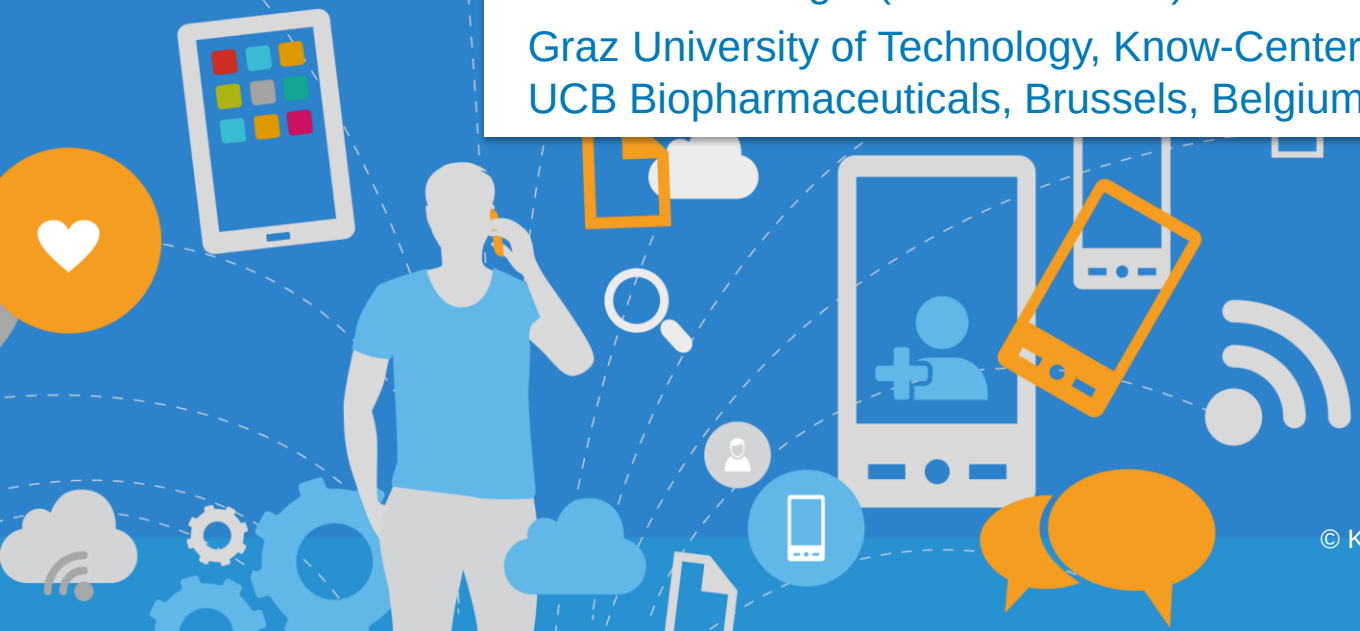
KNOW Center

AUSTRIA'S LEADING RESEARCH CENTER
FOR DATA-DRIVEN BUSINESS AND BIG DATA ANALYTICS

CUSTOMIZABLE CATEGORIZATION OF DOCUMENTS IN EVIDENCE-BASED RESEARCH FOR BIOPHARMA

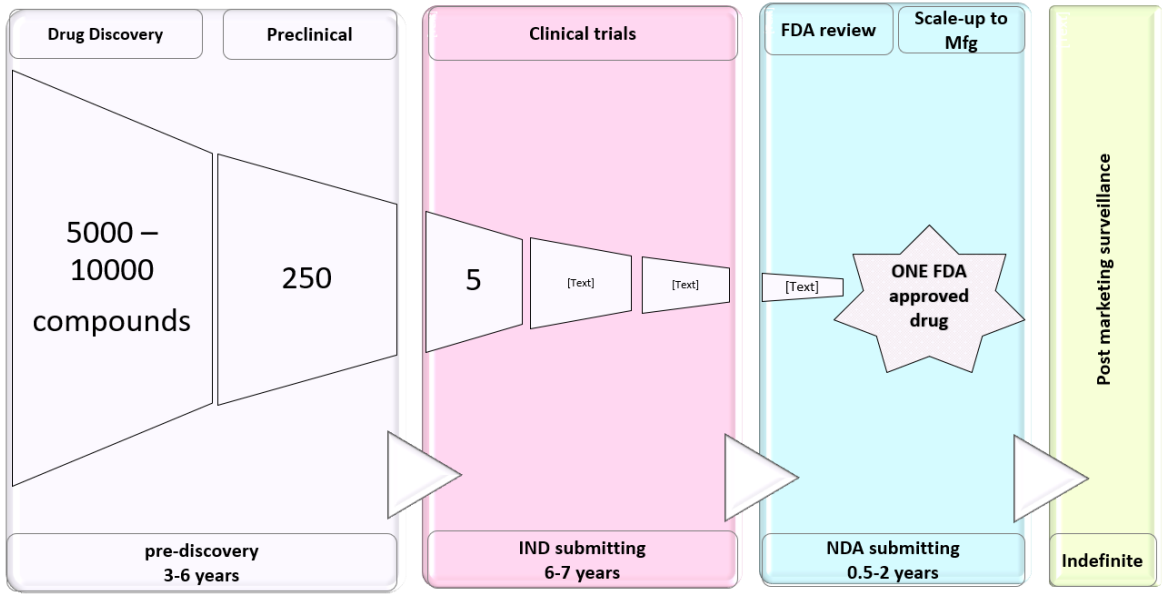
Reihaneh Manteghi (PhD candidate) - Eduardo Veas (supervisor)
Graz University of Technology, Know-Center GmbH, Austria
UCB Biopharmaceuticals, Brussels, Belgium

