

This directory releases all the evaluation data used in experimental section of ***“Focused Query Expansion with Entity Cores for Patient-Centric Health Search”*** paper.

Queries

Inside directory Queries you can find a dataframe containing all queries used in the paper. The selected queries are post published by users on eHealthForum posing questions about health topics with high coverage, such as: Alzheimer Disease, Acid Reflux, Arthritis, Asthma, Back Pain, Carpal Tunnel Syndrome, Crohn Disease, Diabetes, Skin Cancer, Fibromyalgia, High/Low Blood Pressure and Hypertension, Insect Bites, Meningitis, Multiple Sclerosis, Pancreas Disorders, Sinusitis, Vision and Eye Disorders. The query file is a tab-separated text file with the following columns:

- *Query_Id*: Unique id of the query
- *Post_Title*: Title text of the query
- *Post_Body*: Post text of the query
- *QE_Entities*: Entities from the "Entity Expansion" method picked for expansion
- *ST_Entities*: Entities picked from the "Steiner Tree" method for expansion
- *ECO_Entities*: Entities picked from the "ECO" method for expansion

Evaluation Data

Inside directory Evaluation_Data you can find individual files for crowd worker judgements for each of the methods mentioned in the paper. Each evaluation file for each method is a tab-separated text file and contains the following information:

- *Id*: Row Id
- *Query_Id*: Id of the query being evaluated
- *Doc_Id*: Id of the document retrieved as a result from executing query number Query_Id using method MethodName
- *Doc_URL*: URL of the document retrieved from executing query number Query_Id using method MethodName
- *Relevance*: Relevance label of the retrieved document. It can be 1 or 0 to present whether the retrieved document is relevant or non-relevant respectively. These labels were collected through a crowd-sourcing experiment conducted by the authors. The crowd workers were asked to judge if the retrieved document is relevant or not for a particular query.
- *Doc_Text*: Thread Text of the document retrieved as a result from executing query number Query_Id using method MethodName

DISCLAIMER: This dataset is released for research purposes only. We can not publish all 1.1 million thread/clinical_trials open source, because they come from massively crawled websites where the content is the legal property of third parties. Instead, we publish only top 10 results for <query, method> pair.