Software Practicals
Summer Semester 2023

Database Systems Research Group
Heidelberg University
April 19, 2023

Slides Online

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

Organization

Outline

- Overview of topics (today)
  - Send application for a topic until **Monday, April 24, 1pm**
  - Assignment of topics by April 27
- 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 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”
    - 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 2023
- Report & Gitlab upload: August 7, 2023
- No extension possible

Not finished = failed (grade 5,0)!

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
- Satya Almasian (SA)
  almasian@informatik.uni-heidelberg.de
- Jayson Salazar (JS)
  salazar@informatik.uni-heidelberg.de
- John Ziegler (JZ)
  ziegler@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)

Overview of Topics

1. Interface for Quantity Extractor, **AP** (Almasian)
2. Quantity and Concept Extraction with ChatGPT, **BP** (Gertz/Almasian)
3. Extracting Scientific Documents from Wikipedia, **BP** (Almasian)
4. Table of Content Crawler for Proceedings, **AP** (Gertz)
5. Office Document Reader and Analyzer, **BP**/**AP** (Gertz)
6. Trend Exploration UI, **AP** (Ziegler)
7. Dynamic Network Exploration, **AP** (Ziegler)
8. Acquisition, Analysis and NER on OPS Codes **AP/BP** (Salazar)
9. Package Integration for NER in Patient Records **BP/AP** (Salazar)
10. Crawler and Analyzer for PubMed Article Full Text, **BP** (Chouhan)
11. Aspect Based Temporal Clustering, **AP** (Chouhan)
12. Machine Learning in Web Browsers, **AP/BP** (Reuter)
13. Creating a Domain-Aware Web Table Corpus, **AP** (Reuter)

AP: Interface for Quantity Extractor (SA)

Given:

- A package implemented in previous practicals to identify and normalize quantities in the text
- "The tower is 100m high"
  → value:100, unit: metre, concept: tower, change: equal

Tasks:

- Create a web interface to interact with the package

Subtasks:

- Interface for on-demand extraction on the website
- Possibility of uploading documents and getting extractions as XML

Languages / Tools:

- Python; Flask; frontend development skills (css, JS, svelte…)

AP: Quantity and Concept Extraction with ChatGPT (MG/SA)

Given:

- Sentences from news articles tagged with quantity information

Tasks:

- Use ChatGPT to extract quantity information for a sentence, containing: (value, unit, change, concept)
- Apple hires 200 people:
  → value=200, unit=people, change=equal, concept=Apple.

Subtasks:

- Examine ChatGPT's ability to standardize values, normalize unit and find relevant nouns (concepts)
- Build a pipeline that gets sentences and outputs quantity information

Languages / Tools:

- Python

BP: Extracting Scientific Documents from Wikipedia (SA)

Given:

- List of “quantity heavy” topics from Wikipedia, e.g., physics, medicine or math.

Tasks:

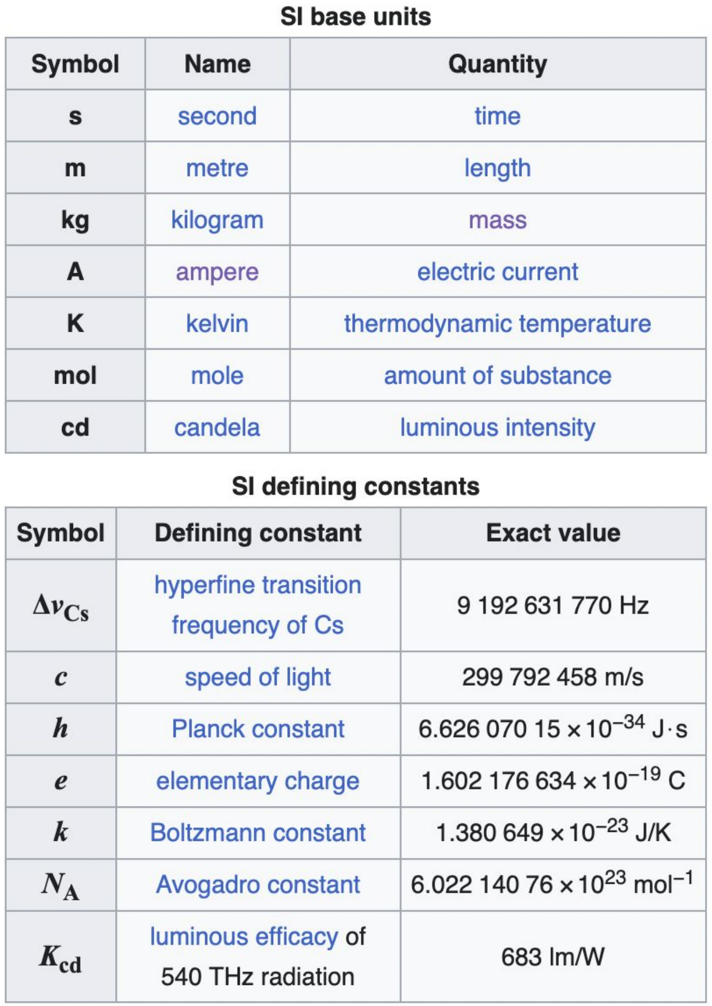
- Extract the pages and text data associated with these topics
- Analyze quantity extractions

