

Digital humanities

Cvičení: velké jazykové modely

Jindřich Marek

1. Stejný dotaz pro různé modely

Nejčastější typy dotazů

- faktické otázky
 - Jaké je hlavní město Francie?
- otevřené otázky
 - Jaké jsou výhody každodenního cvičení?
- otázky vyžadující zdůvodnění nebo vyvození
 - Pokud venku prší, co mám dělat?

Příklad složitějšího dotazu (česky/angl.)

Představte si, že jste dostali za úkol navrhnout projekt v oblasti digitálních humanitních věd, který zahrnuje kvalitativní i kvantitativní data. Popište, jak byste v tomto projektu použili velký jazykový model, a uveďte jeho konkrétní přínos a případné etické aspekty, které je třeba vzít v úvahu. Svoji odpověď napište česky.

Imagine you have been tasked with designing a digital humanities project that incorporates both qualitative and quantitative data. Describe how you would use a large language model in this project, outlining its specific contributions and any ethical considerations that must be taken into account.

Modely k vyzkoušení dotazu

- ChatGPT 3.5 <https://chat.openai.com/>
- Mixtral-8x22B-Instruct <https://labs.perplexity.ai/>
- Llama-3-70B-Instruct <https://labs.perplexity.ai/>
- Phi3 (lokálně)

2. Obohacování dat

- prezentovaný výstup: model Llama 3 8B Instruct, Q8_0

BUDOVA PZO CENTROTEX

**A78/1047**

**VÁCLAV HILSKÝ, OTAKAR JURENKA,
JIŘÍ NÁHLÍK
1972 - 1978
NÁMĚSTÍ HRDINŮ 3, PRAHA 4
FOTO: PAVEL HRUBÝ
TEXT: PAVLÍNA KARBANOVÁ
KRÁSNÁ**

Komplex dvou budov přezdíváný "Kráva a tele" patří do série reprezentativních budov postavených pro Podnik zahraničního obchodu. Je tvořen výraznou výškovou budovou a menší budovou v sousedství, ve které se nacházel výzkumný ústav. Budova v sobě dnes mimo kanceláří Ministerstva vnitra skrývá vestibul metra, supermarket a řadu menších obchodů. Střešní nástavba, která připomíná korunu zakrývá technologická zařízení. V rámci projektu stavby byl navržen i betonový výdech metra pojatý jako abstraktní plastika.



Obohacování dat – vstup

I will give you a text copied from the web and you will encode it in JSON.

BUDOVA PZO CENTROTEX

A78/1047

VÁCLAV HILSKÝ, OTAKAR JURENKA, JIŘÍ NÁHLÍK

1972 - 1978

NÁMĚSTÍ HRDINŮ 3, PRAHA 4

FOTO: PAVEL HRUBÝ

TEXT: PAVLÍNA KARBANOVÁ KRÁSNÁ

Komplex dvou budov přezdívaný "Kráva a tele" patří do série reprezentativních budov postavených pro Podnik zahraničního obchodu. Je tvořen výraznou výškovou budovou a menší budovou v sousedství, ve které se nacházel výzkumný ústav. Budova v sobě dnes mimo kanceláři Ministerstva vnitra skrývá vestibul metra, supermarket a řadu menších obchodů. Střešní nástavba, která připomíná korunu zakrývá technologická zařízení. V rámci projektu stavby byl navržen i betonový výdech metra pojatý jako abstraktní plastika.

Obohacování dat – výstup

```
{
  "text": [
    {
      "title": "BUDOVA PZO CENTROTEX",
      "address": "NÁMĚSTÍ HRDINŮ 3, PRAHA 4"
    },
    "A78/10", "47",
    {
      "authors": ["VÁCLAV HILSKÝ", "OTAKAR JURENKA", "JIŘÍ NÁHLÍK"]
    },
    "1972 - 1978",
    {
      "photo": "Pavel Hrubý"
    },
    "TEXT: PAVLÍNA KARBANOVÁ KRÁSNÁ"
  ],
  "description": [
    "Komplex dvou budov přezdívaný \"Kráva a tele\" patří do série reprezentativních budov postavených pro Podnik zahraničního obchodu.",
    "Je tvořen výraznou výškovou budovou a menší budovou v sousedství, ve které se nacházel výzkumný ústav.",
    "Budova v sobě dnes mimo kanceláři Ministerstva vnitra skrývá vestibul metra, supermarket a řadu menších obchodů.",
    "Střešní nástavba, která připomíná korunu zakrývá technologická zařízení.",
    "V rámci projektu stavby byl navržen i betonový výdech metra pojatý jako abstraktní plastika."
  ]
}
```

