UNIVERSITÄT HEIDELBERG
ZUKUNFT SEIT 1386

Software Practicals
Winter Semester 2024/25

Data Science Group
Heidelberg University
October 16, 2024

Slides Online

The slides are available on our webpage
https://ds.ifi.uni-heidelberg.de/teaching/current/

Organization

Outline

UNIVERSITÄT HEIDELBERG
ZUKUNFT SEIT 1386

- Overview of topics (today)
  - Send application for a topic until **Monday, October 21, 1pm**
  - Assignment of topics by October 23
- First milestone (before Christmas break)
  - Prototype / part of software
  - Summary of research (literature and related systems/tools)
  - Further milestones in agreement with supervisor
- End of practical (mid/end February)
  - Code has to be in local Gitlab of the database group
  - Presentation / demo of practical and software (10-12 minutes)
  - Report / documentation as Gitlab document (README.md)

Application

- Apply directly to supervisor via mail
  - Program of study, semester of study, matriculation number
  - List relevant course experience, including course grades
  - List other experience:
    - Side projects you are working on
    - “Anwendungsgebiet” / Application Field
    - Job and project experience
  - Send your tentative schedule and milestones for the practical
  - Group work is not possible!
- It is recommended to apply for multiple topics (e.g., “top-3 list”)

Application is binding!
Don’t apply if you don’t want to do the practical!

Deadlines

- In general: biweekly meetings with supervisor
- Presentation: end of February 2025
- Report & Gitlab upload: end of February 2025
- No extension possible

<u>Not finished = failed (grade 5,0)!</u>

Assessment

UNIVERSITÄT HEIDELBERG
ZUKUNFT SEIT 1386

- Credit points (Leistungspunkte)
  - Beginners Practical (IAP, 2 CP + 4 FÜK) [Bachelor students]
    - workload: 180 h (~1 ½ days/week)
  - Advanced Practical (IFP, 8 CP)
    - workload: 240 h (~2 days/week)
- Grading based on
  - code (readability, structure, functionality; code in local Gitlab)
  - documentation (README.md, code comments, documentation in Gitlab)
  - commitment and self-reliance
  - cool ideas!!
- IMPORTANT
  - talk to / communicate with your advisor (at least biweekly meetings)

Supervisors

- Michael Gertz (MG)
  gertz@informatik.uni-heidelberg.de
- Marina Walther (MW)
  walther@informatik.uni-heidelberg.de
- Ashish Chouhan (AC)
  chouhan@informatik.uni-heidelberg.de
- Nicolas Reuter (NR)
  reuter@informatik.uni-heidelberg.de

Project Topics

AP = Advanced Topic

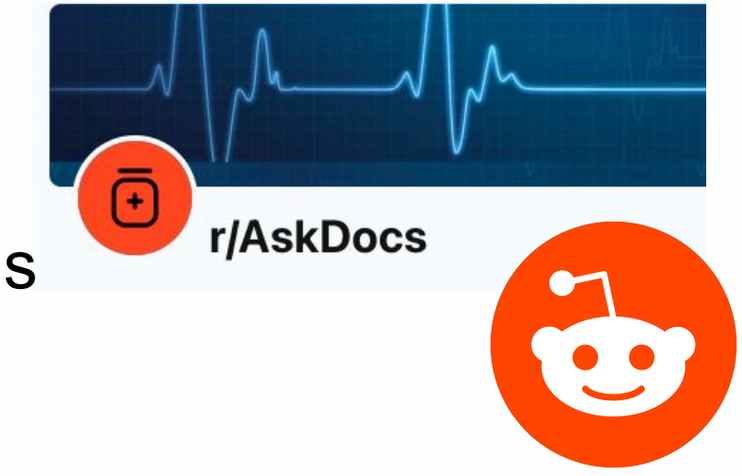
BP = Beginners Topic (for BSc students only)

