MŰEGYETEM 1782

Dokumentumfeldolgozáson alapuló magyar nyelvű chat AI RAG technológiával

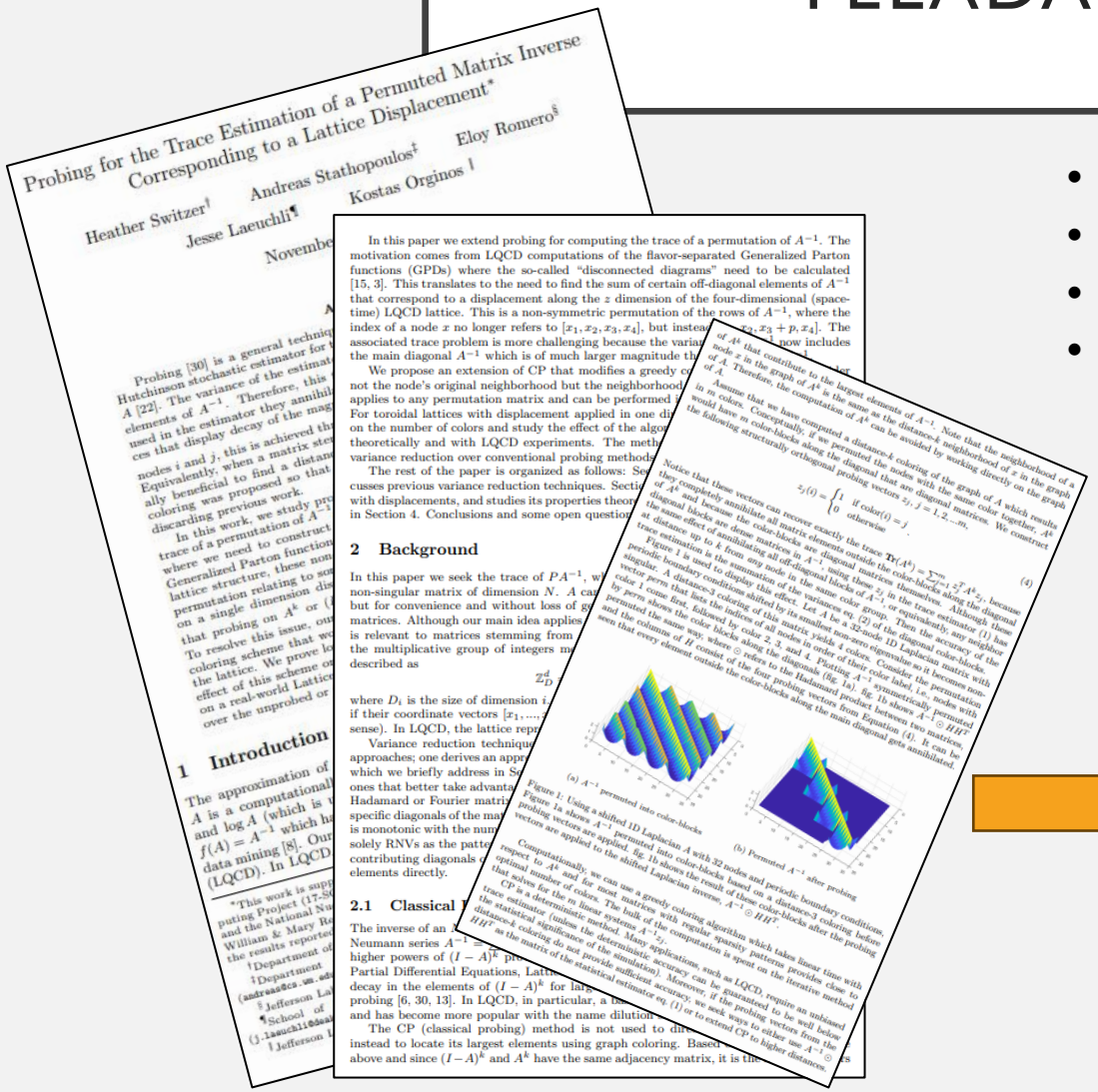
Önálló laboratórium 2 – Szladek Máté (TGPZTT)

Konzulensek: Gyires-Tóth Bálint és Ónozó Lívia

https://github.com/TGPZTT/Onlab2

2024.05.21.

