Software Practicals
Summer Semester 2024

Data Science Group
Heidelberg University
April 17, 2024

Slides Online

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

Organization

Outline

- Overview of topics (today)
    - Send application for a topic until **Monday, April 22, 1pm**
    - Assignment of topics by April 25
- First milestone (end of May)
    - Prototype / part of software
    - Summary of research (literature and related systems/tools)
    - Further milestones in agreement with supervisor
- End of practical (mid/end July)
    - Code has to be in local Gitlab of the Data Science 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”
        - Job experience
    - Send your tentative schedule and milestones for the practical
    - Group work is not possible!
- It is recommended to apply for multiple topics (“top-3 list”)

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

Deadlines

- Generally meetings with supervisor every week
- Presentation: last week of July 2024
- Report & Gitlab upload: August 5, 2024
- No extension possible

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

Assessment

- 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

- Ashish Chouhan (AC)
  chouhan@informatik.uni-heidelberg.de

- Nicolas Reuter (NR)
  reuter@informatik.uni-heidelberg.de

- Marina Walther (MW)
  walther@informatik.uni-heidelberg.de

Project Topics

AP = Advanced Topic

BP = Beginners Topic (for BSc students)

Overview of Topics

1. QA System to Unlock Legal Insights, **BP/AP** (Chouhan)
2. ArXiv Abstracts Clustering Analysis, **AP** (Chouhan)
3. Quantity and Concept Extraction with ChatGPT **BP**/**AP** (Gertz)
4. Voice Assistant for Retrieval Augmented Generation, **AP** (Gertz)
5. Apothekenumschau Knowledge Graph, **BP/AP** (Walther)
6. YouTube Video Scraping of Medfluencer Channels, **AP** (Walther)
7. Query Databases using Natural Language, **AP/BP** (Reuter)
8. Evaluation of General Purpose LLMs on Table Data, **BP** (Reuter)

BP/AP: QA System to Unlock Legal Insights (AC)

Given:

- People post information about Legal domain on LinkedIn (see, e.g., Martin Ebers)

Tasks:

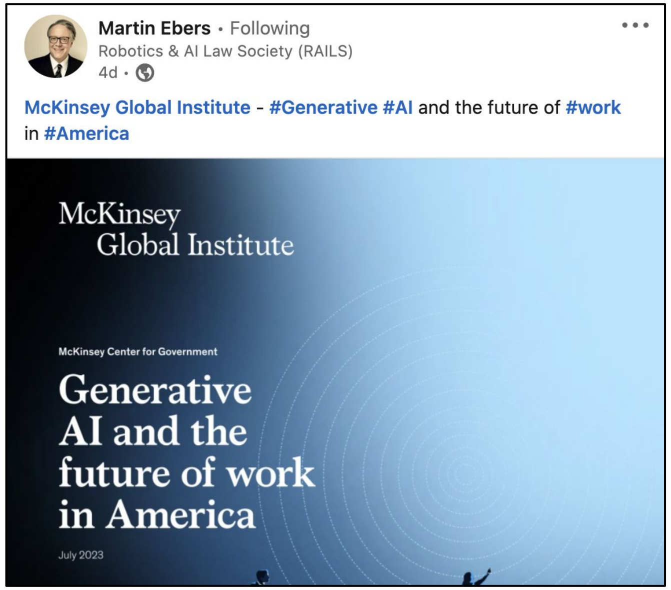
- Extract articles, documents, and posts related to legal domain and store them in OpenSearch

Subtasks:

- Realize backend to store extracted information using OpenSearch
- Question Answering (QA) system to interact with the stored information

Languages / Tools:

