Decision/Health Misinformation Track

Summary
The Decision Track provides a venue for research on retrieval methods that promote better decision making with search engines and in developing new evaluation methods, both online and offline to predict the decision making quality induced by search results. The track has been renamed to TREC Health Misinformation in 2020. (https://trec-health-misinfo.github.io/)
Search Engine Results Pages (SERP) are the pages displayed by search engines in response to a query by a searcher. The main component of the SERP is the listing of results that are returned by the search engine in response to a keyword query. Search engine results support many consequential decision-making tasks and people use search technologies to seek health advice online. Time-pressured clinicians also rely on search results to decide upon the best diagnosis for a patient.
The key problem in decision-making tasks using search engines is filtering out and discerning authoritative information from unreliable information. Another problem is the search occurs within uncontrolled and on unreliable data collections such as the web, where information can be generally misleading and too technical. Research shows that increasing the amount of incorrect information about a topic presented in a SERP can impel users to take incorrect decisions (Pogacar et al., 2017).
Evaluation measures for these search tasks also need to be developed and improved.

Task
The coordinators plan the track over multiple years, with data and resources created in one year flowing into the next year. The track is planned to run for at least 3 years.
Search technologies are built to promote correct information over incorrect information. This task is more than simply a new definition of what is relevant.
There are 3 types of results: correct and relevant, incorrect, and non-relevant.
Search results should avoid containing incorrect results, and ranking non-relevant results above incorrect is preferred. Evaluation measures consider relevance beyond topicality and include information and credibility correctness. A dual goal is to return the relevant and correct information. Following the year 1 assessment, the organizers recruit test subjects to perform a decision making task using a selection of the year 1 runs. Test subjects are given a fixed result list which is selected from the participating teams' submitted runs and a decision task.

From 2020 onward, in addition to a ranking task, the track will have evaluation tasks. Given a query, a document ranking (results list) and interaction data of real users (collected right after year 1), researchers have to predict the decisions along with the confidence score that users may take at the end of the search process. This simulates an online evaluation process.

Data
The Track focuses on topics within the consumer health search domain and related to topics where people seek health advice online to form user stories.

Search Topics
The assessors are provided with the topic query and narrative as XML files instead of creating their own topic statements.

Dataset
The collection used in TREC Decisions 2019 is ClueWeb12-B13 (https://lemurproject.org/clueweb12/). The dataset contains a collection of about 50 million pages.
ClueWeb12 B13 Summary Statistics

- Size compressed: 389 GB
- Size uncompressed: 1.95 TB
- Number of WARC files: 33,447
- Number of documents: 52,343,021

The link provided doesn't provide much detail about the dataset but the description provided by CLEF' 2018 for ClueWeb12-B13 is as follows:
The webpages are obtained from the CommonCrawl(https://commoncrawl.org/) and 50 medical queries issued by the general public and gathered from Health on the Net (HON)(http://hon.ch/) search engine.

Assessing guidelines
The documents are assessed in three categories:
Relevance: whether the document is relevant to the topic.
Credibility: whether the document is considered credible by the assessor.
Treatment Efficacy: whether the document contains correct information regarding the topic's treatment.

Evaluation of Runs
The submitted runs are evaluated with respect to the following measures proposed by Lioma et al. (ICTIR'17, https://doi.org/10.1145/3121050.3121072):

- Normalised Local Rank Error (NLRE) - Compares the rank position of documents pairwise and checks for errors
- Normalised Weighted Cumulative Score (nWCS) - Generates a single label out of the multiple aspects and computes the standard nDCG.

- Convex Aggregating Measure (CAM) - Considers each aspect separately and computes either AP or nDCG. Finally average AP or nDCG is computed across the aspects.
- Normalized Discounted cumulative gain (nDCG) and MAP

Approach to the task

The query can be expanded using Entity Query Feature Expansion model (EQFE). Documents can be retrieved using BM25 approach and then re-ranked using Deep Structured Semantic Models (DSSM) or CDSSM models.