Subtasks:

- Investigate different quantity types (statistics)
- Store pages and quantity extraction results
- in OpenSearch

Languages / Tools:

- Python, OpenSearch



SI base units

Symbol	Name	Quantity
s	second	time
m	metre	length
kg	kilogram	mass
A	ampere	electric current
K	kelvin	thermodynamic temperature
mol	mole	amount of substance
cd	candela	luminous intensity

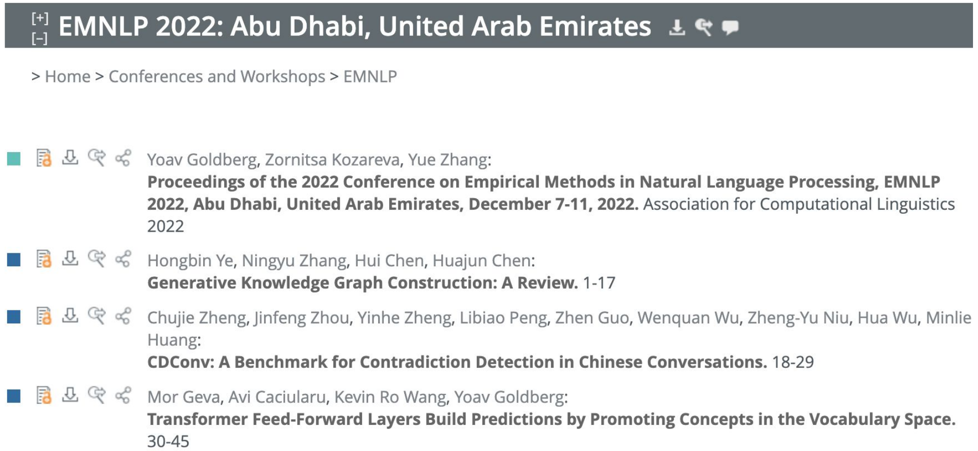
SI defining constants

Symbol	Defining constant	Exact value
$\Delta\nu_{\text{Cs}}$	hyperfine transition frequency of Cs	9 192 631 770 Hz
c	speed of light	299 792 458 m/s
h	Planck constant	$6.626\,070\,15 \times 10^{-34}$ J⋅s
e	elementary charge	$1.602\,176\,634 \times 10^{-19}$ C
k	Boltzmann constant	$1.380\,649 \times 10^{-23}$ J/K
N_{A}	Avogadro constant	$6.022\,140\,76 \times 10^{23}$ mol^{-1}
K_{cd}	luminous efficacy of 540 THz radiation	683 lm/W

AP: Table of Content Crawler for Proceedings (MG)

Given:

- Many web sites list papers accepted for conferences in unstructured fashion (see, e.g., https://aclanthology.org/)



[+]
[–] EMNLP 2022: Abu Dhabi, United Arab Emirates

> Home > Conferences and Workshops > EMNLP

Yoav Goldberg, Zornitsa Kozareva, Yue Zhang:
Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022. Association for Computational Linguistics 2022

