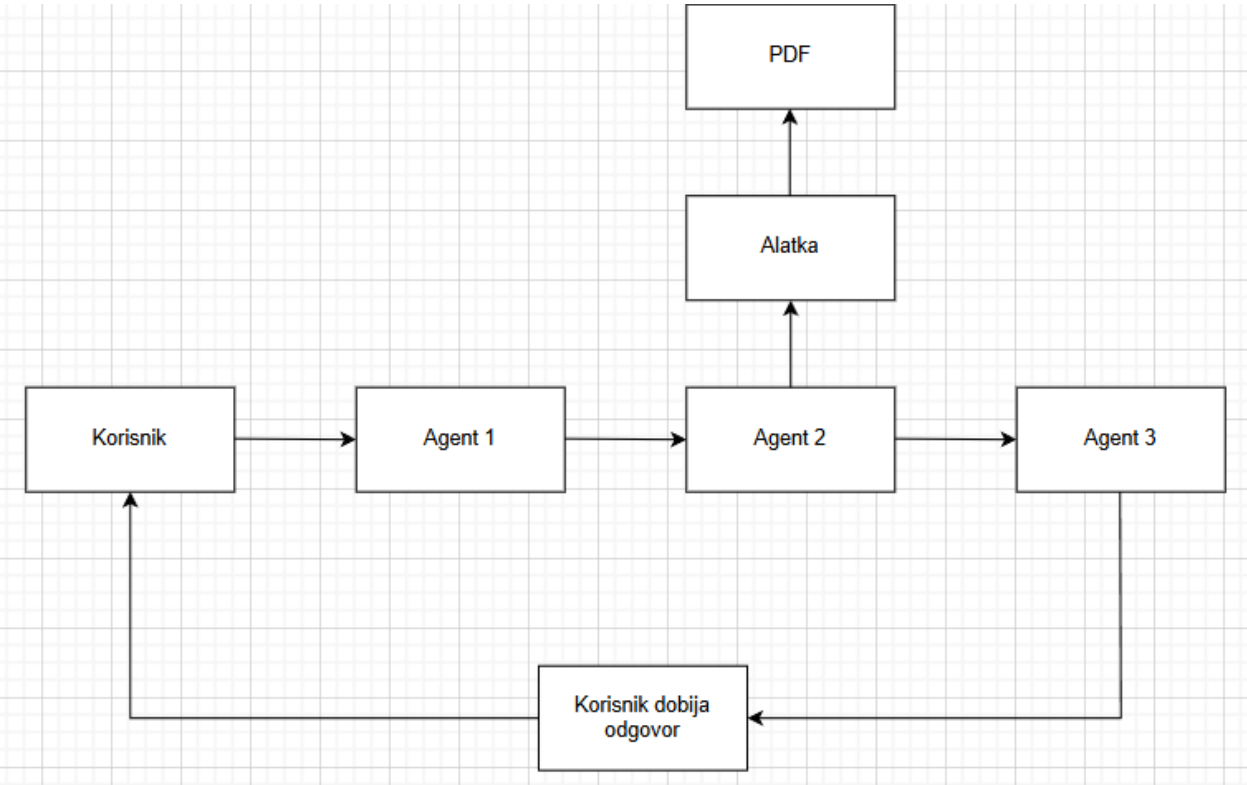
Četvrti agent (Upravljanje kontekstom)

- Uloga: Praćenje toka razgovora i pružanje mogućnosti korisniku da postavlja dodatna pitanja na osnovu prethodnih odgovora. Četvrti agent omogućava sistemu da održi kontekstualnu konzistentnost tokom interaktivne sesije.

- Tehnologije: Python strukture podataka za čuvanje istorije sesija, algoritmi za analizu sličnosti pitanja i odgovora.
- Implementacija:
    - Četvrti agent čuva sve prethodne upite i odgovore u memoriji sistema.
    - Pri svakom novom pitanju, upoređuje ga sa prethodnim kako bi prepoznao kontekstualnu povezanost.
    - Ako je pitanje povezano sa prethodnim, koristi sačuvane podatke kako bi poboljšao rezultate Trećeg agenta ili ponudio dodatna objašnjenja.
- Ograničenja: Iako je implementiran, četvrti agent nije uključen u osnovnu demonstraciju zbog složenosti procene korisničkog konteksta u realnom vremenu.