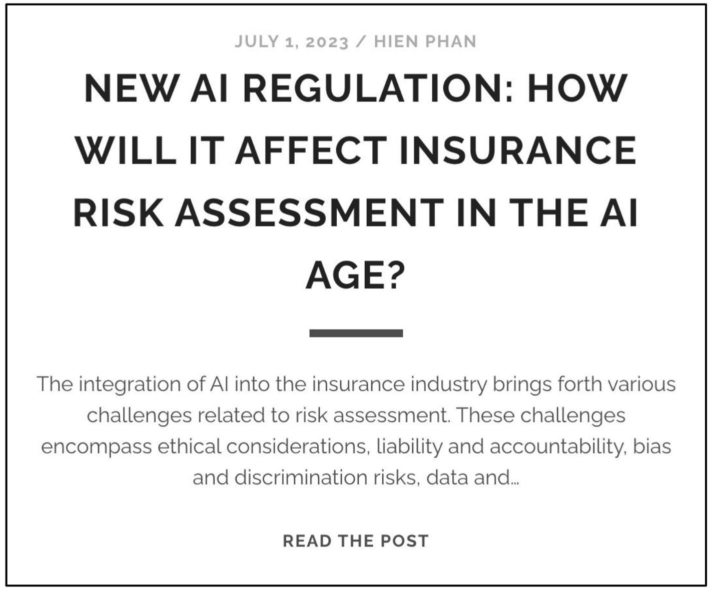
- Python, OpenSearch, LangChain or Llamaindex, Hugging Face



Martin Ebers • Following
Robotics & AI Law Society (RAILS)
4d •

McKinsey Global Institute - #Generative #AI and the future of #work in #America

McKinsey
Global Institute
McKinsey Center for Government
Generative AI and the future of work in America
July 2023



JULY 1, 2023 / HIEN PHAN

NEW AI REGULATION: HOW WILL IT AFFECT INSURANCE RISK ASSESSMENT IN THE AI AGE?

The integration of AI into the insurance industry brings forth various challenges related to risk assessment. These challenges encompass ethical considerations, liability and accountability, bias and discrimination risks, data and...

READ THE POST

Link

AP: ArXiv Abstract Clustering Analysis (AC)

Given:

- ArXiv corpus
- Reference Paper, Demo

Tasks:

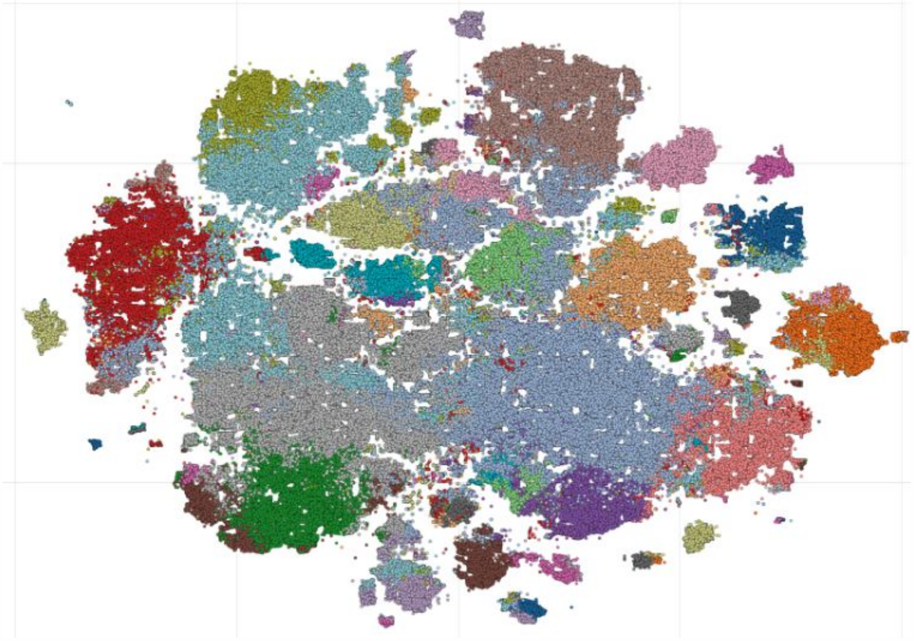
- Develop a pipeline in Python to create an arXiv corpus
- Identify groups using clustering algorithms

Subtasks:

- Realize backend to store extracted text using OpenSearch
- Visualization and exploration of clusters

Languages / Tools:

- Python, OpenSearch, Hugging Face, SvelteKit



Link

BP: Quantity and Concept Extraction with ChatGPT (MG)

Given:

- Sentences from news articles that contain quantity information

Tasks:

- Use ChatGPT to extract quantity information for a sentence in the form <value, unit, change, concept>
- “SAP lays off 2500 employees”
  → value=2500, unit=employees, change=equal, concept=SAP.