Obohacování dat – výstup (po upřesnění)

```
{  
  "text": [  
    {  
      "title": "BUDOVA PZO CENTROTEX",  
      "address": "NÁMĚSTÍ HRDINŮ 3, PRAHA 4"  
    },  
    "A78/10", "47",  
    {  
      "authors": ["Václav Hlinský", "Otakar Jureňka", "JIŘÍ Náhlík"]  
    },  
    {  
      "date": "1972 - 1978",  
      "textAuthor": "Pavlína Karbanová Krásná"  
    },  
    {  
      "photo": "Pavel Hrubý"  
    }  
  ]  
}
```

Obohacování dat – výstup (po 2. upřesnění)

```
{
  "text": [
    {
      "title": "BUDOVA PZO CENTROTEx",
      "location": {"name": "NÁMĚSTÍ HRDINŮ 3, PRAHA 4"}
    },
    "A78/10", "47",
    {
      "authors": [
        {"name": "Václav Hliský"},
        {"name": "Otakar Jureňka"},
        {"name": " Jiří Náhlík"}]
    },
    {
      "date": {"start": "1972", "end": "1978"},
      "textAuthor": {"name": "Pavlína Karbanová Krásná"}
    },
    {
      "photo": {"name": "Pavel Hrubý"}
    }
  ]
}
```

Použité prompty

- upřesnění
 - Encode the line beginning with TEXT and stating the text author as a string property and string value. Change the string property accordingly.
 - Add the string property to the datation.
 - Normalize the letter case for all personal names.
- druhé upřesnění
 - And now try basic named entity recognition in the text and change the JSON accordingly.

2. Rozpoznávání motivů na obrázcích

Multimodální modely

- llama-v1.5-7B lokálně přes llamafile
 - <https://github.com/Mozilla-Ocho/llamafile>

llama.cpp



User: What can you see in the picture?

Llama: The image is a colorful, old-fashioned painting of three women dressed in Renaissance costumes. They are standing next to each other and appear to be engaged in conversation or enjoying themselves together. Each woman has a unique outfit, adding variety to the scene.

In addition to their clothing, one of the women is holding a rose, which adds an elegant touch to the painting. The overall atmosphere of the image suggests that it may have been created during the Renaissance period and serves as a visual representation of social interactions or artistic expression from that time.

Say something...

3. Práce s vlastními daty

MZ - 800:

OSOBNÍ OSMIBITOVÝ SHARP

Vloni se na našem trhu objevily počítače MZ-821 s vestavěným magnetofonem a MZ-811 bez magnetofonu. Nemýlíte se, jde skutečně o počítače SHARP typové řady MZ-800, která se již rok předtím objevovala v nabídce objednávkové služby PZO Tuzex. První, co zájemce o tento atraktivní počítač překvapilo, byly poměrně příznivé ceny obou typů — 7800 a 6600 Kčs (ve srovnání s kuriózními 2490 TK v PZO Tuzex). A další příjemné překvapení přišlo už při seznamování se základními technickými daty počítače.

Osvědčený procesor Z 80A, paměť RAM 64 KB, ROM 16 KB a VideoRAM 16 KB s možností rozšíření až na 32 KB — to jsou základní rysy řady MZ-800. A dále: dvakrát vyvedená sběrnice pro připojení periférií a rozšiřujících zařízení pomocí přímých konektorů, dva konektory pro připojení křížových ovladačů slučitelných s Atari, paralelní rozhraní Centronics.

Obrazové výstupy: barevný monitor připojený přímo pomocí konektoru RGB nebo televizor připojený buď přímo přes Video-výstup, nebo pomocí modulovaného výstupu do pásma kanálů 33 až 39 s možností individuálního jemného doladění.

Grafické rozlišení: buď 320 × 200 bodů nebo 640 × 200 bodů, kterému odpovídá 40 nebo 80 znaků na 25 řádek textu. V prvním případě lze v normální konfiguraci (po rozšíření VideoRAM) zobrazit čtyři (šestnáct) barev, ve druhém případě dvě (čtyři) barvy.

Spínaný zdroj zabudovaný přímo do počítače zaručuje, že na stole nevznikne nechvalně známý kabelový salát. Do zátěže mimo samotný počítač dodá proud až 2,5 A, takže lze připojit poměrně velké množství periférních zařízení, aniž by se uživatel musel starat o jejich zvláštní napájení.

