Web2Table: Using LLMs to Generate Tables from the Web

The Web contains information about arbitrary topics, including:

- products, hotels, recipes, events, organizations, …

It is often desirable to combine data from multiple web sources into a clean and comprehensive table.

Required steps:

1. Retrieve relevant web pages
2. Decide on the schema for the table
3. Extract data according to schema
4. Combine data from multiple web pages into a single table

"Generate a table containing hotels that are close to Frankfurt fair."

Hotel Name	Street	City
Radisson Blu Hotel	Franklinstraße 65	Frankfurt
Holiday Inn – the niu	Leonardo-da-Vinci-Allee 30	Frankfurt
LyvInn Hotel Frankfurt Messe	Europa-Allee 64	Frankfurt

Question: How good are LLMs at table generation from the Web?

Benchmarking LLM-based Table Generation

Project Goal

Develop a benchmark for testing the table generation capabilities of LLMs.

Involves

- Design LLM test cases that require content retrieval and table generation based on the retrieved documents.
- Model evaluation and error analysis

Questions to answer:

- How good are LLMs at understanding table creation queries?
- How good are LLMs at each task of the table creation pipeline?

Requirements

- Programming skills in Python
- Relevant courses: Data Mining I, Text Analytics recommended

Organization: 4-9 people, 5 months

Instructors: Alexander Brinkmann, Christian Bizer