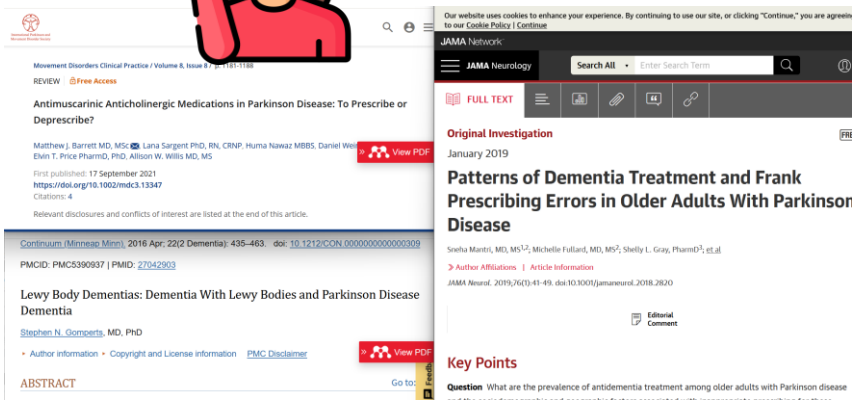


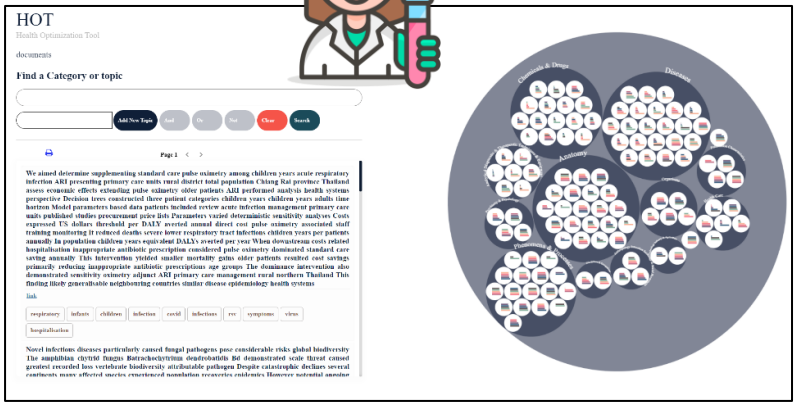
© Know-Center GmbH, www.know-center.at

What is the problem?

Drug development







© Know-Center GmbH • Research Center for Data-Driven Business and Big Data Analytics

Overall research question

Can interactive search tools with intelligent user interfaces help domain experts in **decision-making** processes?

Application of intelligent interactive tools in decision support

Use case: Drug development

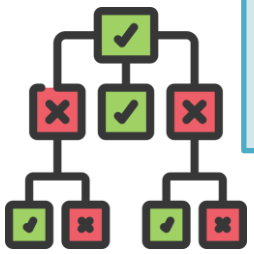
Decision-making problems

Sub-problems

Find paths within complex constraints in decision trees

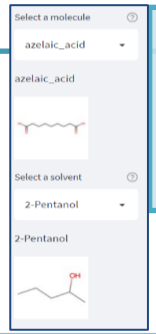
Use case: Drug formulation

Item configuration with constraint satisfaction



A modular and interactive framework for integrating prediction models.

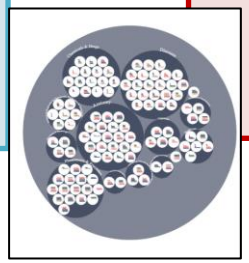
Use case: Solvent prediction



Interactive machine learning for predicting and comparing items on multiple criteria

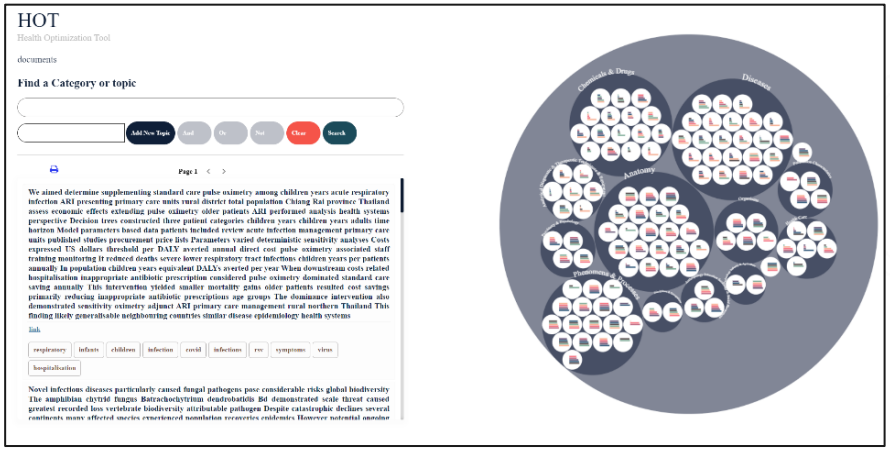
Research within the state-of-the-art publications

Use case: Explore biopharma literature



Literature research

© Know-Center GmbH • Research Center for Data-Driven Business and Big Data Analytics



Literature research assistant

Goal:
Customized classification and data exploration through scientific documents to improve the decision-making process

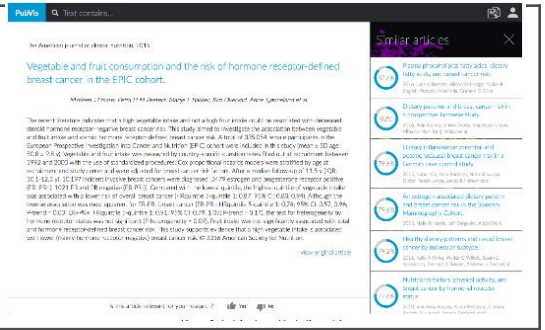
Used in:
Different domains

Use case: Explore biopharma literature

Similar tools

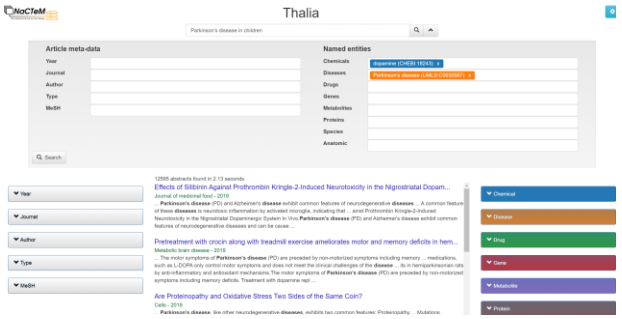
PubVis[1]

Uses **interactive machine learning** and combines essential features to help researchers find the content most relevant to them-articles can be tailored to the user's interests



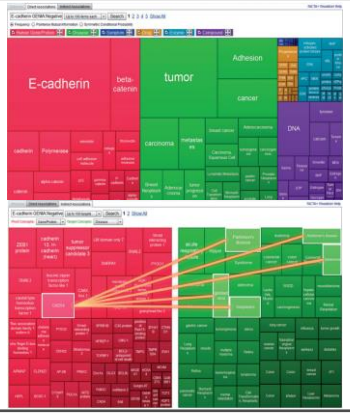
Thalia[2]