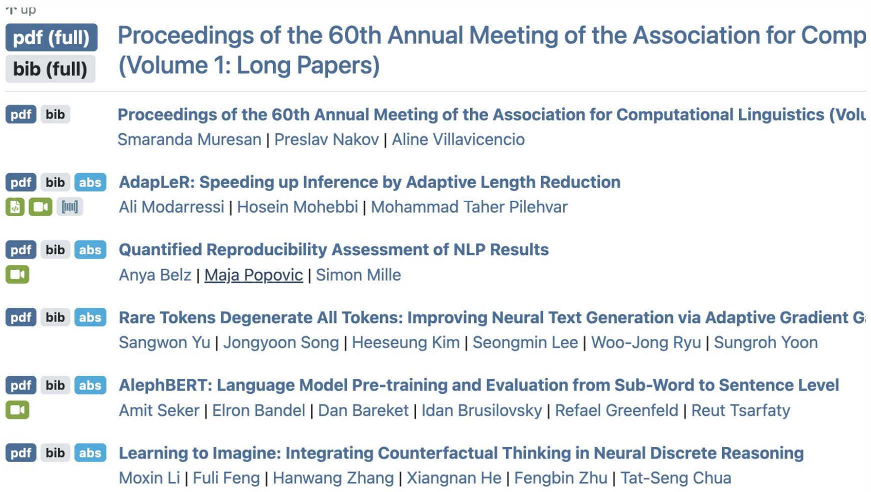
Hongbin Ye, Ningyu Zhang, Hui Chen, Huajun Chen:
Generative Knowledge Graph Construction: A Review. 1-17

Chujie Zheng, Jinfeng Zhou, Yinhe Zheng, Libiao Peng, Zhen Guo, Wenquan Wu, Zheng-Yu Niu, Hua Wu, Minlie Huang:
CDConv: A Benchmark for Contradiction Detection in Chinese Conversations. 18-29

Mor Geva, Avi Caciularu, Kevin Ro Wang, Yoav Goldberg:
Transformer Feed-Forward Layers Build Predictions by Promoting Concepts in the Vocabulary Space. 30-45

Tasks:

- Develop crawler that extracts paper information from web sites
- Develop frontend that allows to search, explore, and cluster papers



Proceedings of the 60th Annual Meeting of the Association for Comp
(Volume 1: Long Papers)

Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volu
Smaranda Muresan | Preslav Nakov | Aline Villavicencio

AdapLeR: Speeding up Inference by Adaptive Length Reduction
Ali Modarressi | Hosein Mohebbi | Mohammad Taher Pilehvar

Quantified Reproducibility Assessment of NLP Results
Anya Belz | Maja Popovic | Simon Mille

Rare Tokens Degenerate All Tokens: Improving Neural Text Generation via Adaptive Gradient G
Sangwon Yu | Jongyoon Song | Heeseung Kim | Seongmin Lee | Woo-Jong Ryu | Sungroh Yoon

AlephBERT: Language Model Pre-training and Evaluation from Sub-Word to Sentence Level
Amit Seker | Elron Bandel | Dan Bareket | Idan Brusilovsky | Refael Greenfeld | Reut Tsarfaty

Learning to Imagine: Integrating Counterfactual Thinking in Neural Discrete Reasoning
Moxin Li | Fuli Feng | Hanwang Zhang | Xiangnan He | Fengbin Zhu | Tat-Seng Chua

Subtasks:

- Design and implement document store based on OpenSearch.

Languages / Tools:

- Python; OpenSearch

BP/AP: Office Document Reader and Analyzer (MG)

Given:

- Office documents based on Office Open XML (.docx, .pptx, .xlsx)

Tasks:

- Develop pipeline to detect type of document and convert it according some document model for search and downstream NLP tasks
- **AP**: convert documents from Opensearch to .docx

Subtasks:

- Design and implement document store based on OpenSearch.

