Project DIS18

04 – Projects

Jüri Keller and Philipp Schaer

24-11-29 – Cologne, Germany
https://ir.web.th-koeln.de

Technology
Arts Sciences
TH Köln

Provisional Schedule

	Datum	Notiz	
W1	18.10.24		
W2	25.10.24		Einführung zu Temporal IR
W3	01.11.24	Entfällt (Allerheiligen)	
W4	07.11.24		Queries and Topics
W5	15.11.24		Topics selbst erstellen
W6	22.11.24	Projektwoche	
W7	**29.11.24**		**Relevanzbewertungen und Projekte**
W8	06.12.24		Projektphase
W9	13.12.24		...
W10	20.12.24		Projekte präsentieren
		Weihnachten	
W11	03.01.25		Home Office
W12	10.01.25		LongEval
W13	17.01.25		

TASK:

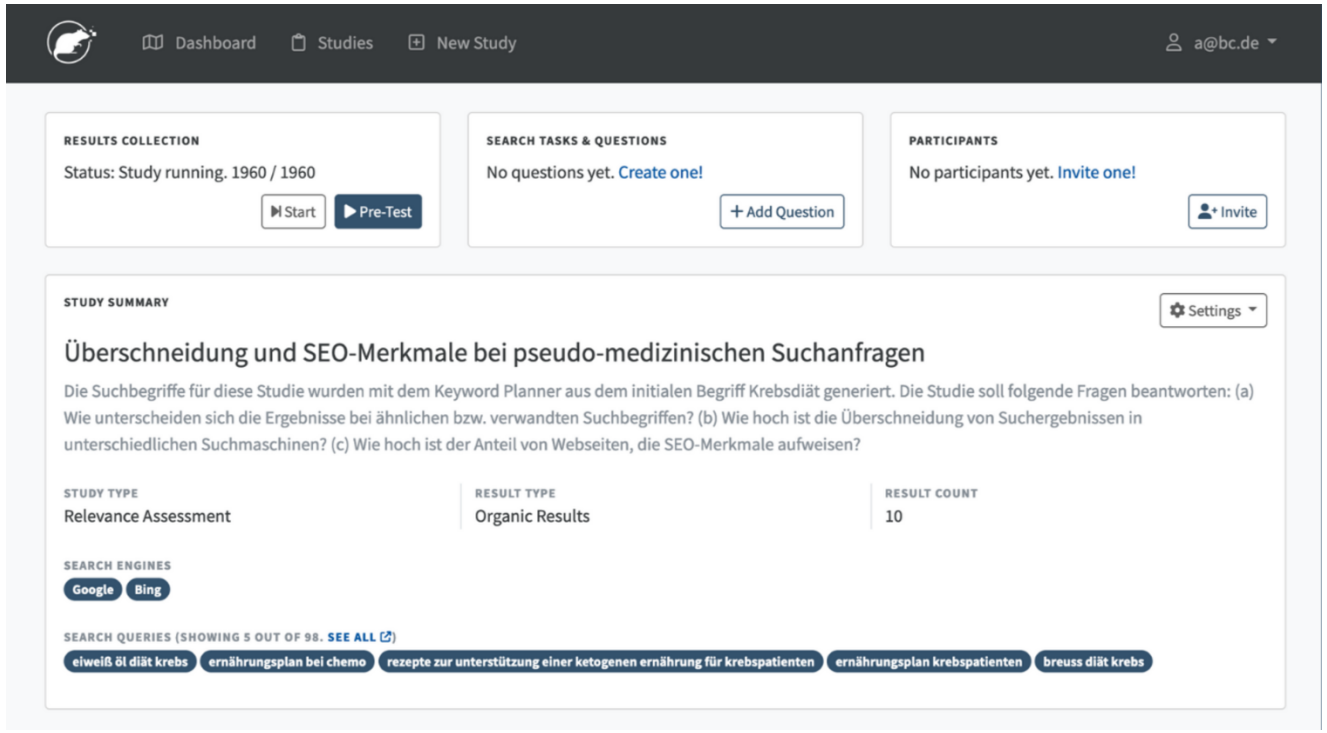
1. Überarbeiten Sie ihre eigenen Topics nach den gemeinsam vereinbarten Richtlinien
2. Bewerten Sie die ersten 10 Google Ergebnisse für ihre eigenen Topics
3. Bewerten Sie die ersten 10 Google Ergebnisse für die ihnen zugeteilten Topics
4. Geben Sie Feedback zu den zugeteilten Topics
5. Überarbeiten Sie ihre Topics auf Basis des Feedbacks

Laden Sie die überarbeiteten Topics bis spätestens zum **22.11.24** in das GitHub Repository hoch.