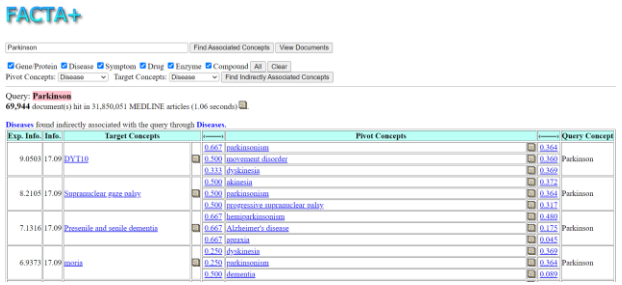
A semantic search engine for exploring **biomedical** abstracts



FACTA +[3]

- Discovering useful **associations** between **biomedical** concepts
- Visualization of **directly** and **indirectly** associated concepts

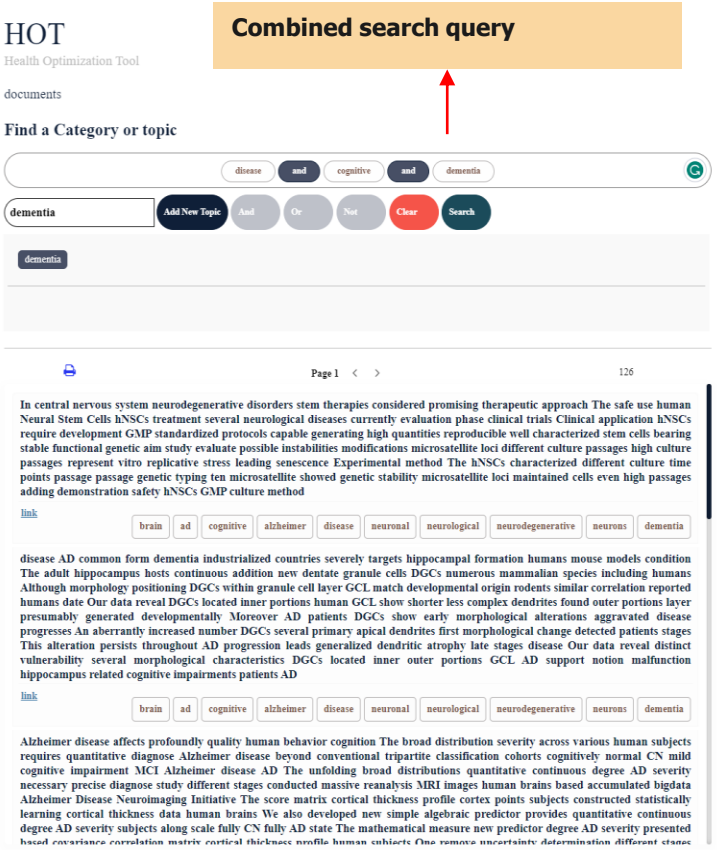




1. Horn F. Interactive Exploration and Discovery of Scientific Publications with PubVis. arXiv e-prints. 2017 June. doi: 10.48550/arXiv.1706.08094.
2. Soto AJ, Przybyła P, Ananiadou S. Thalia: semantic search engine for biomedical abstracts. Bioinformatics. 2019 May 15;35(10):1799-1801. doi: 10.1093/bioinformatics/bty871. PMID: 30329013; PMCID: PMC6513154.
3. Tsuruoka Y, Miwa M, Hamamoto K, Tsujii J, Ananiadou S. Discovering and visualizing indirect associations between biomedical concepts. Bioinformatics. 2011 Jul 1;27(13):i111-9. doi: 10.1093/bioinformatics/btr214. PMID: 21685059; PMCID: PMC3117364.

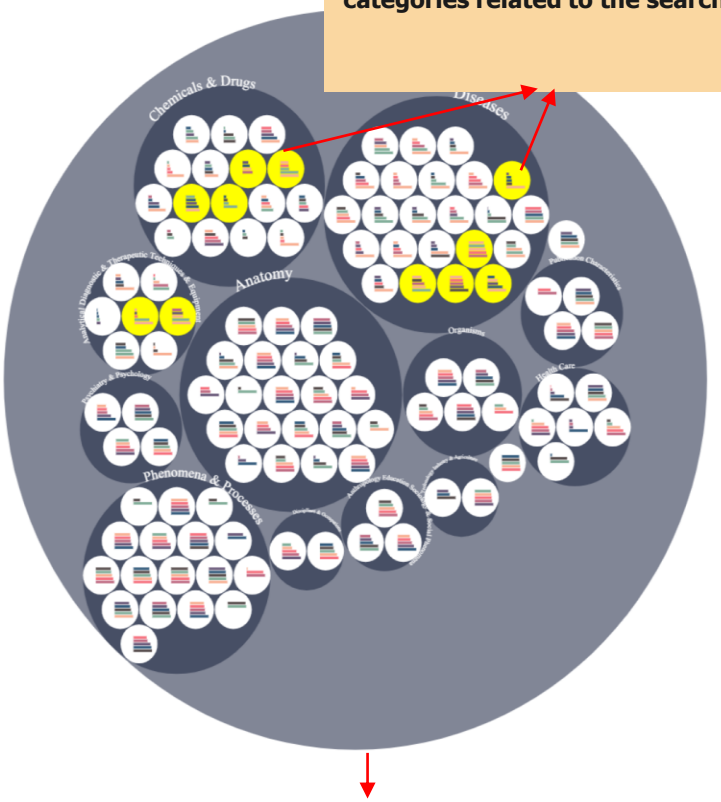
Literature research assistant

How it works:
In-house data exploration tool with search and customized classification option which can be used on different literature data