Languages / Tools:

- Python; OpenSearch

AP: Trend Exploration UI (JZ)

Given:

- REST API to access trends
- Twitter trends given as hashtag networks
- Political domain → see EPINetz project

Task:

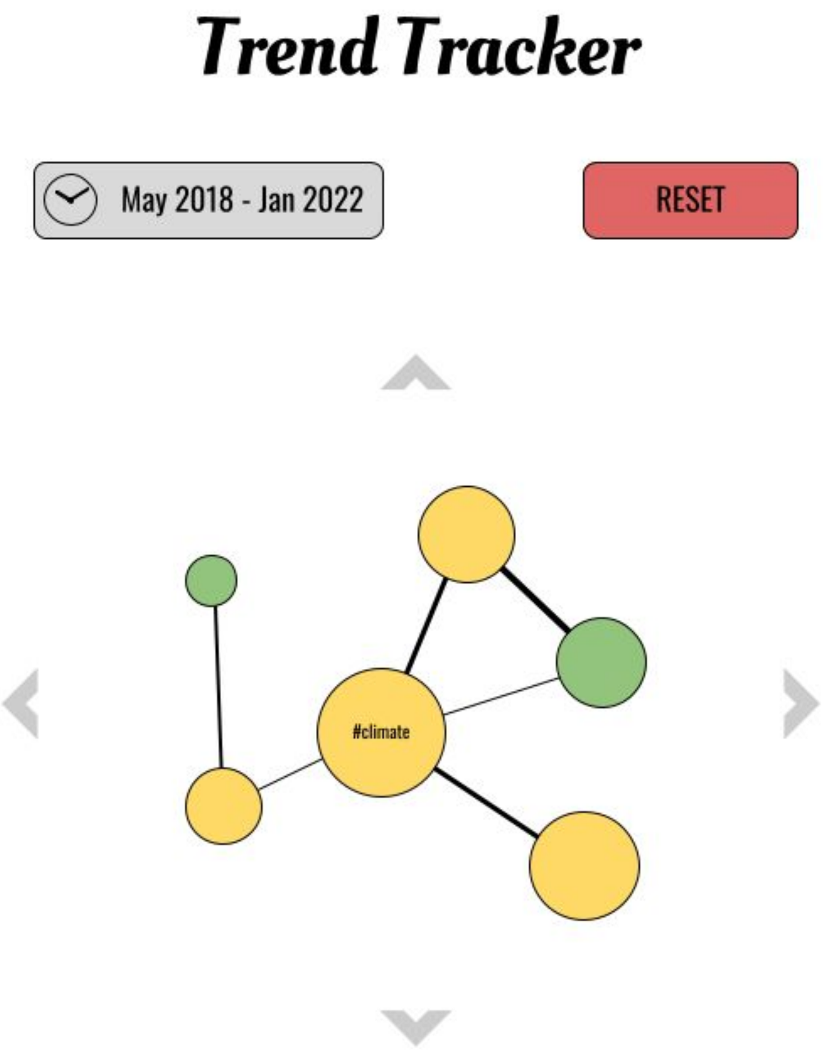
- Extension of existing UI to explore trends

Subtasks:

- Feature to compare trends
- Highlighting of nodes/edges

Languages / Tools:

SvelteKit, Chart.js, D3.js, TypeScript



AP: Dynamic Network Exploration (JZ)

Given:

- Concept of app to visualize and explore network dynamics

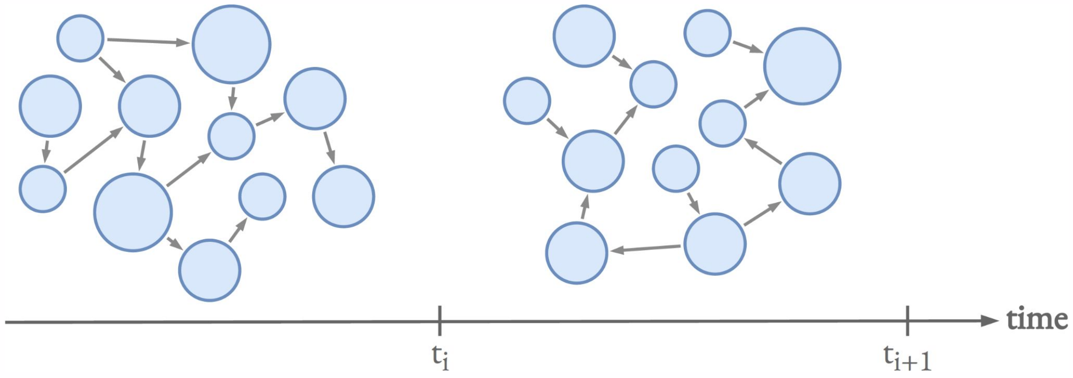
→ Creative ideas are welcome!