Result Assessment Tool (RAT)

RAT
RESULT ASSESSMENT TOOL

1. Studie erstellen
2. Ergebnisse sammeln
3. Bewerten & Analysieren
4. Ergebnisse exportieren

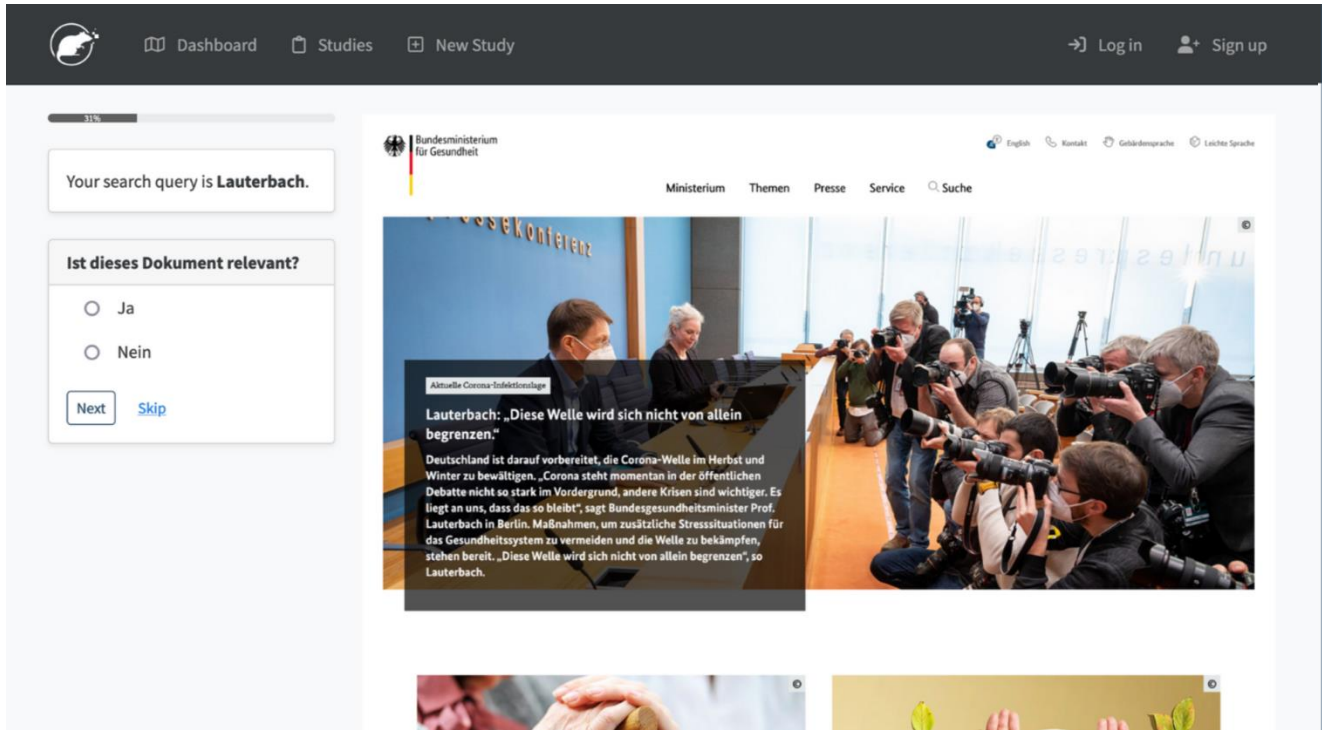


Source: https://searchstudies.org/research/rat/

Result Assessment Tool (RAT)

RAT
RESULT ASSESSMENT TOOL

1. Studie erstellen
   1. Eine Studie pro Person
2. Ergebnisse sammeln
   - Top 10 Google Results
3. Bewerten & Analysieren
   - Not Relevant (0)
   - Relevant (1)
   - Highly relevant (2)
4. Ergebnisse exportieren
   - Wöchentlicher Export



Source: https://searchstudies.org/research/rat/

TASK:

1. Annotieren Sie einmal in der Woche die Relevanz für die ersten 10 Google Ergebnisse für Ihre selbst erstellten Topics nach der Guideline.

GUIDELINE

Highly relevant (2): Die Website entspricht voll und ganz dem Informationsbedürfnis, welches durch das Topic ausgedrückt wird, d. h. sie beantwortet die Frage im Topic. Die Website muss nicht alle Informationen zum Topic enthalten, aber sie muss für sich genommen eine Antwort auf die Frage geben.