Combined search query

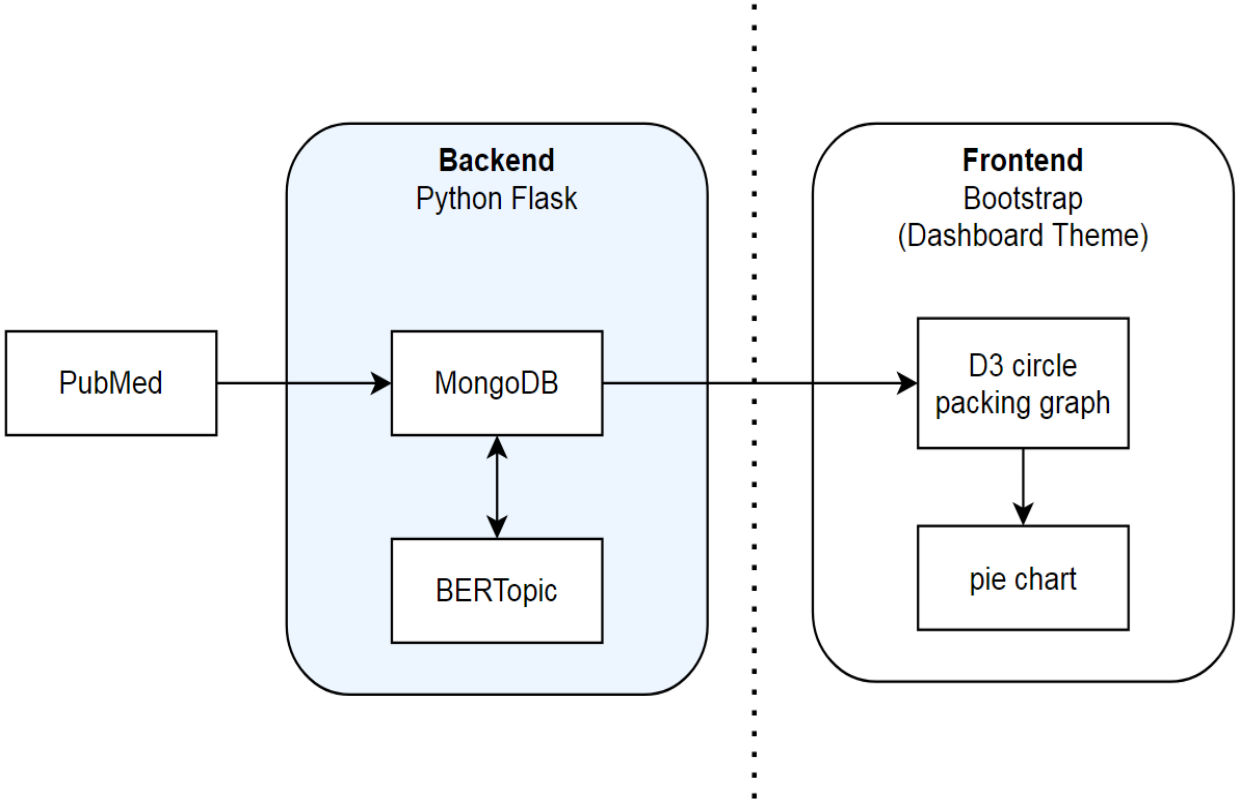
Data table including existing publications and the links



Existing documents within the categories related to the search query

Zoomable graph to find publications by categories

Architecture



© Know-Center GmbH • Research Center for Data-Driven Business and Big Data Analytics

Architecture

Story behind the categories

https://meshb.nlm.nih.gov/treeView

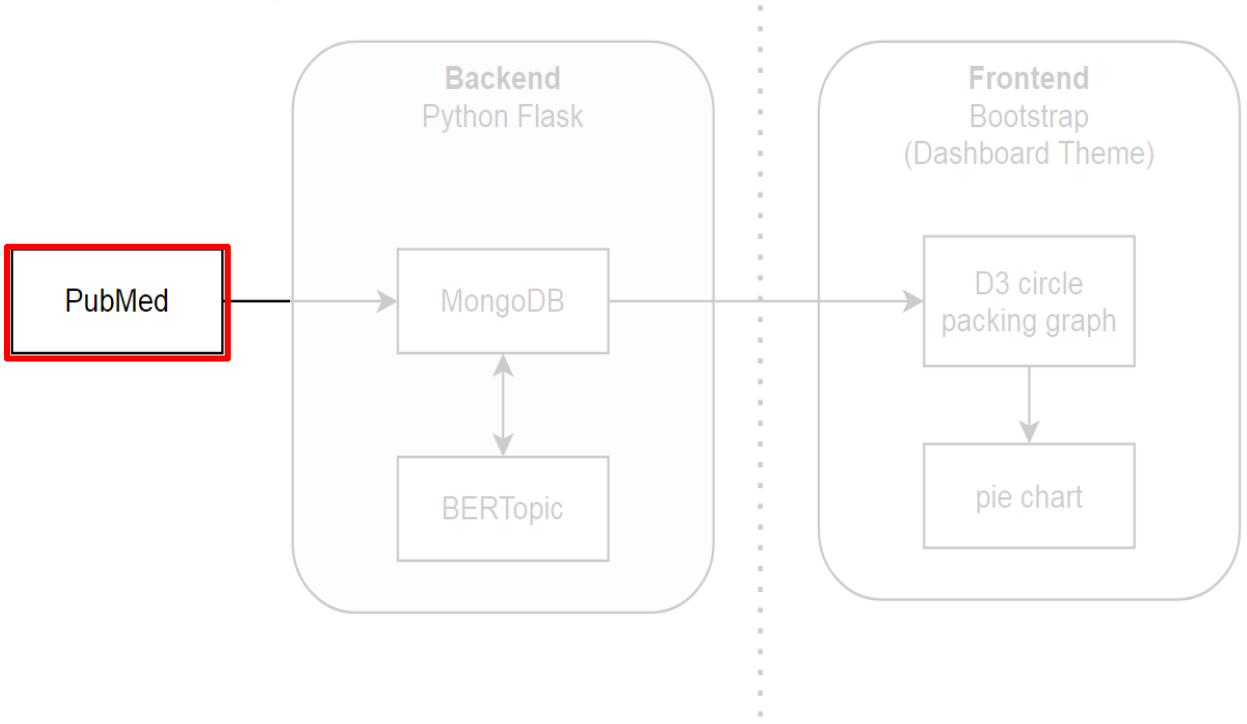
Anatomy [A]
Organisms [B]
Diseases [C]
Chemicals and Drugs [D]
Analytical, Diagnostic and Therapeutic Techniques, and Equipment [E]
Psychiatry and Psychology [F]
Phenomena and Processes [G]
Disciplines and Occupations [H]
Anthropology, Education, Sociology, and Social Phenomena [I]
Technology, Industry, and Agriculture [J]
Humanities [K]
Information Science [L]
Named Groups [M]
Health Care [N]
Publication Characteristics [V]
Geographicals [Z]

