605.744 Information Retrieval: Class Project

Goals

The class project allows you the opportunity to investigate a particular topic of information retrieval in greater depth than we could cover in the classroom. Projects are normally individual endeavors, and require advance planning and progressive effort to complete successfully. The majority of projects tend to involve writing or using software to conduct an experiment or analyze a textual dataset. However, I have occasionally had students work on projects that are more theoretical vs. empirical. Example projects[1] include: comparing open source retrieval software using one or more IR test collections; analyzing a collection of police crime reports and performing data mining; and, implementing an IR algorithm (*e.g.,* cover density ranking, nextword indexing) and comparing results against baselines. On empirical projects the relevant literature must be reviewed to give context to your work, a hypothesis must be developed, experiments must be designed, conducted, and analyzed, and finally, results must be presented in a report, and to the rest of class. "Theory" projects are expected to include a much more extensive review and study of the literature; they should exhibit independent thinking, and the written report is expected to be longer and more comprehensive (~ 15 pages vs. 6 to 8 pages).

To earn a grade of A- or higher in the course, students must complete a project; however, completing a project does not guarantee receiving an A- or higher. Students are permitted to opt out of submitting a project, in which case the other coursework will determine the final grade, as discussed in the course syllabus.

Grading Criteria

Project grades are based on the proposal (10%), the work performed as documented in the written report (60%), the class presentation (30%), and plus or minus modifiers (these vary between -5% to +3%). A checklist that I use for scoring presentations is attached.

Proposal

A written proposal must be submitted for approval by the instructor. The proposal should have a title, must identify a topic of interest, should briefly motivate why this is an interesting or important problem, identify some relevant scientific literature for the problem of interest, identify sources of data, and outline planned work for the project. Sufficient details about data, experimental design, and evaluation methodology should be provided. Proposals are usually less than 1 page in length. If you have a topic that interests you, but you have questions or are not sure how to proceed, you are strongly encouraged to contact me informally for ideas or feedback in advance of when the proposal is due. I give feedback (and approval) on proposals, so it is important to think about an idea and submit the proposal on time. You can find useful links to conferences and sources of resource papers on the resource page that I host at: http://pmcnamee.net/ir.html

Written Report

The written report is the most significant project deliverable – it is where you document the work that you have performed and it counts for most of the project grade. Reports should be scientifically-

[1] A long list of ideas is attached below.

oriented and should include an abstract, an introduction to the problem, a review of related work (extensive if a theory paper), discussion of ideas (extensive if a theory paper), experimental results with analysis (extensive if an empirical paper), conclusions supported by your work, and appropriate references. You have flexibility in the style of formatting; however, do include headings, and use a font between 10-12 points. Suitable tables and figures are highly encouraged. You should take suitable care to clearly communicate the scale and quality of your work. Reports are due at the end of Module 14 (as a PDF file).

Presentation

A short video presentation must be shared with the class using the discussion forum in Canvas. When multiple sections are being taught you submit your presentation in a thread for your assigned group (i.e., Group A/B/C – the same groups used for paper summaries). The video is due during Module 13. Presentations should be about 10 minutes in length. A common format is to use prepared electronic slides (*e.g.,* PowerPoint, Keynote, OpenOffice, PDF), and to provide voice narration using a tool such as Camtasia. You may prefer instead to provide a video of you speaking or demonstrating a system. The choice of format is yours.

Presentations should clearly introduce the problem under discussion, briefly review prior work from the literature, explain in detail your contribution and results. Experiments are not always successful, and you can achieve a good score on the project, even with negative results; however, your design should be good and you need to articulate what was learned. If your work is theoretical you should illustrate how it might be evaluated and what applications could benefit from your ideas.

Modifiers

Project grades will receive modifiers as follows:

"Best video delivery award" (1 per group):	+ 3 points
"Runner up" for video delivery (1 per group):	+ 2 points
"Best project" (1 per group):	+ 3 points
"Runner up" for best project (1 per group):	+ 2 points
Submission of the presentation video by 11:59pm on Wednesday of Module 13:	+ 3 points
Submission of the presentation video by 11:59pm on Saturday of Module 13:	+ 1 point
Late submission of the presentation video:	- 5 points
Video length > 12 minutes:	- 2 points
Video length > 15 minutes:	- 5 points

Every student in the class, whether or not they are submitting a project of their own, will watch the project videos and complete a short review form for each project in their group. Students whose projects receive the most votes for best video or best project will earn the extra credit modifiers listed above. You are not allowed to vote for your own project!