Overview of Topics

UNIVERSITÄT HEIDELBERG
ZUKUNFT SEIT 1386

1. Expert vs. LLM generated Medical Advice, **BP** (Walther)
2. KG Enrichment using Semi-Structured Data, **AP** (Walther)
3. Diversified Answer Generation QA System, **AP** (Chouhan)
4. Agentic Environmental Conversation System, **AP** (Chouhan)
5. Query Databases using Natural Language, **AP/BP** (Reuter)
6. Evaluation of General Purpose LLMs on Table Data, **BP** (Reuter)
7. Image Text Extractor, **AP** (Gertz)
8. Crawler for Legal Advice Websites, **AP** (Gertz)

BP: Expert vs. LLM generated Medical Advice (MW)



Given:

- r/AskDocs (Reddit) dataset containing QA pairs between consumers and physicians

Tasks:

- Statistical analysis of dataset
- Perform experiments using different LLMs to compare expert and generated answers

Subtasks:

- Find suitable analyses, metrics, and visualizations.
- Store dataset + results & extend existing UI

Languages / Tools: Python, OpenSearch, LangChain, FastAPI, Svelte, MongoDB

AP: KG Enrichment using Semi-Structured Data (MW)

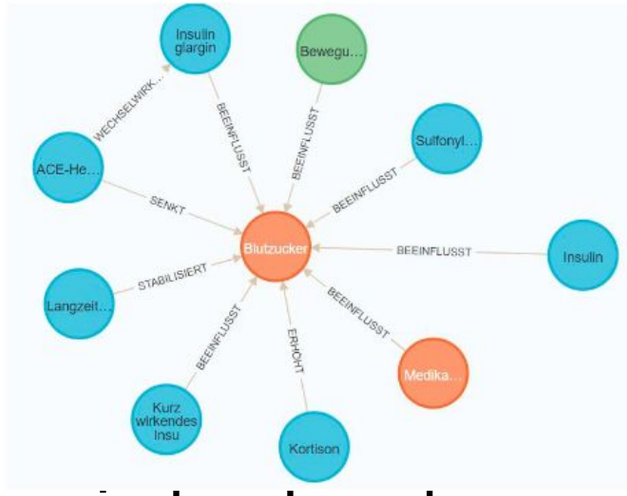
UNIVERSITÄT HEIDELBERG ZUKUNFT SEIT 1386

Given:

- (small) Medical KG constructed from Apotheken Umschau

Tasks:

- Extend existing KG with another data source, i.e. active ingredient (drug) profiles
- Explore enhanced KG, e.g. drug interactions



Subtasks:

- Process and store German pharmacological semi-structured dataset
- Conceptualize and implement methods to update existing KG

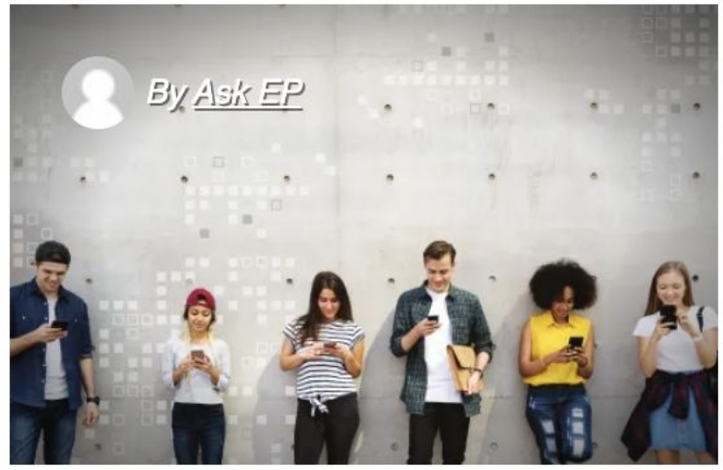
Languages / Tools: Python, Neo4J, bs4, …

AP: Diversified Answer Generation QA System (AC)

UNIVERSITÄT HEIDELBERG
ZUKUNFT SEIT 1386

Given:

- Ask EP answer questions raised by citizens on different topics



BLOG, EP ANSWERS / 3 MONTHS AGO

Regulating social media: What is the European Union doing to protect social media users?

