CS410 Project Proposal [Otto Piramuthu]

NetID: obp2@illinois.edu

[Free] Topic: *Online Algorithms for Technology-Assisted Reviews*

Please give a detailed description. What is the task? Why is it important or interesting?
Exhaustive manual review of documents to determine their relevancy for a given purpose is error-prone and resource intensive. This has led to the consideration of computer-aided processes for litigation support where only a small subset of the entire set of documents is manually reviewed with comparable performance as exhaustive manual review, resulting in the reduction of human-introduced error and allocated resources. As more evidence for the superiority of Technology-Assisted Reviews (TAR) becomes available, researchers and practitioners have resorted to the exploration of various methods to improve process efficiency without significant degradation in output quality. Of particular interest in this process is the decision on when to stop reviewing additional documents. I evaluate online algorithms, which are a natural fit for this purpose.

What is your planned approach? What tools, systems or datasets are involved? What is the expected outcome? How are you going to evaluate your work?
I have tried counting processes to determine the total number of relevant documents (R), which is necessary for online learning. The next step is to try at least one other method to determine R and incorporate this information in online algorithms. This is followed by comparison of the performance of these methods. The dataset used is from CLEF 2017 e-Health Lab Task. The hoped outcome is that the performance of online algorithms for TAR is competitive with those from existing methods. I plan to compare results from this study with that from a published study

Which programming language do you plan to use?
Python.

Please justify that the workload of your topic is at least 20*N hours, N being the total number of students in your team.
I am doing this project by myself and have written relevant code for online algorithms with counting processes. I have written up the results so far as a paper that is about 12 pages long.

You may list the main tasks to be completed, and the estimated time cost for each task.
I want to identify at least one other method to determine R and incorporate that method with online algorithms. I suspect that reading papers for their relevance, selecting an algorithm for further consideration, writing code for this selected algorithm, and then incorporating this in online learning would take me at least 20 hours.