Relevant (1): Die Website beantwortet einen Teil der Frage, müsste aber mit anderen Informationen kombiniert werden, um eine vollständige Antwort zu erhalten.

Not relevant (0): alles andere

Projekte

1. Klassifizieren von temporalen Querys mit LLMs
2. Literaturrecherche zu Trendanalyse und Klassifikation
3. Klassifizieren von temporalen Querys durch Google Trends
4. Queries clustern basierend auf dem Querytext
5. Queries clustern basierend auf relevanten Dokumenten
6. Implementierung eine RAT Scrapers für weitere Suchmaschinen

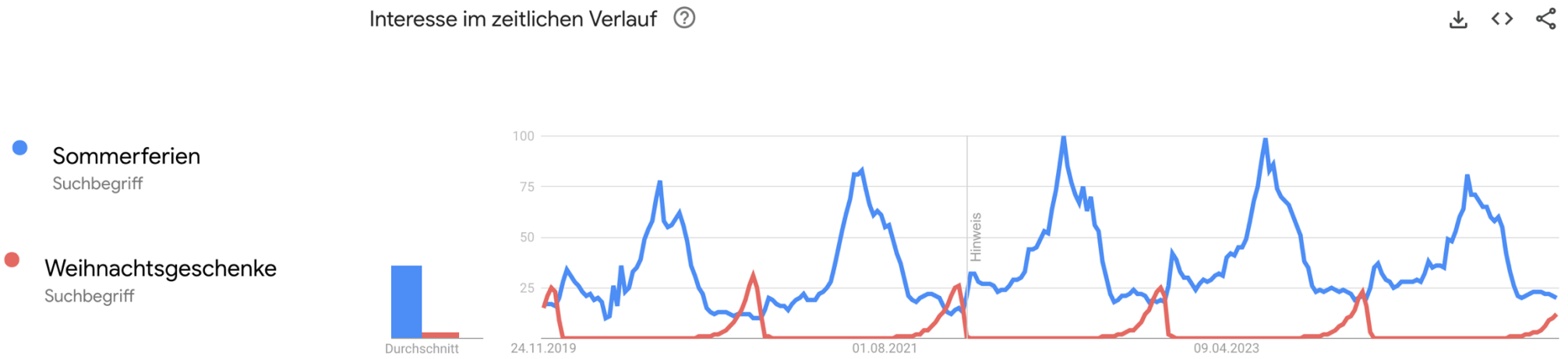
1. Klassifizieren von temporalen Querys mit LLMs

- Gemeinsam erstellte Taxonomie als Grundlage für einen Prompt
- Prompt tuning: Single vs. multi-shot
- Implementierung zur Automatisierung der Klassifikation
- Vergleich zu unseren eigenen Ergebnissen

