TaxoGen:
Innovating WebIsA Taxonomies

Team Project FSS 2025

UNIVERSITY OF MANNHEIM
Data and Web Science Group

Motivation

UNIVERSITY OF MANNHEIM
Data and Web Science Group

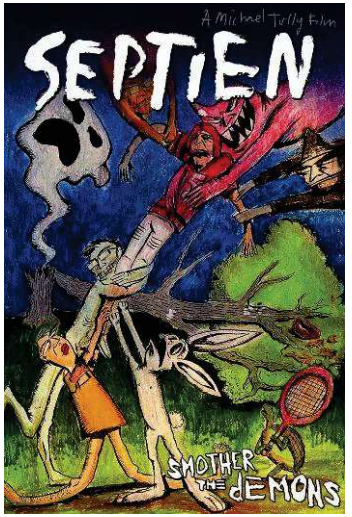
- Many systems need to know type of entities

row_id	name	description	director
0	Septien	Writer-director	{'name': 'Michael Tully'}
1	The Finest Hours	Recounting one	{'name': 'Craig Gillespie'}
2	United	A devastating pl	{'name': 'James Strong'}
3	The Haunting in Con	In this supernat	{'name': 'Peter Cornwell'}

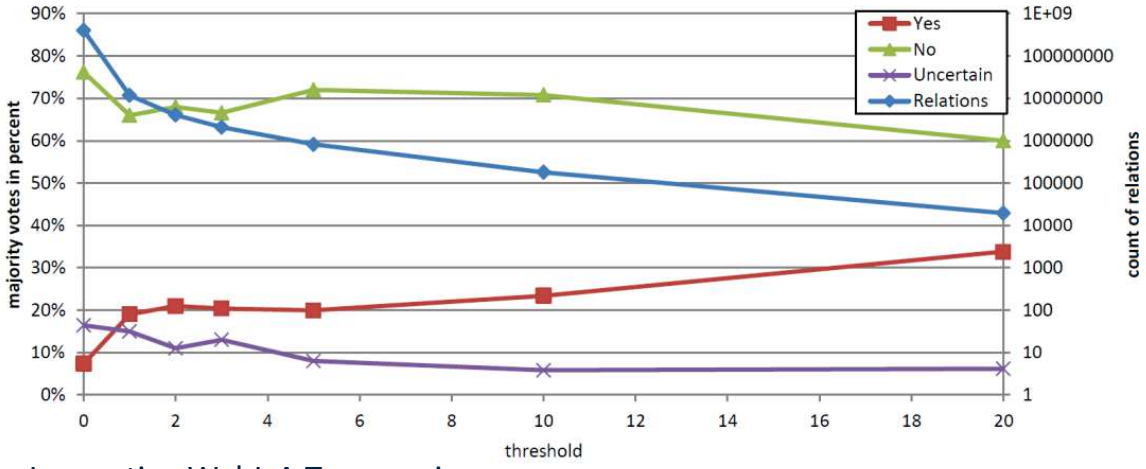
movies

Motivation

UNIVERSITY OF MANNHEIM
Data and Web Science Group



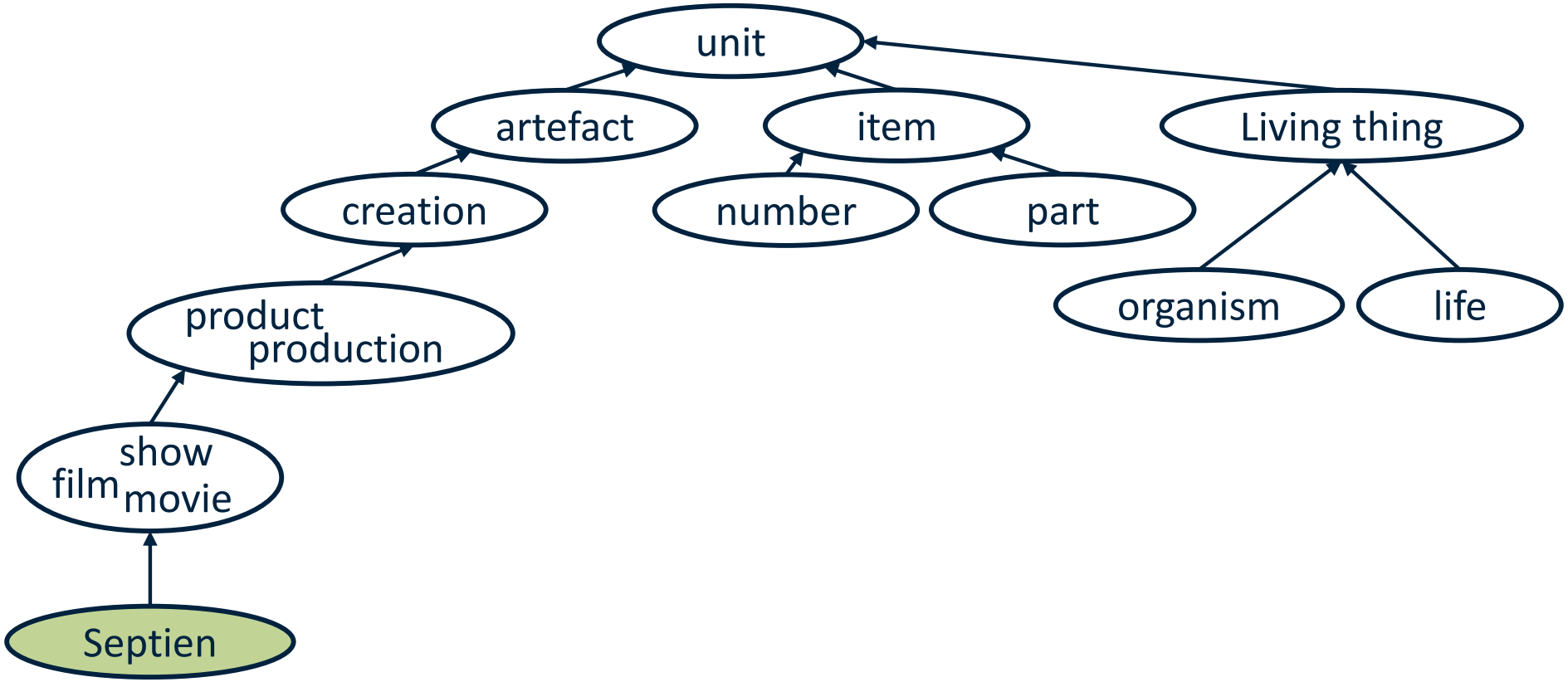
- WebIsaDB
  - Analyzed large web crawl with Hearst patterns
    - „Septien and other movies“ -> Septien ISA movie
  - 400,533,808 relations extracted (2.1 billion pages)
  - Collected information
    - pre modifier, head noun, post modifier for hypernym and hyponym
    - set of patterns, sentences, pay-level domains
    - absolute number of occurrences