FELADAT BEMUTATÁSA



- Nem feltétlenül strukturált kérdések
- Pontos válaszok
- Hivatkozások (forrás)
- Magyar nyelvű válasz (lehet angol a forrás)

MEGOLDÁSOK

Kulcsszavas keresés

Nagyon egyszerű
Könnyen implementálható
Kontextus nem vizsgál, pontatlan

Szemantikus keresés

A szöveg jelentését is ismeri (vektoros reprezentációk)
Módszertani korlátok: Lexikai homonímia, Out-of-vocabulary

Q&A rendszerek

Információkinyerési technológia
Mély tanulás megjelenésével óriási fejlődésnek indult

Témamodellezés

Transzformer modellek fine-tuningolása

Még sok más…

RAG - Retreival Augmented Generation

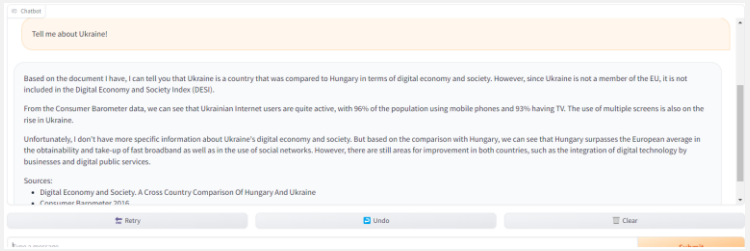
- Generatív- és információ-visszakeresési (IR) modell ötvözése
- Egyszerre ad választ és tanul is. (A legtöbb megoldás külön lépésekben csinálja)
- Kontextusérzékeny
- Iteratív válaszfinomítás
- Tudásintegrálás – a nyelvi modell meglévő tudása + a visszakeresett információk

FELHASZNÁLT KOMPONENSEK

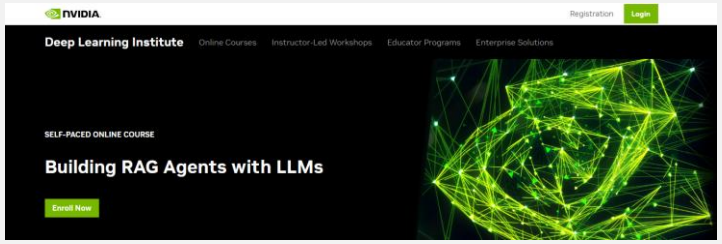
- arxiv
- FIASS vektortároló (META – facebook)
- RAG-os LangChain
- Gradio chatbot

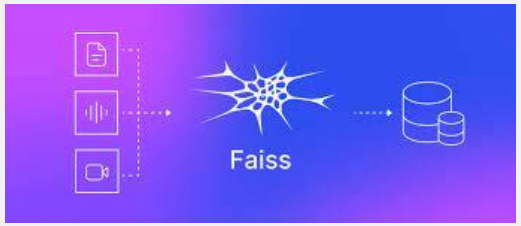
```
## Start kérdés
test_question = "Tell me about Hungary!"
```

```
Sure, I'd be happy to tell you about Hungary! When it comes to the digital economy
  and society, Hungary has made progress in recent years. According to the Digital Economy
  and Society Index (DESI), Hungary ranks 14th in connectivity, which is an improvement
  from 16th place in 2016. The country has seen an increase in fast broadband coverage,
  which now stands at 81%, and the government has launched initiatives to increase demand
  for broadband, such as a preferential VAT rate for subscriptions and a "digital welfare
  basic tariff" for non-users.

Hungary's human capital in terms of digital skills is a bit of a mixed picture. While
  only half of the population has at least basic digital skills, the country has a higher
  share of ICT specialists in the labor force (4.9%) compared to the EU average (3.7%).
  However, the study notes that integration of digital technology by businesses is a major
  issue in Hungary, and that Hungarian firms should make better use of online business,
  social media, and cloud-based applications.

When it comes to e-commerce, the most popular device in Hungary is still the computer,
  but the popularity of smartphones and tablets is on the rise. The use of multiple
  screens is also becoming more common in Hungary.

As for the use of the internet, Hungary performs well above the European average. In
  fact, Hungarian internet users are avid consumers of online content, with 96% of the
  population using mobile phones and 93% having TV. The ratio of STEM graduates per 1000
  individuals aged 20-29 is quite low in Hungary, but the government has implemented a new
  Digital Competences Strategy to increase the ratio of STEM graduates and address lifelong
  learning.

Overall, Hungary's DESI score is above the EU average in the use of internet and
  somewhat below the average in connectivity and human capital. The country needs to improve
  the integration of digital technology by businesses and digital public services.

Sources:

* Digital Economy And Society. A Cross Country Comparison Of Hungary And Ukraine
* E-commerce in Hungary: A Market Analysis
* The Digital Scoreboard (EC 2016b)
* Consumer Barometer 2016
```



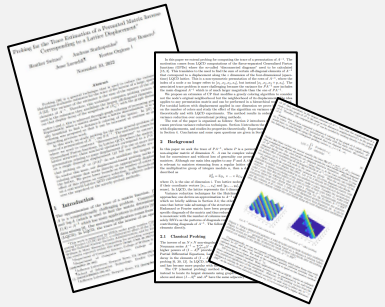

```
{'input': 'Tell me about Hungary!',
'history': '', 'context': '[Quote
from Digital Economy And Society. A
Cross Country Comparison Of Hungary
And Ukraine] Table 2. Connectivity
scores Source: EC (2017)\nHungary
ranks 14th in this component,
compared to\n16th in 2016. Hungary
has made progress both in the
supply \nand the demand side. Fast
broadband coverage increased to
\n81% from 78%. The Hungarian
government launched two
\ninitiatives to increase demand
[...] }
```





