Cmpe 493 Introduction to Information Retrieval, Fall 2020
Term Project - Information Retrieval for Covid-19

In this project, you will work on information retrieval from the *Covid-19* related scientific literature. This task has been recently addressed in the *TREC-COVID* Challange (`https://ir.nist.gov/covidSubmit/`).

You will use the TREC-COVID Complete data set (`https://ir.nist.gov/covidSubmit/data.html`). The data set includes documents (scientific papers) from the *July 16, 2020 release of the CORD-19* corpus[1]. You should download the document collection from `https://ai2-semanticscholar-cord-19.s3-us-west-2.amazonaws.com/historical_releases/cord-19_2020-07-16.tar.gz`. It is around 3.7 GB. We will only use the titles and abstracts of the papers.

A set of 50 topics (i.e., queries) is provided at `https://ir.nist.gov/covidSubmit/data/topics-rnd5.xml`. You will use the odd numbered topics (i.e., 1, 3, 5, 7,..., 49) for developing your systems and the even numbered topics (i.e., 2, 4, 6, 8,...,50) for evaluation.

The relevance judgements for the documents with respect to the topics are provided at `https://ir.nist.gov/covidSubmit/data/qrels-covid_d5_j0.5-5.txt`. Each line in the file contains the relevance judgement for a document. The first column is the topic-id, the second column is iteration (you will NOT be using this field), the third column is the document-id (cord-id), and the last column is the relevance judgement, where 0 means not-relevant, 1 means partially relevant, and 2 means fully relevant.

We will use Mean Average Precision (MAP), Normalized Discounted Cumulative Gain (NDCG), and Precision of top 10 results (P@10) for evaluation. You should use the official evaluation tool available at `https://github.com/usnistgov/trec_eval`.

Some relevant papers are provided below and more papers are avaiable at `https://ir.nist.gov/covidSubmit/bib.html`. I also suggest you to look for other relevant publications on the Web.

- Roberts, Kirk, et al. ”TREC-COVID: Rationale and Structure of an Information Retrieval Shared Task for COVID-19.” Journal of the American Medical Informatics Association (2020). Available at `https://academic.oup.com/jamia/article/27/9/1431/5828938`.

- Voorhees, Ellen, et al. ”TREC-COVID: Constructing a Pandemic Information Retrieval Test Collection.” ACM SIGIR Forum (2020). Available at `https://ir.nist.gov/covidSubmit/papers/Forum_TRECCOVID1.pdf`.

- Chen, Jimmy, and William Hersh. ”A Comparative Analysis of System Features Used in the TREC-COVID Information Retrieval Challenge.” medRxiv (2020). Available at `https://www.medrxiv.org/content/10.1101/2020.10.15.20213645v1`

- Zhang, Edwin, et al. ”Covidex: Neural Ranking Models and Keyword Search Infrastructure for the COVID-19 Open Research Dataset.” Proceedings of the First Workshop on Scholarly

[1] Wang, Lucy Lu, et al. ”CORD-19: The Covid-19 Open Research Dataset.” ArXiv (2020).

Document Processing. 2020. Available at https://www.aclweb.org/anthology/2020.sdp-1.5/.

- Esteva, Andre, et al. ”Co-search: Covid-19 information retrieval with semantic search, question answering, and abstractive summarization.” arXiv preprint arXiv:2006.09595 (2020). Available at https://arxiv.org/abs/2006.09595.

Deliverables:

1. Project progress presentation (December 28, 2020 (in the lecture hour); 30% of your project score): You should prepare a 10min presentation describing what you have done so far and what your plan is for the remaining time period. You should have completed at least the preprocessing of the data set and implemented and tested a baseline approach (such as TF-IDF based cosine similarity). You should also have clear plans about how you will improve your system by the end of the semester.
2. Project final presentation (On the final exam date/slot; 70% of your project score): You should prepare a 10min presentation describing your final system and your results. I also suggest you to include an error analysis.
3. Prior to each presentation (latest 1 hour before the presentations start) you should send me by email your slides and all source code and accompanying readme documents.

Honor Code: You should work in teams of two or three people. Each team member should contribute equally to the development of the project and to the presentations. All team members will get the same score. You are allowed to use external libraries/resources for the project. However, you SHOULD properly acknowledge and cite these in your presentations and source code.

Late Submission: Late submissions are NOT allowed.