Idea

UNIVERSITY OF MANNHEIM
Data and Web Science Group

- Use WordNet as top level taxonomie



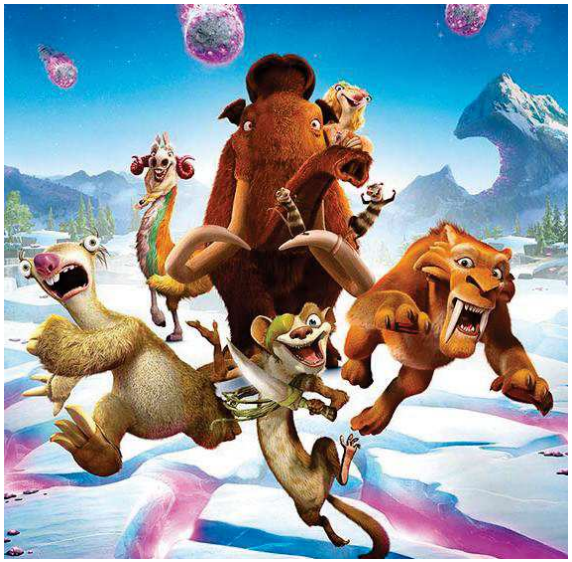
- Optional: Redo the extraction with improved methods

Organisational information

UNIVERSITY OF MANNHEIM
Data and Web Science Group

- The project runs for one semester (six months)
- 3-4 students
- There will be regular meetings
  - Some in person, most can be online
- Computing infrastructure is provided by the DWS group

- **Required skills of the team**
  - Good programming skills e.g., Python or Java
  - Optionally: attended (or attending)
    - Knowledge Graphs (Semantic Web Technologies)
    - Data Mining (or comparable ML course)



Supervisors

UNIVERSITY OF MANNHEIM
Data and Web Science Group

Dr. Sven Hertling
Substitute Professor for Data Science

E-Mail: sven.hertling@uni-mannheim.de

Prof. Dr. Heiko Paulheim
Chair for Data Science

E-Mail: heiko.paulheim@uni-mannheim.de

Thank you

UNIVERSITY OF MANNHEIM
Data and Web Science Group

- Hopefully you are not afraid