Medical Subject Headings 2022

The files are updated each week day Monday-Friday by 8AM EST

Search MeSH... FullWord

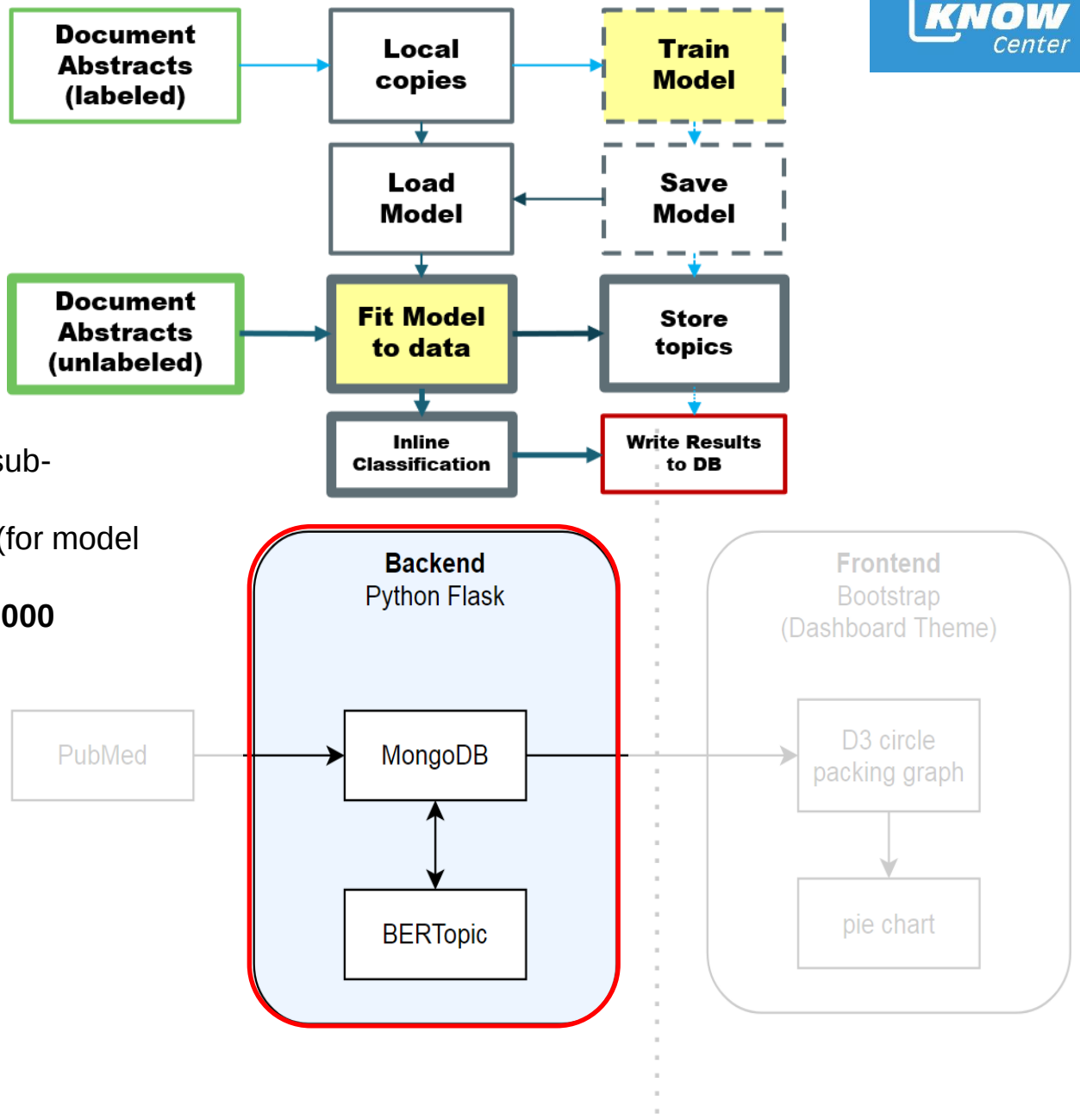
- All Terms
  - Main Heading (Descriptor) Terms
  - Qualifier Terms
  - Supplementary Concept Record Terms
- MeSH Unique ID
- Search in all Supplementary Concept Record Fields
  - Heading Mapped To
  - Indexing Information
- Pharmacological Action
- Search Related Registry and CAS Registry/EC Number/UNII Code/NCBI Taxonomy ID Number (RN)
  - Related Registry Search
  - CAS Registry/EC Number/UNII Code/NCBI Taxonomy ID Number (RN)
- Search in all Free Text Fields
  - Annotation
  - ScopeNote
  - SCR Note



© Know-Center GmbH • Research Center for Data-Driven Business and Big Data Analytics

Architecture

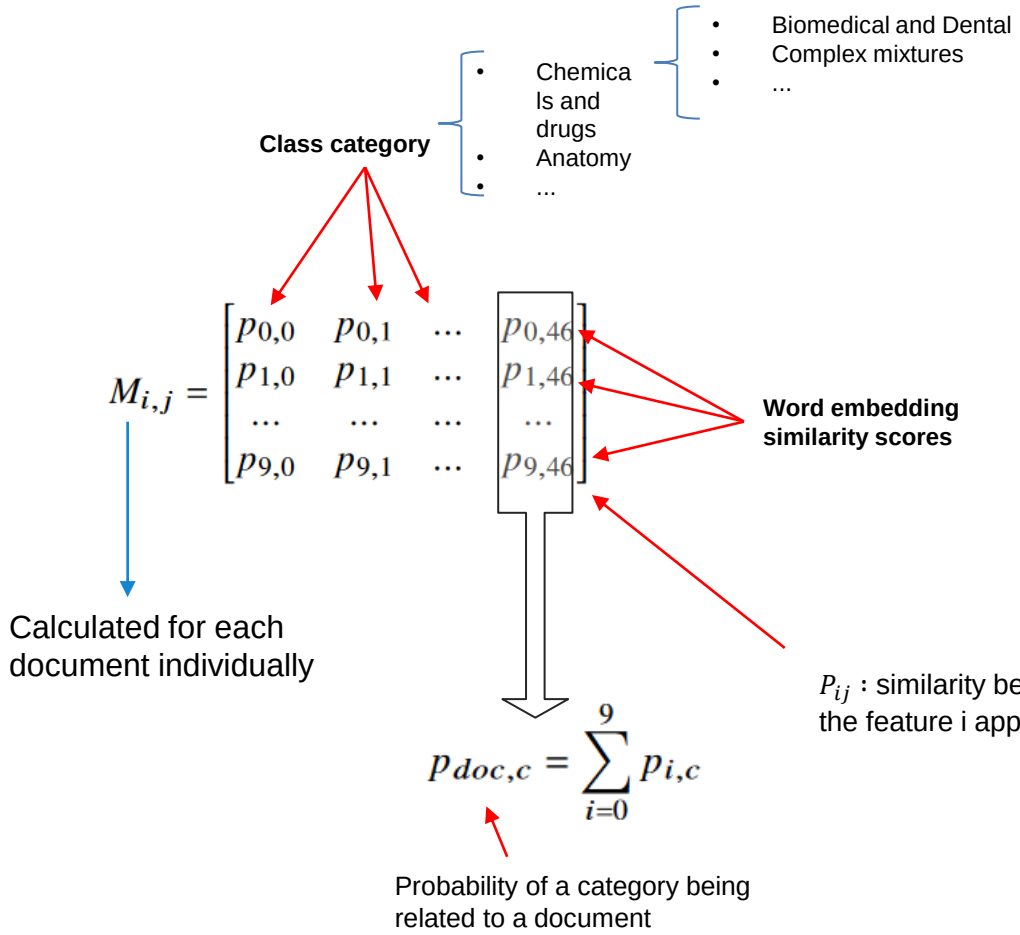
Back-end model



Number of classes (categories and sub-categories) : **47**
Number of documents per category (for model training):**10000**
Number of unlabeled documents: **10000**

