**New Graduate Course Proposal**

**Stevens Institute of Technology**

Please submit this information electronically to the Office of Graduate Academics ([lromano@stevens.edu](mailto:jcuddy@stevens.edu)) requesting the proposal be added to the Graduate Curriculum Committee agenda. Allow at least 2 months for approval. There will be a first meeting to introduce and discuss the proposal, and a second meeting to consider for approval. In the case of suggested major revisions or collaboration across programs, a longer period of consideration may ensue. Please remember that the Committee’s agenda is set a week before each scheduled meeting. The GCC’s calendar fills rapidly.  ALL submissions, original or revised, must be sent to Lena Romano one week before the meeting at which they will be discussed.  Late submission will result in proposals being heard at subsequent meetings.

Proposals must be brought to the Committee by the departmental representative, but relevant stakeholders may attend meetings and present the proposals. If the proposal was developed in a different format for approval at the school level, the Committee still requires that this form be completed.

If there is parallel or similar course running at the undergraduate level, proposals must clearly indicate how the proposed course is a graduate course either in the course description itself or in a supplemental section below it.  What differentiates the proposed course from an undergraduate offering? Insuring clarity for the GCC will streamline discussion.

**School**: Schaefer School of Engineering and Science (SES)

**Course Title**: Text mining and information retrieval

**Program(s)**: MS in Computer Science / MS in Machine Learning / Ph.D. in Computer Science / Ph.D. in Data Science     

**Proposed Course # or Level**: CS 589    

**Catalog Description:** (This paragraph should succinctly summarize the course objectives and outcomes as described in this proposal, and list pre- or co-requisite courses in the relevant section below. Keep the length to no more than 100 to 150 words. Do not restate the list of topics.)

**Objective**: This course is a graduate-level course on fundamental techniques in information retrieval and text mining. By taking this course, students learn how to crawl, clean, process, mine, and infer knowledge from a massive amount of text data; how to build a search engine from scratch, including indexing, building retrieval models, and evaluating the performance of a search engine; they will also learn important machine learning and deep learning techniques for text data, including topic model, LSTM and BERT; finally, they will learn state-of-the-art research topics in text mining and information retrieval, and get research experience in these topics by working on the final project.

**Prerequisite**:

Undergraduate: CS115

Graduate: no prerequisite

**Course Objectives:** (This section should provide a description of what students will get out of the course beyond the course itself – i.e., how it will prepare them for their profession, or how the course fits in with the overall curriculum of the program(s) that the course belongs to.)

1. Understanding on the fundamental techniques behind a modern search engine
2. Hands-on experience on industry-standard technique(s) and infrastructure(s) for building a search engine.
3. Fundamentals for conducting research in text mining and information retrieval
4. This course would be the first course on information retrieval in the CS curriculum, thus filling this gap. Being related to CS584 (natural language processing), it can strengthen the natural language aspect of the CS program.

**List of Course Outcomes:** *A course outcome is an ability or skill that each student is expected to develop by taking the course.* Each course must have a list of outcomes developed by the instructor. The course outcome should be worded actively.[[1]](#footnote-1)

Students will be able to:

1. Perform basic text processing tasks such as crawling data from the web, pre-processing and cleaning natural language sentences;
2. Evaluate ranking algorithms by using information retrieval evaluation techniques, and implement text retrieval models such as TF-IDF and BM25;
3. Use Elastic search to implement a prototypical search engine on Twitter data;
4. Derive inference algorithms for the maximum likelihood estimation (MLE), implement the expectation maximization (EM) algorithm;
5. Use state-of-the-art tools such as LSTM/Bert for text classification tasks;

**Prerequisites**:[[2]](#footnote-2) Undergraduate students must have taken CS115, no prerequisite for graduate students

**Cross-listing**:        — show cross-listed course number(s)

**Grading Percentages**: HW 50% Mid-term 20% Projects 30%

Other  (specify both percent and kind of work)

**Credits**:  3 credits  Other

**For Graduate Credit toward Degree or Certificate**  
  **Yes**  No  Not for Dept. Majors  Other

**Textbook(s) or References** (List required and recommended texts including publisher and year in a recognized format such as APA, AIP, Chicago or MLA):

 Zhai, C., & Massung, S. (2016). Text data management and analysis: a practical introduction to information retrieval and text mining. Association for Computing Machinery and Morgan & Claypool.