Task:

- Software to visually explore temporal networks

Subtasks:

- Data import
- Network visualization
- Timeline exploration feature
- Snapshot sampling



Languages / Tools:

- JS framework (e.g., React), Cytoscape.js, TypeScript

BP(AP): Acquisition, Analysis and NER on OPS Codes (JS)

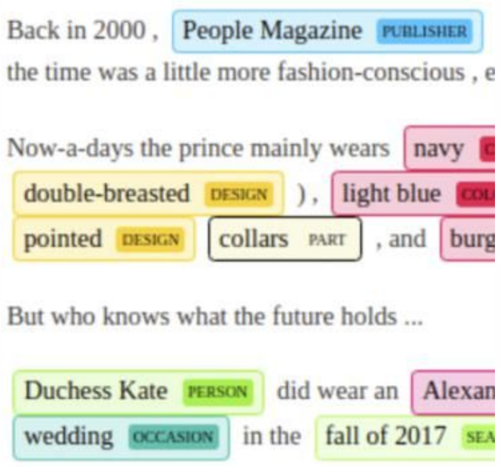
Federal Institute
for Drugs
and Medical Devices

Given:

- Python library for Medical Thesaurus Correlation (MedKEET)
- Access to OPS-Database and related medical thesauri

Tasks:

- Acquire and analyze the structure and contents of the OPS dataset.
- Extend MedKEET to process and annotate (naively) OPS entries

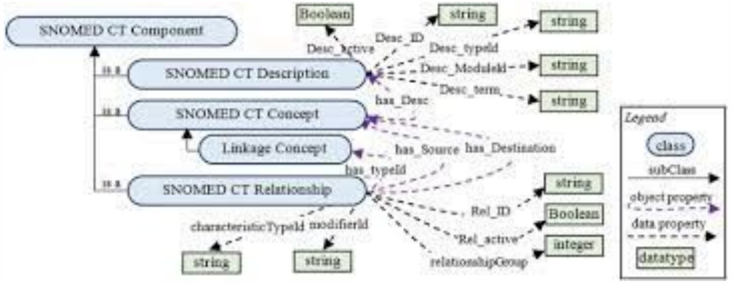


Back in 2000 , People Magazine PUBLISHER
the time was a little more fashion-conscious , e

Now-a-days the prince mainly wears navy C
double-breasted DESIGN), light blue COL
pointed DESIGN collars PART , and burg

But who knows what the future holds ...

Duchess Kate PERSON did wear an Alexan
wedding OCCASION in the fall of 2017 SEA



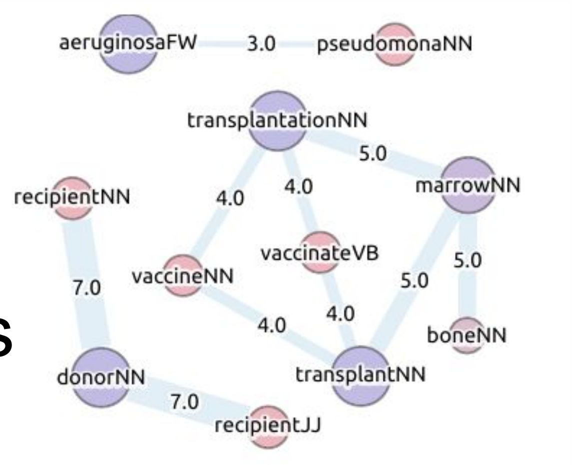
Languages / Tools:

- Python, Pandas, Neo4J

BP(AP): Package Integration for NER in Patient Records (JS)

Given:

- PyCoNet, a prototype package aimed towards the extraction of concept relationships from raw (medical) text.
- FHIRPACK, an open source FHIR data Python processing toolkit
- Both developed jointly in the DBS and the UK-Essen



Tasks:

- Integrate and extend both packages so is-relationships (NER) present in raw, semi-structured patient records can be extracted and stored seamlessly

Languages / Tools:

- Python, PostgreSQL

BP: Crawler and Analyzer for PubMed Article Full Text (AC)

Given:

- Access to PubMed metadata and abstract extraction script

Tasks:

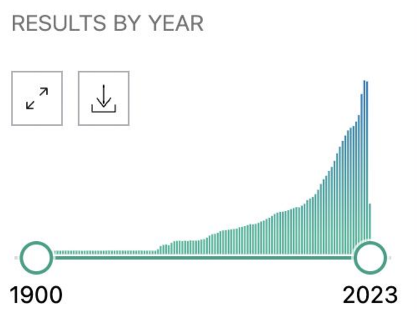
- Extract full text from PubMed and store it in OpenSearch

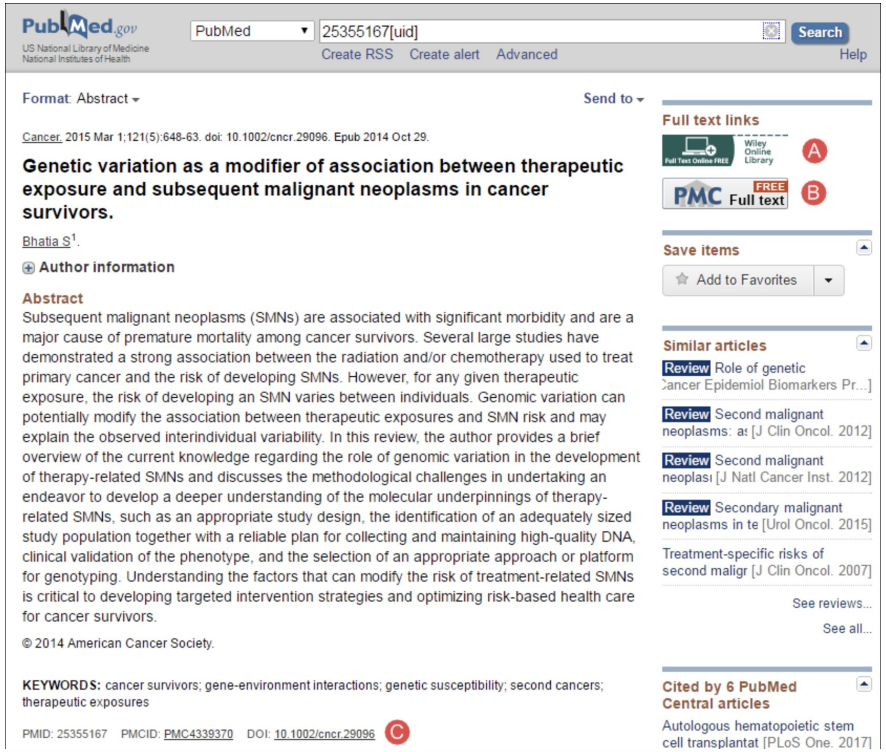
Subtasks:

- Realize backend to store extracted text using OpenSearch
- Analyze the dataset and compute basic properties

Languages / Tools:

- Python, OpenSearch





Cancer. 2015 Mar 1;121(5):648-63. doi: 10.1002/cncr.29096. Epub 2014 Oct 29.

Genetic variation as a modifier of association between therapeutic exposure and subsequent malignant neoplasms in cancer survivors.

Bhatia S[1].

Author information

Abstract

Subsequent malignant neoplasms (SMNs) are associated with significant morbidity and are a major cause of premature mortality among cancer survivors. Several large studies have demonstrated a strong association between the radiation and/or chemotherapy used to treat primary cancer and the risk of developing SMNs. However, for any given therapeutic exposure, the risk of developing an SMN varies between individuals. Genomic variation can potentially modify the association between therapeutic exposures and SMN risk and may explain the observed interindividual variability. In this review, the author provides a brief overview of the current knowledge regarding the role of genomic variation in the development of therapy-related SMNs and discusses the methodological challenges in undertaking an endeavor to develop a deeper understanding of the molecular underpinnings of therapy-related SMNs, such as an appropriate study design, the identification of an adequately sized study population together with a reliable plan for collecting and maintaining high-quality DNA, clinical validation of the phenotype, and the selection of an appropriate approach or platform for genotyping. Understanding the factors that can modify the risk of treatment-related SMNs is critical to developing targeted intervention strategies and optimizing risk-based health care for cancer survivors.

© 2014 American Cancer Society.

KEYWORDS: cancer survivors; gene-environment interactions; genetic susceptibility; second cancers; therapeutic exposures

PMID: 25355167 PMCID: PMC4339370 DOI: 10.1002/cncr.29096

Similar articles

Review Role of genetic ;ancer Epidemiol Biomarkers Pr...]