Subtasks:

- Examine ChatGPT’s ability to standardize values, normalize unit, and find relevant nouns (concepts); employ few-shot learning
- Build a Streamlit frontend for text input and output

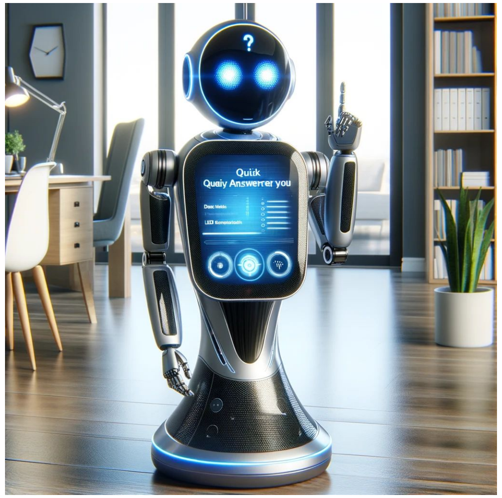
Languages / Tools / Platforms:

- Python, LangChain, Streamlit, OpenAI

AP: Voice Assistant for **R**etrieval **A**ugmented **G**eneration (MG)

Given:

- Standard pipeline for RAG-based questions answering over some text corpus, including Web frontend



Task:

- Instead of typing questions, users should use voice interface to interact with RAG system
- Design and deploy (open source) voice assistant

Subtasks:

- Develop framework to integrated different open source components into RAG pipeline
- Deploy and evaluate different components

Languages / Tools / Platforms:

- Python, LangChain or Llamaindex

BP/AP: Apothekenumschau Knowledge Graph (MW)

Given:

- Corpus of German health information extracted from https://www.apotheken-umschau.de/

Task:

- Construct knowledge graph, for example, using OpenAI functions
- Evaluate resulting knowledge graph

Subtasks:

- Implement information extraction pipeline using LangChain and an open source graph database, for example Neo4j
- Analyze, explore, and visualize knowledge graph

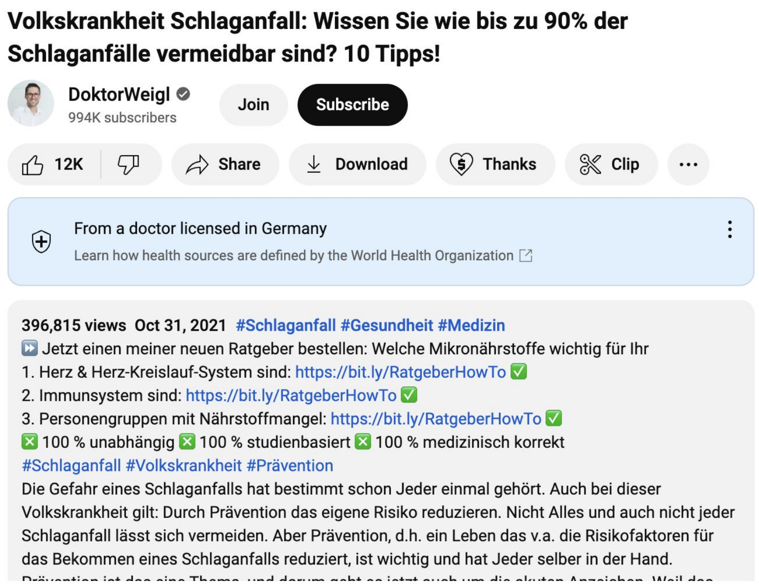
Languages / Tools / Platforms:

- Python, Neo4j, LangChain

AP: YouTube Video Scraping of Medfluencer Channels (MW)

Given:

- Medical and health education for non-professionals is largely provided in social media as videos, e.g., Doktor Weigl.



Volkskrankheit Schlaganfall: Wissen Sie wie bis zu 90% der Schlaganfälle vermeidbar sind? 10 Tipps!