Schedule

Module 7	Select a topic and submit a proposal.
Module 11	Send a brief status report by email (one paragraph can be enough)
Module 13	Video presentation provided to the class
Module 14	Written report due. Reviews of presentations due.

Literature

Numerous resources are available to you. Online papers can be found via Google Scholar, arXiv, CiteSeer, or various websites for conferences; pointers to the TREC conferences and the ACL Anthology are on the course resource page. The JHU libraries provide no-cost access to the ACM and IEEE digital libraries (you may have to VPN'd to JHU to access those collections).

Sample projects of former students

- Detecting misinformation on Reddit
- Computing Document Similarity using Cloud Computing
- Evaluating indexing and retrieval of Hindi song titles
- Analysis of online police crime reports and classification of crime narratives.
- Sentiment analysis from the 2016 elections
- Parallelization of machine learning algorithms and cloud-based classification.
- Exploring methods to compress indexes using document identifier reassignment
- Extracting obituary information from news sources for genealogical purposes
- Analyzing released records about the Kennedy assassination
- Distributed indexing using Bloom filters
- Predicting author gender, time of authorship, or identify of author
- Analyzing sentiment about pharmaceutical drugs
- Extraction of apartment rental information from Craigslist ads
- Proposing a theoretical model for detecting click-through fraud
- Attempting the NetFlix challenge
- Collaborative filtering using beer recommendation reviews (from pintley.com)
- Attempting to predict the market using publicly disclosed financial documents (SEC filings, transcripts)
- Exploitation of open source information for maritime domain awareness - matching pictures of ocean-going vessels (from Flickr) to Coast Guard databases.

Empirical Ideas

- Can POS-tagging be used to improve IR performance?
- Can a given NLP technique improve performance (*e.g.,* keyword phrases or stemming)?
- Using electronic thesauri to automatically augment user queries
- Learning to spell correct or phrasify (add quotes to) user's web queries.
- Apply a machine learning algorithm (or algorithms) such as SVMs/NNs/Decision Trees for a problem in text classification
- Develop and test a method for spam filtering.
- Implement an algorithm for document similarity that we did not cover extensively in class, such as Cover Density Ranking or Latent-Semantic Indexing. Compare your results to some baseline method such as vector-cosine or an out-of-the-box IR package.

- Experiment with cross-language retrieval by manually translating some of the HW#3 queries into another language and using a bilingual dictionary or on-line translation system to translate queries back to English prior to search.
- Implement phrase-indexing efficiently (see work by Bahle et al.)
- Build a system for detecting when online reviews (Yelp, Travelocity) are likely to be genuine vs. fake.
- Build and evaluate a translation resource (*e.g.,* dictionary or parallel corpus) obtained from the Web
- Obtain a speech recognition package and run it on web-available audio files to support spoken retrieval.
- Help users visualize textual information (such as a retrieved document set)
- Develop a system for retrieval of music
- Build a Web collection of 100k HTML documents. Analyze it in significant detail.
- Attempt retrieval of stored images using image embeddings.
- Develop an information extraction system that learns a particular kind of fact from unstructured documents (*e.g.,* crimes: perpetrator, victim, date, officers involved)
- Build a system for collaborative filtering, to match people with similar interests, or to suggest movies, books, wines, etc... to an individual based comparing their profile with others
- Question answering for one particular type or question (e.g., how many or who).
- Using continuous vector representations (i.e., word embeddings) to improve text classification

Theory / Application Ideas

- What problems are current large-scale evaluations (like TREC) susceptible to?
- Develop a framework for retrieval against scanned document images
- Investigate retrieval of a specialized type of document (*e.g.,* a collection of source code or job openings).
- How can the relative efficacy of two *web* search engines be established?
- How can spam filtering software adapt to new changes (trends) from commercial spammers?
- How can commercial engines counter attempts to falsify click-throughs?

Instructor's Video Presentation Scoring Sheet

Student:

Topic:

1. Were the project's goals and motivation sufficiently explained? (1-10)

2. Was suitable and meaningful background information presented (e.g., prior work)? (1-10)

3. Did the presentation provide sufficient technical detail (1-5) and articulate a novel contribution or insight? (1-5)

4. Clarity of the presentation, quality of slides or AV/materials, video length is appropriate (1-10).

5. Was the work well thought out? Did conclusions follow from the argument or experimental results? (1-10)

6. Other comments.