Dijagram rada sistema


PDF
Alatka
Korisnik
Agent 1
Agent 2
Agent 3
Agent 4
(Nastavlja razgovor
koristeći kontekst)
Memorija
Korisnik dobija
odgovor



Primer:

1.Korisnik unosi pitanje: "What are the applications of CNNs?"

2.Agent 1 preuzima korisničko pitanje,analizira ga,i vraća prepoznate ključne reči,kao I namenu pitanja.


```
{
  "original_query": "What are the applications of CNNs?",
  "keywords": [
    "What",
    "CNNs",
    "the applications"
  ],
  "question_type": "definition"
}
```


3.Agent 2 preuzima odgovor od prvog agenta uz putanju do pdf fajla,nakon čega počinje pretragu po ključnim rečima u pdf fajlu.Kada završi obradu,agent vraća odgovor u obliku json-a,gde se nalaze svi pronađeni rezultati,sa ocenom koja označava relevantnost pronađenih podataka,brojem stranice na kojoj se nalazi taj tekst,I tekst koji je izvukao.

```
{
  "original_query": "What are the applications of CNNs?",
  "question_type": "definition",
  "search_results": [
    {
      "score": 0.6647507548332214,
      "page_number": 4,
      "sentence": "Key components of CNNs include:\n\uf0b7 Convolutional Layers: Use filters (
    },
    {
      "score": 0.6171004772186279,
      "page_number": 4,
      "sentence": "Convolutional Neural Networks (CNNs)\nConvolutional Neural Networks (CNNs)
    },
    {
      "score": 0.6143110990524292,
      "page_number": 3,
      "sentence": "Some prominent applications include:\n\uf0b7 Computer Vision: Object detect
    },
    {
      "score": 0.5894181728363037,
      "page_number": 4,
      "sentence": "Extensions\nof CNNs, such as ResNet, Inception, and EfficientNet, have furt
    },
    {
      "score": 0.5653724670410156,
      "page_number": 4,
      "sentence": "CNNs excel due to their ability to capture hierarchical patterns in images,
    },
```


4.Agent 3 preuzima odgovor od drugog agenta,čisti tekst od nepotrebnih karaktera,i generiše smislen odgovor na osnovu analize sadržaja koji je dobio od drugog agenta,uzimajući u obzir odgovore koji prelaze score od 0.5.

{

'original_query': 'What are the applications of CNNs?',

'generated_answer': 'Convolutional Neural Networks (CNNs) are specialized neural networks designed for processing data with a grid-like topology, such as images. CNNs excel due to their ability to capture hierarchical patterns in images, starting from low-level features (edges) to high-level representations (shapes or objects) Some prominent applications include: Computer Vision, facial recognition, medical imaging, and autonomous vehicles.'

}

Zaključci

Rad na ovom projektu omogućio je uspešnu implementaciju sistema za podršku pri istraživanju, zasnovanog na višestepenoj analizi i obradi podataka putem agenata. Sistem se pokazao kao efikasan u obradi korisničkih upita, pretrazi relevantnih informacija i generisanju smislenih odgovora. Tokom izrade projekta,

primenjene su napredne tehnike obrade prirodnog jezika (NLP) i dubokog učenja, koje su integrisane u koherentan i modularan sistem.

Ključne prednosti sistema uključuju:

1. Modularnu arhitekturu koja omogućava lako proširenje i prilagođavanje.
2. Korišćenje savremenih tehnologija poput Transformer modela (BART) i biblioteka za semantičku analizu.
3. Preciznu i korisniku prilagođenu interakciju sa sistemom.

Iako je sistem uspešno realizovan, tokom rada identifikovani su određeni izazovi, poput potrebe za optimizacijom performansi i unapređenjem rukovanja kompleksnim korisničkim upitima. Takođe, razvijen je i četvrti agent za nastavak razgovora sa korisnikom, koji predstavlja potencijalnu buduću iteraciju sistema.