1. Nicht Zeitlich	2. Explizit Zeitlich	3. Ereignis	4. Mehrdeutig	5. Aktualität

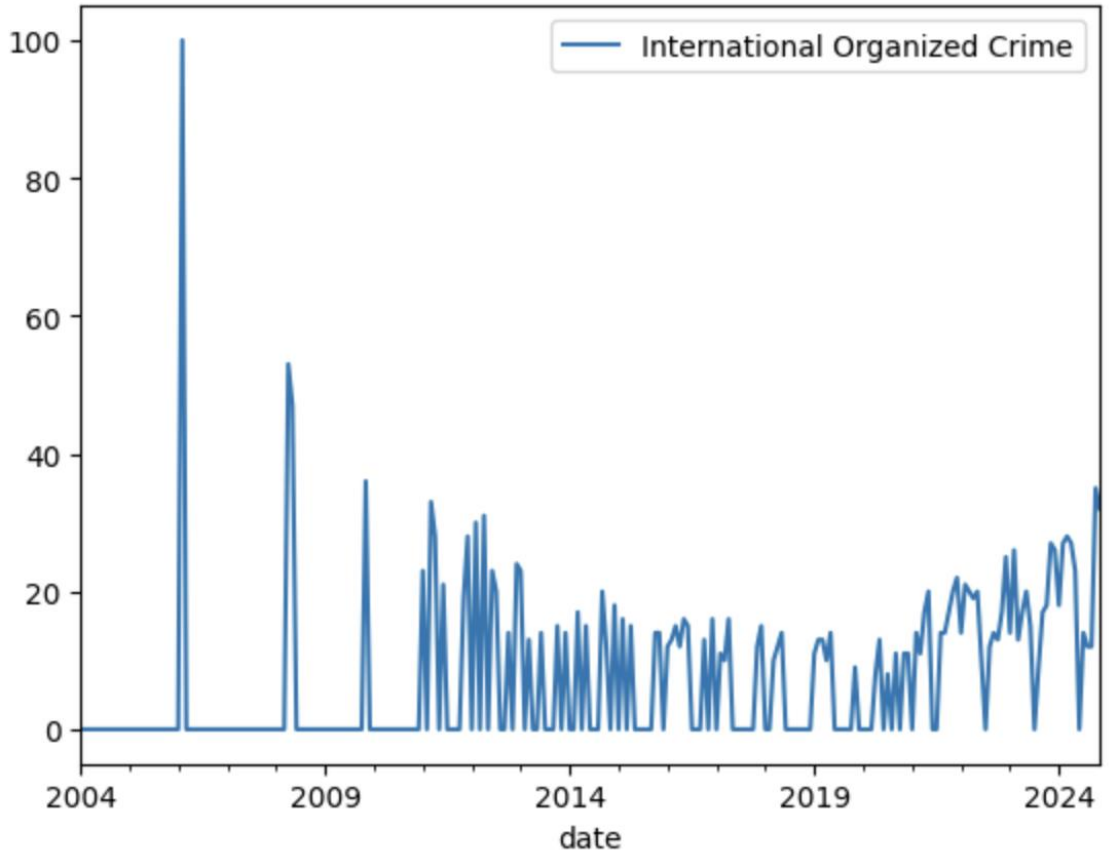
2. Literaturrecherche zu Zeitreihenanalyse und Klassifikation

- Wie werden Querytrends in der Literatur analysiert?
- Welche Methoden werden zur Analyse oder Klassifikation von Zeitreihen verwendet?
- Welche Muster werden in der Literatur genannt?
- Wie lassen diese sich identifizieren?



3. Klassifizieren von temporalen Querys durch Google Trends

- Scrapen der Daten von Google Trends über externe APIs oder toolkits
  - https://serpapi.com/google-trends-api
  - https://github.com/GeneralMills/pytrends
- Aggrigation und Visualisieren der Zeitreihen
- Clustering und Klassifikation der Zeitreihen



4. Queries clustern basierend auf dem Querytext

- Viele Queries sind sehr ähnlich
- Mehrere Querys drücken dasselbe Suchinterresse aus. Sie sind also Queryvarianten eines Topics
- Basierend auf dem Querytext können diese Queries zusammengefasst werden
  - Text Normalisierung
  - Text Embeddings

- gateau a la banane
- gateau **à** la banane
- g**â**teau **à** la banane
- g**â**teau a la banane

- gateau
- gateau recipe
- gateau chocolat banane
- gateau au cacao

4. Queries clustern basierend auf relevanten Dokumenten

- Die **relevanten Dokumente** können zusätzlich Aufschluss über die Ähnlichkeit von Querys geben