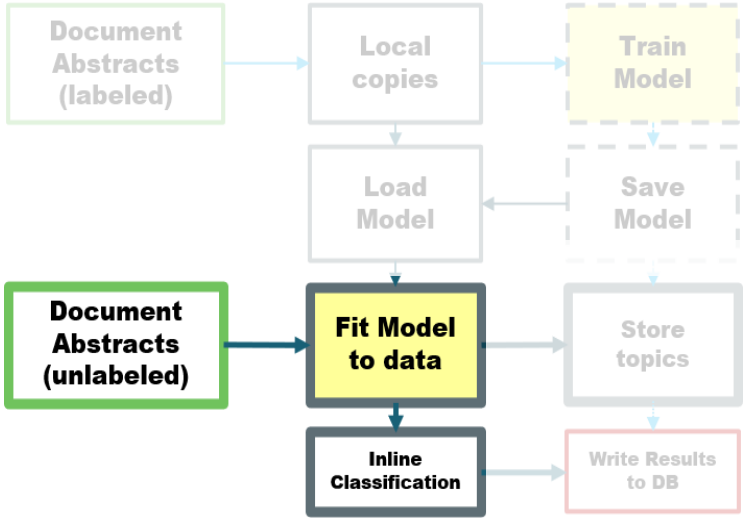
© Know-Center GmbH • Research Center for Data-Driven Business and Big Data Analytics

Architecture
Back-end model
Similarity detection and classification

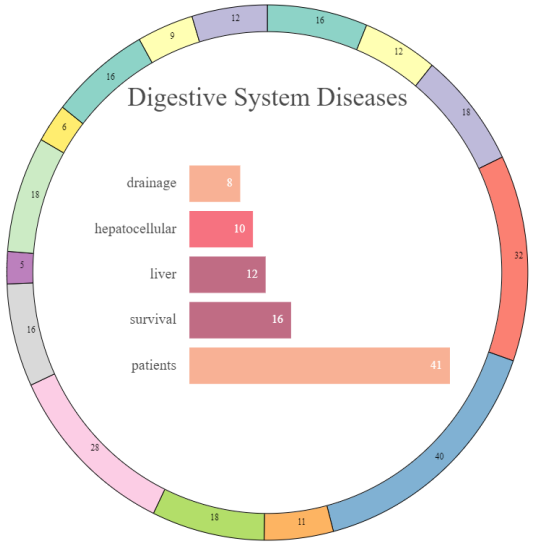


P_{ij} : similarity between features for the feature i appearing in class j

Top 10 Features are extracted from introduced categories using bertopic



Architecture
Front-end



patients - group - analysis - clinical - in - using - treatment - study - the - survival - liver - hepatocellular - prognostic - hepatectomy - afp - tace - hcc - drainage - robotic - underwent - surgical - complications - surgery - laparoscopic - postoperative - cancer - carcinoma - proliferation - crc - immune - expression - tumor - cell - cells - intestinal - ticagrelor - thrombophilia - platelets - vwd - coagulation - thrombotic - platelet - thrombosis - bleeding - polarization - aunps - reversal - hepatocytes - macrophages -

BERTopic

Each individual color is assigned to a related category

Frontend
Bootstrap
(Dashboard Theme)