Tříkanálový zvukový a jednocanálový šumový generátor SN 76489 včetně zabudovaného zesilovače s regulací hlasitosti a reproduktoru poskytují dostatečně kvalitní zvuk v rozsahu šesti oktáv. Velice pohodlná kláves-

MZ - 800:
OSOBNÍ
OSMIBITOVÝ
SHARP



Zpracování vlastních dat (text, tabulka)

- vnoření textu (text embedding)
 - model pro vnořování
 - důležitý je rozsah podpory jazyků (angličtina, mezinárodní)
- uložení vnořeného textu ve vektorové databázi
- dotazování na základě databáze
 - model pro dotazování (obecné modely)

localGPT: instalace

- <https://github.com/PromptEngineer/localGPT>
- potřebuje git, VS Build Tools, cmake a conda (Miniconda)

```
git clone https://github.com/PromptEngineer/localGPT.git
```

```
cd localGPT
```

```
conda create -n localGPT python=3.10.0
```

```
conda activate localGPT
```

```
pip install -r requirements.txt
```

```
pip install llama-cpp-python
```

localGPT: nastavení

- v souboru constants.py

- zvolit model pro embedding:

```
EMBEDDING_MODEL_NAME = "intfloat/multilingual-e5-large"
```

(u méně výkonných počítačů a pro rychlejší a méně přesné zpracování je možné zvolit intfloat/multilingual-e5-base)

- pro interpretaci ponechat výchozí

- přesunout dokumenty do složky SOURCE_DOCUMENTS

localGPT: vložení dokumentů

- vložení dokumentů do vektorové databáze

```
python ingest.py --device_type cpu
```

- parametr `--device_type cpu` pro případ, že nemáme kvalitní GPU
- pro ingest txt souborů upravit v `ingest.py` `loader = loader_class(file_path)`
na `loader = loader_class(file_path, encoding='UTF-8')`

localGPT: spuštění

```
cd localGPT
```

```
conda activate localGPT
```

```
python run_localGPT.py --device_type cpu
```

- parametr `--device_type cpu` pro případ, že nemáme kvalitní GPU
- při vkládání dokumentů a prvním spuštění se stáhnou modely
 - celkem několik GB dat

```

llm_load_print_meta: LF token      = 13 '<0x0A>'
llm_load_tensors:  ggml ctx size =   0.11 MiB
llm_load_tensors:   CPU buffer size = 3891.24 MiB
.....
llama_new_context_with_model: n_ctx      = 4096
llama_new_context_with_model: n_batch   = 512
llama_new_context_with_model: n_ubatch  = 512
llama_new_context_with_model: freq_base = 10000.0
llama_new_context_with_model: freq_scale = 1
llama_kv_cache_init:   CPU KV buffer size = 2048.00 MiB
llama_new_context_with_model: KV self size  = 2048.00 MiB, K (f16): 1024.00 MiB, V (f16): 1024.00 MiB
llama_new_context_with_model:   CPU output buffer size =   0.12 MiB
llama_new_context_with_model:   CPU compute buffer size = 296.01 MiB
llama_new_context_with_model: graph nodes = 1030
llama_new_context_with_model: graph splits = 1
AVX = 1 | AVX_VNNI = 0 | AVX2 = 1 | AVX512 = 0 | AVX512_VBMI = 0 | AVX512_VNNI = 0 | FMA = 1 | NEON = 0 | ARM_FMA = 0 | F16C = 1 | FP
16_VA = 0 | WASM_SIMD = 0 | BLAS = 0 | SSE3 = 1 | SSSE3 = 0 | VSX = 0 | MATMUL_INT8 = 0 |
Model metadata: {'general.name': 'LLaMA v2', 'general.architecture': 'llama', 'llama.context_length': '4096', 'llama.rope.dimension_c
ount': '128', 'llama.embedding_length': '4096', 'llama.block_count': '32', 'llama.feed_forward_length': '11008', 'llama.attention.hea
d_count': '32', 'tokenizer.ggml.eos_token_id': '2', 'general.file_type': '15', 'llama.attention.head_count_kv': '32', 'llama.attentio
n.layer_norm_rms_epsilon': '0.000001', 'tokenizer.ggml.model': 'llama', 'general.quantization_version': '2', 'tokenizer.ggml.bos_toke
n_id': '1', 'tokenizer.ggml.unknown_token_id': '0'}
Using fallback chat format: None

```

Enter a query: |

Dotazy k textu

- What was the price of the MZ-800?
- What problems did people in Czechoslovakia have when they wanted to buy computer accessories?
- What options did people have when they wanted to acquire software?

What was the price of the MZ-800?

Based on the provided context, the price of the MZ-800 was 7800 Kč (approximately 320 USD) in 1988.