große Ähnlichkeit

Q: **gateau a la banane**

- Doc A
- Doc B
- Doc C

Q: **gateau chocolat banane**

- Doc A
- Doc B
- Doc D

geringe Ähnlichkeit

Q: **gateau au cacao**

- Doc D
- Doc E
- Doc F

Q: **gateau chocolat banane**

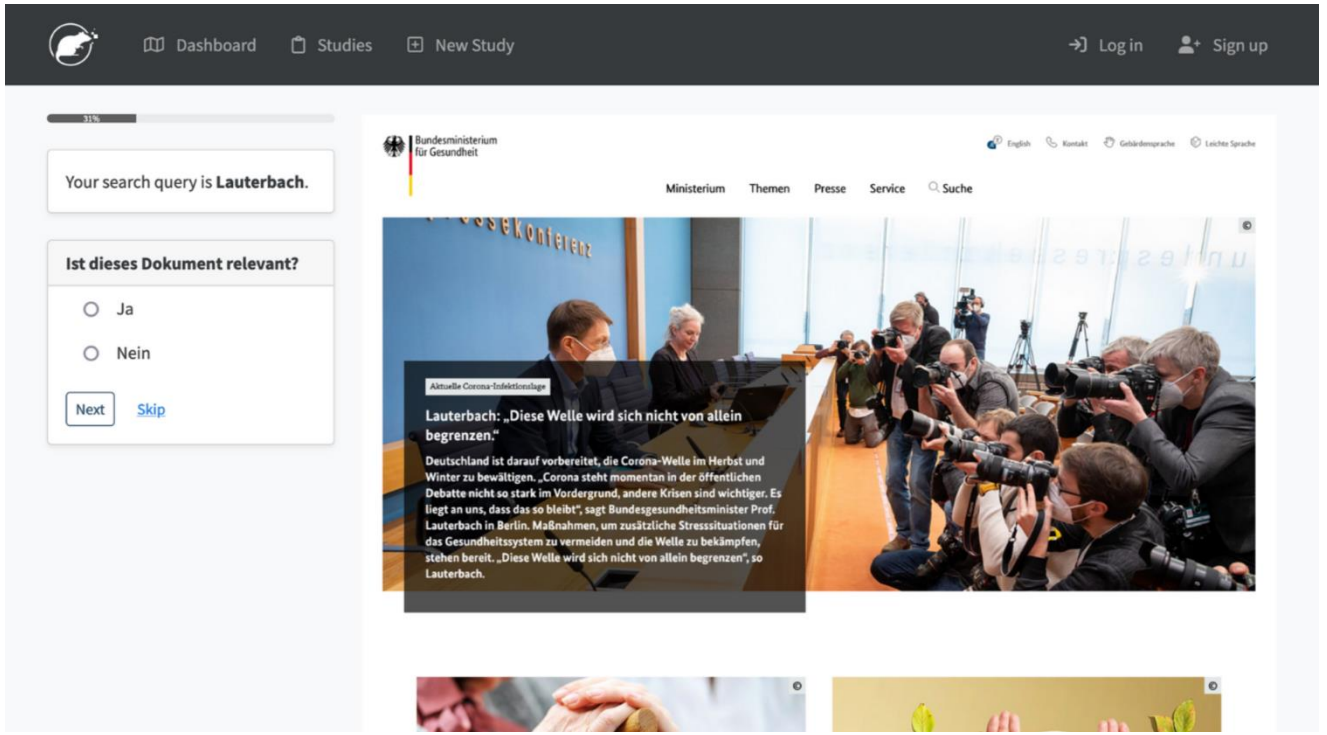
- Doc A
- Doc B
- Doc D

6. Implementierung eine RAT Scrapers für weitere Suchmaschinen

- Das RAT zeigt die Suchergebnisse von ausgewählten Suchmaschinen
- Um Weitere Suchmaschinen verwenden zu können müssen diese automatisiert gescraped werden können.

Suchmaschienen:
CORE: https://core.ac.uk/
EconBiz: https://www.econbiz.de/

GitHub:
https://github.com/rat-software/rat-scrapers



TASK:

1. Annotieren Sie **einmal in der Woche** die Relevanz für die ersten 10 Google Ergebnisse für Ihre selbst erstellten Topics nach der Guideline.

2. Arbeiten Sie **selbstständig** in ihrer Gruppe and den Projekten
   - Stellen Sie am **6.12.24** ersten Ergebnisse und den weiteren Plan vor

How to pass this project

Hard requirements

- Be. Active. And. Participate. In. The. Project.
- Active means to code, label, document, plan, organize, ...
- Milestone presentations after 1st semester
- Do final presentation at the end of the course (2nd semester)
- Write a final term paper / project documentation (and maybe submit it to CLEF).

Soft requirements

- Be. Active. And. Have. Fun. In. The. Project.
- Learn a lot new and interesting stuff about Python, Data Labeling, project management, test collection design, search engines, and other DIS-related topics.

Provisional Schedule

	Datum	Notiz	
W1	18.10.24		
W2	25.10.24		Einführung zu Temporal IR
W3	01.11.24	Entfällt (Allerheiligen)	
W4	07.11.24		Queries and Topics
W5	15.11.24		Topics selbst erstellen
W6	22.11.24	Projektwoche	
W7	**29.11.24**		**Relevanzbewertungen und Projekte**
W8	06.12.24		Projektphase
W9	13.12.24		...
W10	20.12.24		Projekte präsentieren
		Weihnachten	
W11	03.01.25		Home Office
W12	10.01.25		LongEval
W13	17.01.25		

Provisional Schedule

	Datum	Notiz	
W1	24.03		
W2	07.04		
W3	14.04		
W4	21.04		
W5	28.04		
W6	05.05	*Submission Runs*	LongEval Runs einreichen
W7	12.05		
W8	19.05		
W9	26.05	*Submission Notebooks*	LongEval Bericht einreichen
W10	02.06		Abschlusspräsentation
W11	09.06		entfällt
W12	16.06		entfällt
W13	23.06		entfällt