RAG FOLYAMAT

Dokumentumbetöltés → Szövegfeldolgozás → Vektorizálás → Vektortárolás → Lekérdezés → Láncolás (válasz)



arXiv

Akár különböző forrásokból

RecursiveCharacterTextSplitter

Szöveg da rabolás

1000 karakter méretű, de átfedő szövegszeletek

Ha X<200 akkor kuka

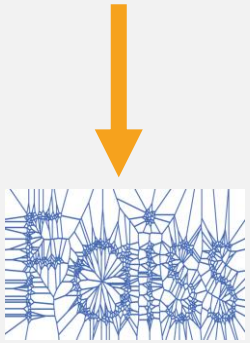
Szöveg

nvolveqa_40k modell

numerikus vektorok

Vektorok:

- Szövegrészek
- Metaadatok
- Extra információk (cím)



- A kérdésnek megfelelő max 5000 karakteres válaszcsomag a vektortárolóból

```
{'input': 'Tell me about Hungary!', 'history': '', 'context': '[Quote from Digital Economy And Society. A Cross Country Comparison Of Hungary And Ukraine] Table 2. Connectivity scores Source: EC (2017)\nHungary ranks 14th in this component, compared to\n16th in 2016. Hungary has made progress both in the supply [...] }
```

- Prompt engineering
- **Input:** kérdés + válaszcsomag
- **Output:** Megfogalmazott válasz
- Ha nincs válasz a dokukban, akkor a modell önálló tudását is használhatja, csak akkor nem hivatkozik

DOKUMENTUM - CHATBOT

embender

- ai-mixtral-8x7b-instruct
- nvolveqa_40k
- LangChain

Képes nagy bemeneteket kezelni (llama3 nem)

Stream chain: chat prompt | llm

Angol válasz

mixtral_8x7b

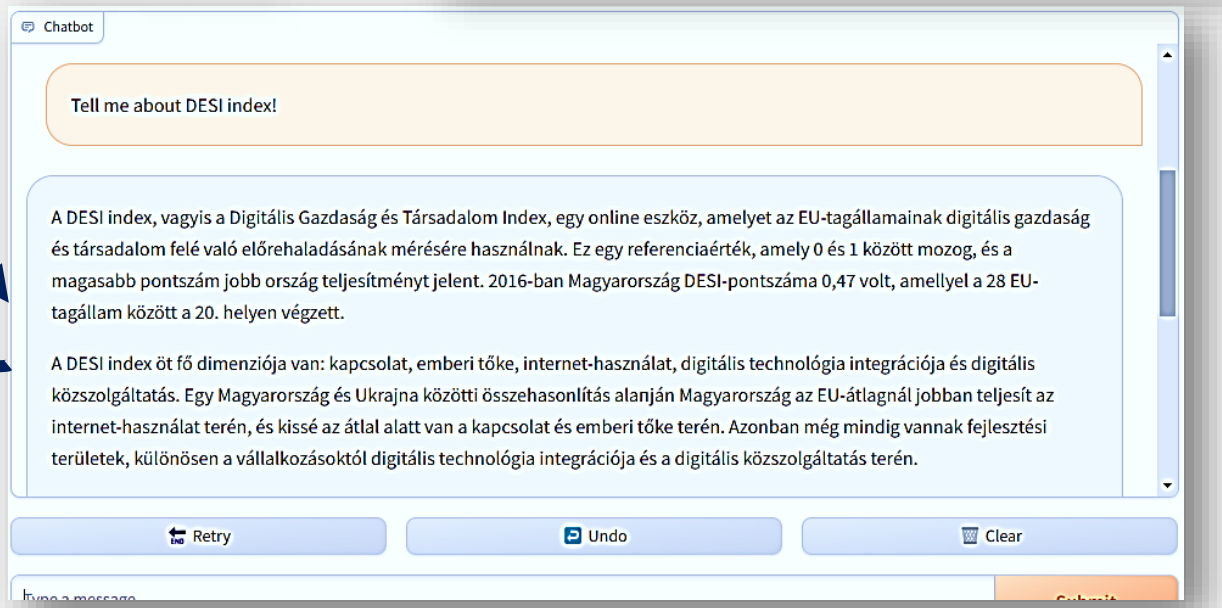
LE CHAT MISTRAL_

Fordító modellnek: llama3

Promptolással jó nyelvezetet ad.

```
print(translate("This apple is red, but not green. "))

Ez az alma piros, de nem zöld.
```

API key kérdés



MŰEGYETEM 1782

KÖSZÖNÖM A FIGYELMET!