Tasks:

- Extract QA pairs and mentioned documents from the posts and store them in OpenSearch

Subtasks:

- Realize backend to store extracted information using OpenSearch
- Diversified answer generation with Retrieval-Augmented Generation to interact with the stored information

Languages / Tools:

- Python, OpenSearch, LangChain, Hugging Face

AP: Agentic Environmental Conversation System (AC)

Given:

- European Commission provides EU action information on different environmental topics

Tasks:

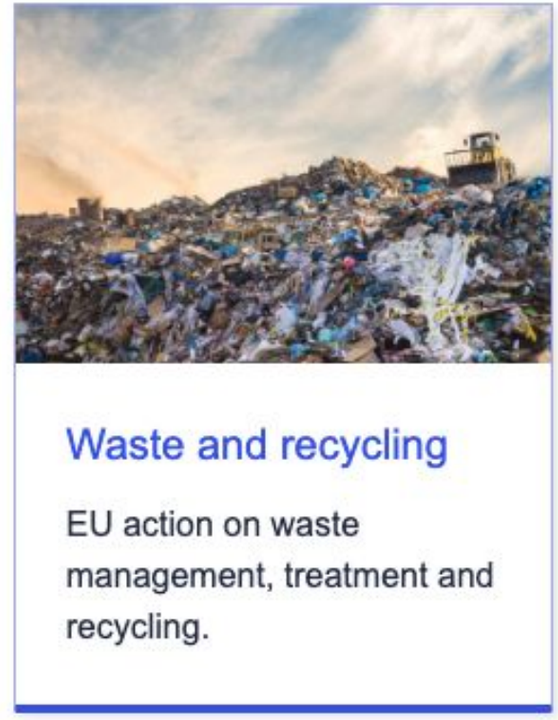
- Create conversational dataset using agents on collected information

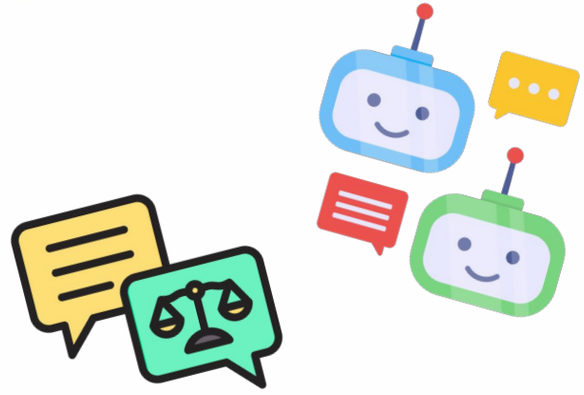
Subtasks:

- Collect information and realize backend to store information using OpenSearch
- Conversation generation on stored information using agents

Languages / Tools:

- Python, OpenSearch, LangChain, LangGraph Hugging Face



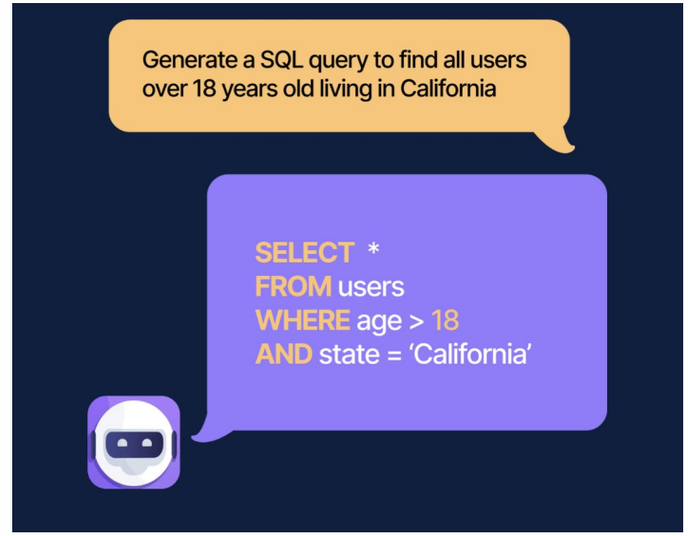


BP/AP: Query Databases using Natural Language (NR)