**Mode of Delivery**  **Class**  Online Modules Other

**Program/Department Ownership:**  Computer Science    

**When first offered**:  This is a new course

**Department Point of Contact and Title:** **Xueqing Liu, Assistant Professor**

**Date approved by individual school and/or department curriculum committee**:        
If your school does not have a curriculum committee, please indicate date of approval by relevant stakeholders and identify them.

**Sample Syllabus**: This syllabus should be sufficiently detailed to allow the Graduate Curriculum Committee to understand and discuss the scope of the course, its aims and assignments. The homework and reading sections should provide sufficient detail for the Committee to judge the amount and kind of work required of students. The Committee understands that this syllabus is a sample of how a course might be organized, not a commitment to always offer the course exactly as described every time. Note that a syllabus is not merely a listing of topics or a restatement of the catalog description.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Topic(s)** | **Reading(s)** | **Class exercises (Optional)** | **HW** |
| Week 1 | Overview and history of IR, vector space model | Chapter 2 |  |  |
| Week 2 | TF-IDF, cosine similarity, probabilistic ranking principle, BM25 | Chapter 6 |  | HW1: data crawling, data cleaning for Stack Overflow data |
| Week 3 | IR evaluation, query expansion/completion | Chapter 9 |  |  |
| Week 4 | IR infrastructure:  inverted index, Elastic search, Page Rank | Chapter 7 |  |  |
| Week 5 | IR infrastructure: HITS; Relevance feedback | Chapter 10 |  | HW2: Elastic search: implement retrieval model using ES, evaluate your model’s performance |
| Week 6 | Pseudo relevance feedback, Feedback retrieval model, Expectation maximization (EM) algorithm: theory | Chapter 7 |  |  |
| Week 7 | EM algorithm: applications | Chapter 17 |  | HW3: implement semi-supervised topic model for SO data tag prediction |
| Week 8 | Midterm |  |  |  |
| Week 9 | Deep learning: recurrent neural network, seq2seq | Chapter 17 |  |  |
| Wee 10 | Deep learning: neural information retrieval |  |  | HW4: use LSTM/BERT for SO tag prediction |
| Week 11 | Frontier topics (for project): opinion mining, sentiment analysis | Chapter 18 |  |  |
| Week 12 | Frontier topics: neural machine translation/text generation | Chapter 11 |  | HW5: Stack Overflow question linking |
| Week 13 | Frontier topics: recommender system |  |  |  |
| Week 14 | Project presentation |  |  |  |

**Description of the final project**: There are two tracks for the final project - research track and engineering track. Research track follows the following 6-step process: (1) After midterm, students choose a topic from one of the topics taught in class, i.e., text classification, opinion mining, recommender system, etc.. (2) for each topic, they pick 2-3 coherent papers from the recent conferences in NLP/data mining/IR, and write a summary for the paper; (3) the students who share the same interest are categorized into groups; (4) each group propose a novel research topic motivated by their survey, (5) deliver a presentation in Week 14, and (6) submit their implementation (code in Python) as well as an 8-page academic paper as their final project. The engineering track follows Step 1-3, and submit an 8-page report as well as an engineering project such as a website.

1. Good examples can be taken from descriptions of Bloom’s taxonomy. For example, see the table located at: <http://www.nwlink.com/~donclark/hrd/bloom.html>. [↑](#footnote-ref-1)
2. You may provide a list of courses, competencies or other criteria (e.g., “Students must have taken CS 6XX” or “Students must have taken a course in thermodynamics,” or “Students must be part of a certain cohort.”) [↑](#footnote-ref-2)