Review Second malignant neoplasms: a: [J Clin Oncol. 2012]

Review Second malignant neoplası [J Natl Cancer Inst. 2012]

Review Secondary malignant neoplasms in te [Urol Oncol. 2015]

Treatment-specific risks of second maligr [J Clin Oncol. 2007]

See reviews...

See all...

Cited by 6 PubMed Central articles

Autologous hematopoietic stem cell transplantat [PLoS One. 2017]

AP: Aspect Based Temporal Clustering (AC)

Given:

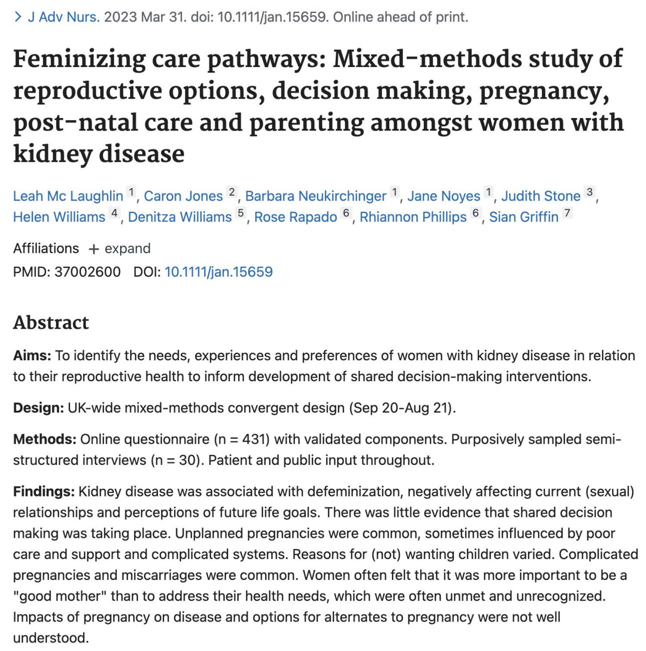
- Dataset containing ~5M PubMed abstracts, metadata, and text embeddings
- Reference Paper

Tasks:

- Develop frontend that allows to perform temporal clustering of PubMed abstracts
- On-demand clustering based on different aspects, i.e., topics, keywords, entities, and many more

Languages / Tools:

- Python, OpenSearch, FastAPI, Svelte



> J Adv Nurs. 2023 Mar 31. doi: 10.1111/jan.15659. Online ahead of print.

Feminizing care pathways: Mixed-methods study of reproductive options, decision making, pregnancy, post-natal care and parenting amongst women with kidney disease

Leah Mc Laughlin [1], Caron Jones [2], Barbara Neukirchinger [1], Jane Noyes [1], Judith Stone [3], Helen Williams [4], Denitza Williams [5], Rose Rapado [6], Rhiannon Phillips [6], Sian Griffin [7]

Affiliations + expand

PMID: 37002600 DOI: 10.1111/jan.15659

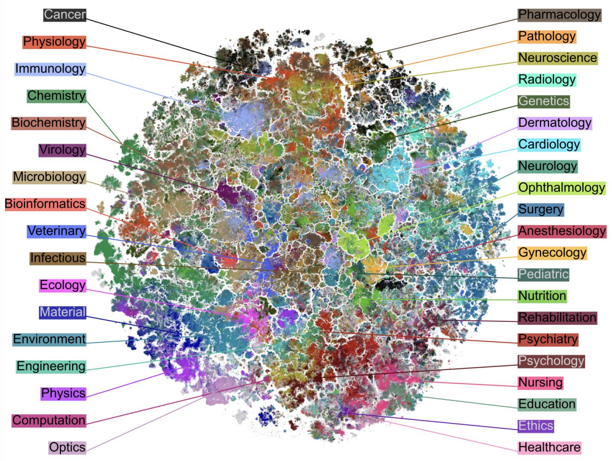
Abstract

Aims: To identify the needs, experiences and preferences of women with kidney disease in relation to their reproductive health to inform development of shared decision-making interventions.

Design: UK-wide mixed-methods convergent design (Sep 20-Aug 21).

Methods: Online questionnaire (n = 431) with validated components. Purposively sampled semi-structured interviews (n = 30). Patient and public input throughout.