UNIVERSITÄT HEIDELBERG
ZUKUNFT SEIT 1386

Given:

- An example database
- Some sample SQL queries and their natural language equivalent



Tasks:

- Build a pipeline to query database using prompts in natural language
- Evaluate pipeline and chosen methods by comparing generated SQL queries with ground truth queries and their results

Subtasks:

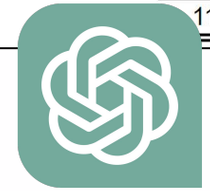
- Integrate database schema/data into the pipeline
- Select different language models and compare their performance

Languages / Tools:

- Python, SQL, LangChain or Llamaindex, OpenAI

BP: Evaluation of General Purpose LLMs on Table Data (NR)

	Three Months Ended		
	December 30, 2023	December 31, 2022	Change
Net sales by category:			
iPhone	$ 69,702	$ 65,775	6 %
Mac	7,780	7,735	1 %
iPad	7,023	9,396	(25)%
Wearables, Home and Accessories	11,953	13,482	(11)%
Services	23,117	20,766	11 %
Total net sales	119,575	$ 117,154	2 %



Given:

- Set of HTML documents with tables

Tasks:

- Evaluate the performance of different LLMs in answering questions using data from tables

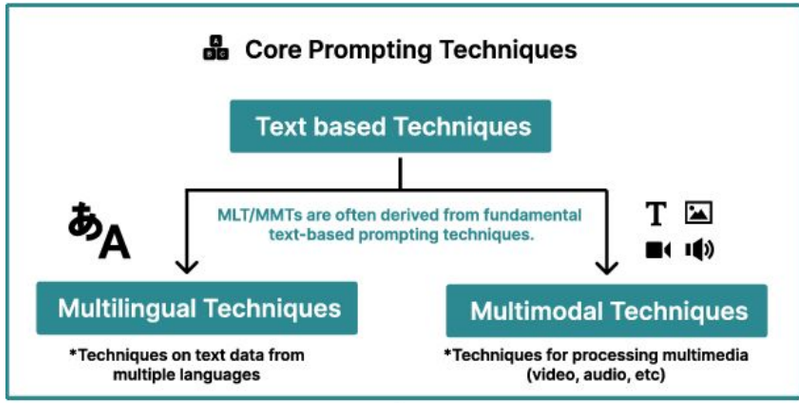
Subtasks:

- Create sample questions for tables from the dataset
- Try different formats for the tables as input to LLMs (e.g., HTML, textual description, comma-separated cells, JSON, ...)

Languages / Tools:

- Python, LangChain or Llamaindex, OpenAI, …

AP: Image Text Extractor (MG)



Given:

- PDF documents that include images;
  such images may contain textual information

Tasks:

- Develop and implement pipeline to extract images from PDF documents
- Manage extracted text (incl. from images) in data store

Subtasks:

- Compare different framework (VLMs) for text extraction from images;
- Build pipeline that ingest PDF documents for processing

Languages / Tools:

- Python; Phi-3-Vision-128K-Instruct and other (V)LMs

BP/AP: Crawler for Legal Advice Websites (MG)

UNIVERSITÄT HEIDELBERG ZUKUNFT SEIT 1386

Given:

- German legal advice websites where users can ask questions and legal experts answer

Frag-einen-Anwalt.de

ANWALT.DE

eins zwei drei 123recht.de

Tasks:

- Design and implement adaptive crawler for select websites (and legal topics)
- Implement data schema in Opensearch or PostgreSQL

Subtasks:

- Design data schema for storing question/answer pairs incl. Metadata.
- Simple frontend to browser data store

Languages / Tools:

- Python; Opensearch/PostgreSQL; Streamlit/Django

Slides Online

The slides are available on our webpage
https://ds.ifi.uni-heidelberg.de/teaching/current/