DoktorWeigl
994K subscribers

From a doctor licensed in Germany
Learn how health sources are defined by the World Health Organization

396,815 views Oct 31, 2021 #Schlaganfall #Gesundheit #Medizin
⏩ Jetzt einen meiner neuen Ratgeber bestellen: Welche Mikronährstoffe wichtig für Ihr
1. Herz & Herz-Kreislauf-System sind: https://bit.ly/RatgeberHowTo ✅
2. Immunsystem sind: https://bit.ly/RatgeberHowTo ✅
3. Personengruppen mit Nährstoffmangel: https://bit.ly/RatgeberHowTo ✅
❎ 100 % unabhängig ❎ 100 % studienbasiert ❎ 100 % medizinisch korrekt
#Schlaganfall #Volkskrankheit #Prävention
Die Gefahr eines Schlaganfalls hat bestimmt schon Jeder einmal gehört. Auch bei dieser Volkskrankheit gilt: Durch Prävention das eigene Risiko reduzieren. Nicht Alles und auch nicht jeder Schlaganfall lässt sich vermeiden. Aber Prävention, d.h. ein Leben das v.a. die Risikofaktoren für das Bekommen eines Schlaganfalls reduziert, ist wichtig und hat Jeder selber in der Hand.
Prävention ist das eine Thema, und darum geht es jetzt auch um die akuten Anzeichen. Weil das

Task:

- Scrape video descriptions, comments, and transcriptions and store it in OpenSearch

Subtasks:

- Construct document model and realize backend to store extracted information using OpenSearch
- Question Answering (QA) system to interact with the stored information

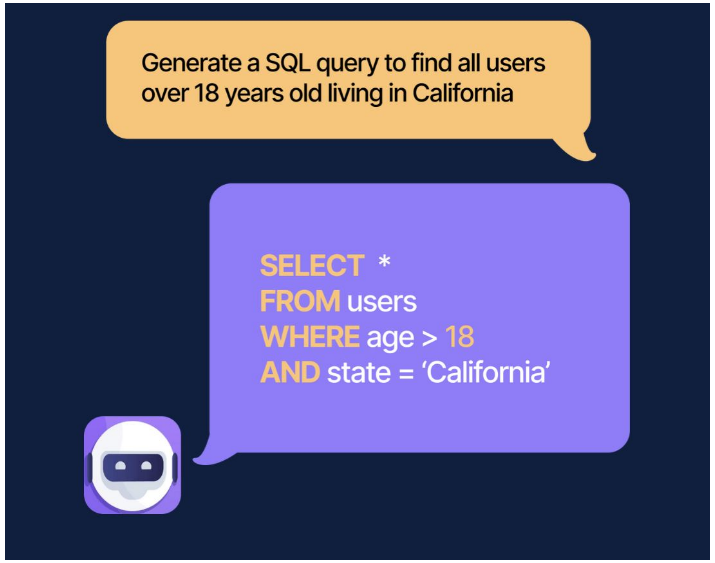
Languages / Tools / Platforms:

- Python, OpenSearch, Selenium, LangChain or Llamaindex

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

Given:

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



Generate a SQL query to find all users over 18 years old living in California

```
SELECT *
FROM users
WHERE age > 18
AND state = ‘California’
```

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

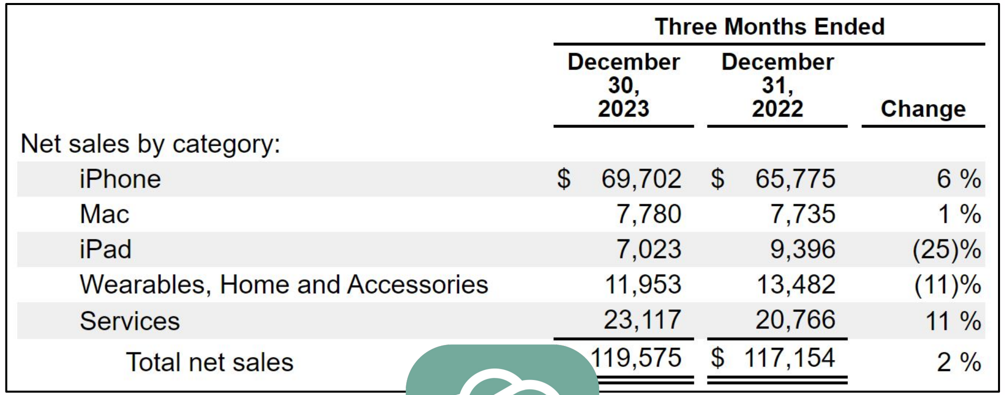
Subtasks:

- 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

Slides Online

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