D3 circle packing graph

pie chart

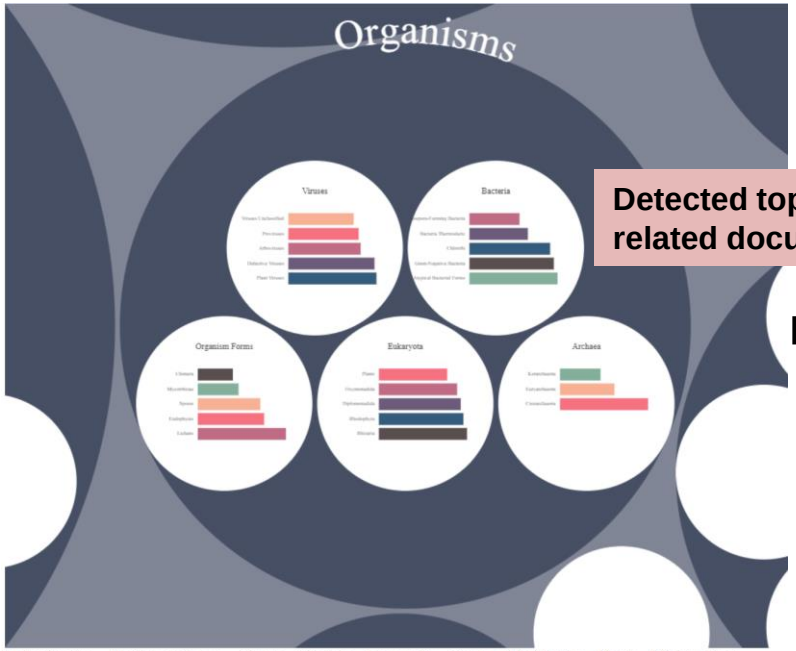
Features in details

Zoom-in/out to explore categories

Visualize relational information using a donut chart

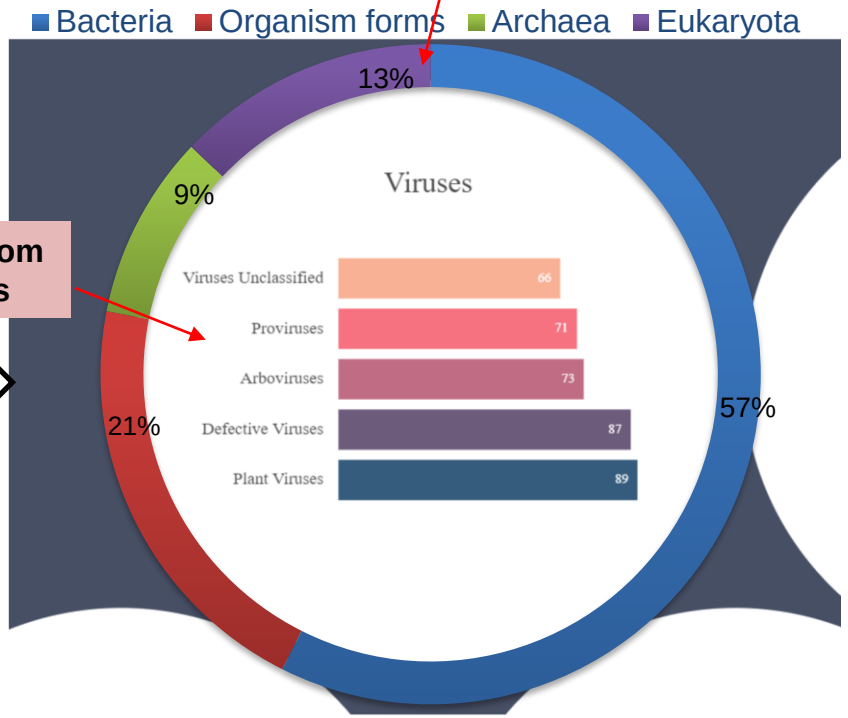
- Provide additional information about the other related categories.
- Each piece of the donut represents the similarity of the selected category to another category based on mutual topics

Percentages of documents being related to other categories



Aquatic Organisms - Blood-Borne Pathogens - Chimera - Endophytes - Extremophiles - Hermaphroditic Organisms - Lichens - Life Cycle Stages - Microfilariae - Mycorrhizae - Organisms Genetically Modified - Sentinel Species - Spores -

Detected topics from related documents



Arboviruses - Archaeal Viruses - Bacteriophages - Defective Viruses - DNA Viruses - Fungal Viruses - Helper Viruses - Hepatitis Viruses - Insect Viruses - Oncogenic Viruses - Oncolytic Viruses - Plant Viruses - Proviruses - Reassortant Viruses - RNA Viruses - Viroids - Viruses Unclassified -

Rest of the detected topics with less frequencies

Evaluations

Goal: Assist pharma scientists in literature research and exploration
Baseline: Conventional methods : (e.g., online database search, in-house document search, ...)

Users: **Three** domain expert users
Task: Find results for given search queries functions

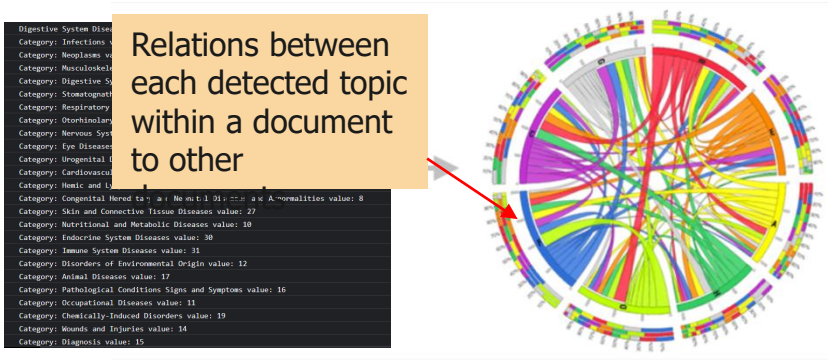
Evaluation method: Usability study using questionnaires and quantitative evaluation

Results:
1. All three users found existing features and the graph representation useful
2. Some users found it difficult to interact with the tool except the user who was in close collaboration with the tool development
3. Users found it highly useful to have manual guide for the tool
4. Users found it very useful to offer different solutions for navigation through the graph (e.g., mouse wheel)

Literature research assistant future work

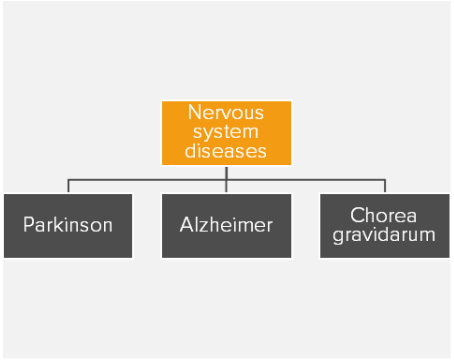
Visualize relational information using a circus chart

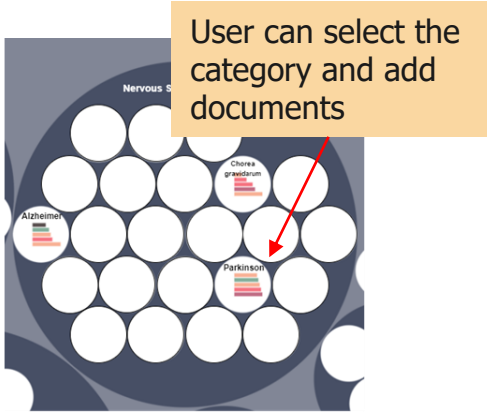
- Visualize the relationship between categories
- Visualize which topic on the bar is related to other categories



Customize the classification by retraining the classifier

Users can upload batches of 100 classified documents to a desired class and retrain the model to update the target class definition