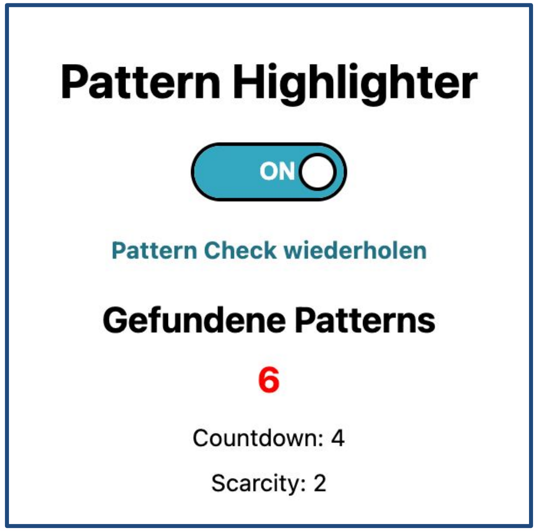
Findings: Kidney disease was associated with defeminization, negatively affecting current (sexual) relationships and perceptions of future life goals. There was little evidence that shared decision making was taking place. Unplanned pregnancies were common, sometimes influenced by poor care and support and complicated systems. Reasons for (not) wanting children varied. Complicated pregnancies and miscarriages were common. Women often felt that it was more important to be a "good mother" than to address their health needs, which were often unmet and unrecognized. Impacts of pregnancy on disease and options for alternates to pregnancy were not well understood.



BP/AP: Machine Learning in Web Browsers (NR)

Given:

- List of types of patterns on web pages, e.g., countdowns, scarcity, ...
- Open-source browser extension that can find some patterns using simple methods



Pattern Highlighter

Gefundene Patterns

6

Countdown: 4

Scarcity: 2

Tasks:

- Implement a machine learning model in the browser extension to find one or more pattern types (as proof-of-concept)

Subtasks:

- Explore available ML libraries for Javascript
- Examine feasibility of running ML directly in the browser (speed, requirements)

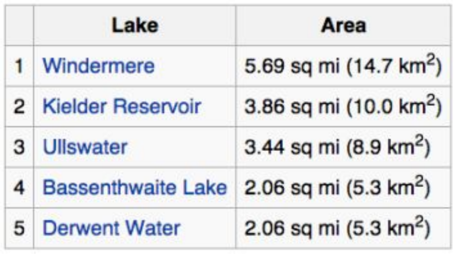
Languages / Tools:

- Javascript, HTML, CSS, Python

AP: Creating a Domain-Aware Web Table Corpus (NR)

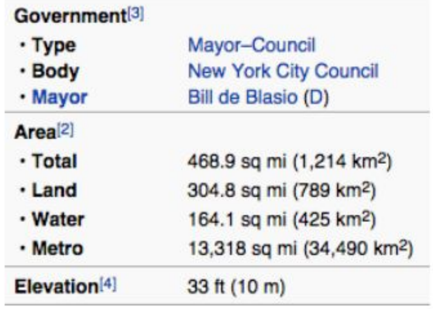
Given:

- Common crawl web corpus
- Dresden Web Table Corpus extractor source code (Java)



	Lake	Area
1	Windermere	5.69 sq mi (14.7 km^2)
2	Kielder Reservoir	3.86 sq mi (10.0 km^2)
3	Ullswater	3.44 sq mi (8.9 km^2)
4	Bassenthwaite Lake	2.06 sq mi (5.3 km^2)
5	Derwent Water	2.06 sq mi (5.3 km^2)

(a) Relational Table



Government[3]	
• Type	Mayor–Council
• Body	New York City Council
• Mayor	Bill de Blasio (D)
Area[2]	
• Total	468.9 sq mi (1,214 km^2)
• Land	304.8 sq mi (789 km^2)
• Water	164.1 sq mi (425 km^2)
• Metro	13,318 sq mi (34,490 km^2)
Elevation[4]	33 ft (10 m)

(b) Entity Table

Tasks:

- Develop a pipeline in Python to create a web table corpus
- Find a method to classify web pages/tables into domain fields

Subtasks:

- Detect tables and classify them by type
- Classify web pages and make a selection
- Optional: Extract tables (convert them to dataframes)

Languages / Tools:

- Python, HTML, (Java)

Slides Online

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