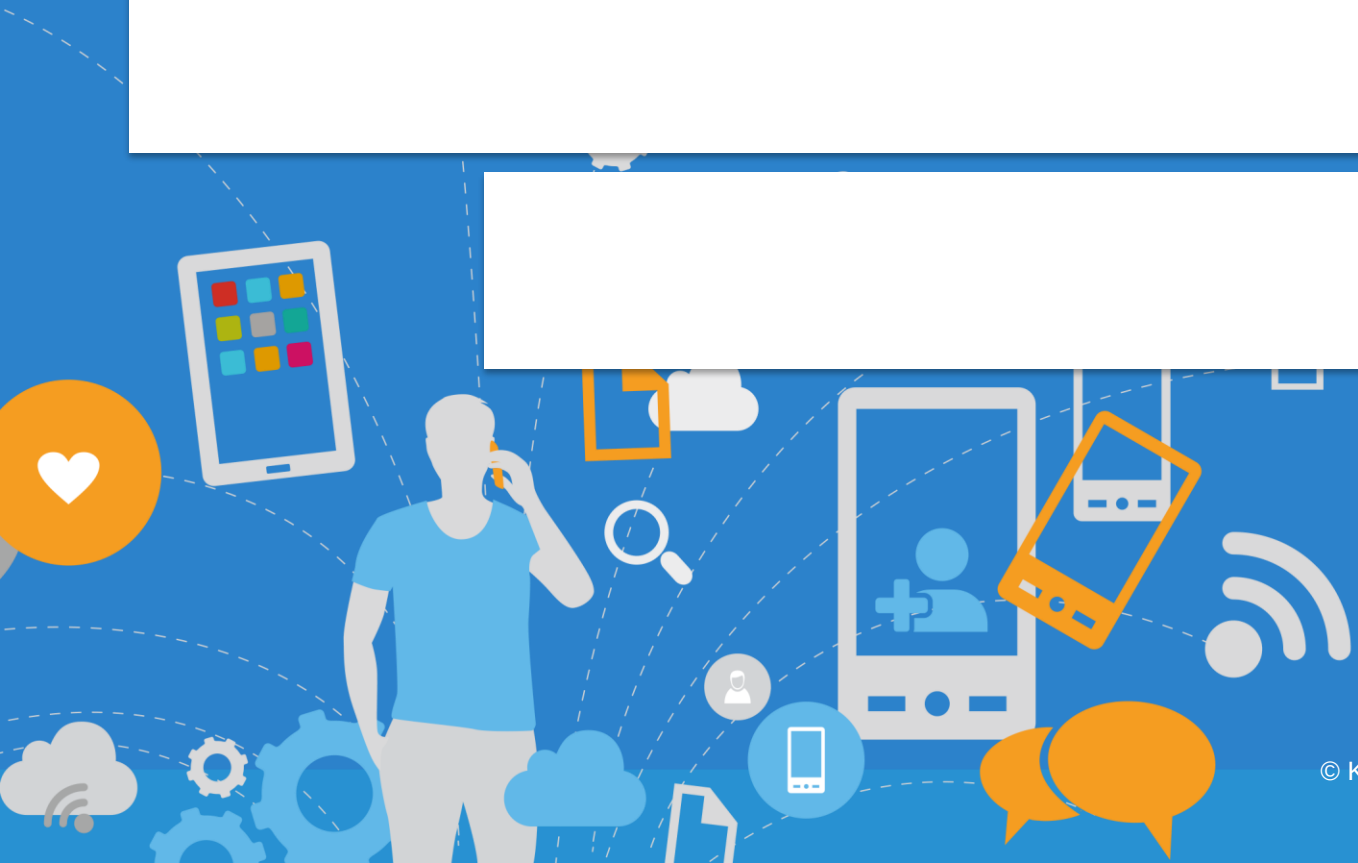




KNOW Center

AUSTRIA'S LEADING RESEARCH CENTER
FOR DATA-DRIVEN BUSINESS AND BIG DATA ANALYTICS

THANK YOU



© Know-Center GmbH, www.know-center.at

Features in details

View other related categories

Health Optimization Tool

Skin and Connective Tissue Diseases

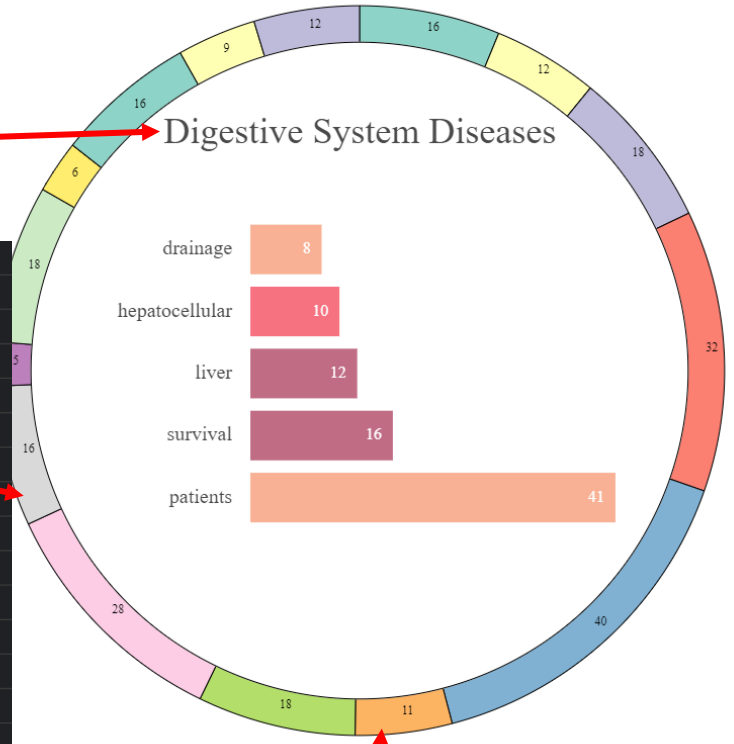
Find a Category or topic

Add New Topic | And | Or | Not | Clear | Search

Selected category

```
Digestive System Diseases
Category: Infections value: 28
Category: Neoplasms value: 28
Category: Musculoskeletal Diseases value: 14
Category: Digestive System Diseases value: 49
Category: Stomatognathic Diseases value: 22
Category: Respiratory Tract Diseases value: 21
Category: Otorhinolaryngologic Diseases value: 21
Category: Nervous System Diseases value: 7
Category: Eye Diseases value: 11
Category: Urogenital Diseases value: 14
Category: Cardiovascular Diseases value: 13
Category: Hemic and Lymphatic Diseases value: 31
Category: Congenital Hereditary and Neonatal Diseases and Abnormalities value: 8
Category: Skin and Connective Tissue Diseases value: 27
Category: Nutritional and Metabolic Diseases value: 10
Category: Endocrine System Diseases value: 30
Category: Immune System Diseases value: 31
Category: Disorders of Environmental Origin value: 12
Category: Animal Diseases value: 17
Category: Pathological Conditions Signs and Symptoms value: 16
Category: Occupational Diseases value: 11
Category: Chemically-Induced Disorders value: 19
Category: Wounds and Injuries value: 14
Category: Diagnosis value: 15
```

Other related categories



nical - in - using - treatment - study - the - survival - liver - hepatocellular - prognostic - hepatectomy - afp - tace - hcc - drainage - robotic - tions - surgery - laparoscopic - postoperative - cancer - carcinoma - proliferation - crc - immune - expression - tumor - cell - cells - intestinal - elets - vwd - coagulation - thrombotic - platelet - thrombosis - bleeding - polarization - aunps - reversal - hepatocytes - macrophages -

Each individual color is assigned to a category