UNIVERSITY OF
SOUTH DAKOTA

CSC 785: Information Storage and Retrieval

Final Project

Objective:

The final project is a capstone assignment that allows students to apply the theoretical and practical knowledge gained throughout the course. Students will design, implement, and evaluate an Information Retrieval (IR) system or conduct original research on an IR topic. The project should demonstrate mastery of indexing, retrieval models, evaluation techniques, and possibly modern topics such as neural IR, semantic search, or domain-specific retrieval.

Project Options

You may choose **one** of the following directions:

1. System Development

Design and build a small-to-medium-scale IR system. This could include:

- A search engine for a domain-specific corpus (e.g., legal texts, scientific articles, reviews)
- A question-answering system over structured or unstructured data
- A personalized or recommendation-based retrieval system
- Integration of traditional IR models (e.g., BM25) with neural re-ranking or embedding-based methods

2. Research Paper

Conduct an in-depth investigation into a focused topic in Information Retrieval. You may choose **one** of the following research directions:

a) Literature Review

Survey and synthesize recent research (ideally from the last 5–7 years) in a specific subfield of IR, such as:

- Neural retrieval methods
- Evaluation metrics and benchmarking
- IR for low-resource languages or specific domains (e.g., biomedical, legal)
- Fairness, bias, or interpretability in search systems

Your paper should:

- Summarize key models, trends, and findings
- Compare and critique different approaches
- Identify open problems and research gaps

b) Empirical Study

Design and carry out an experimental study, such as:

- Reproducing and comparing baseline IR models (e.g., BM25, DPR, ColBERT)
- Evaluating retrieval performance across datasets or domains
- Analyzing the behavior of search algorithms under different conditions
- Studying efficiency–effectiveness trade-offs

Your study must include:

- Clear research questions or hypotheses
- Sound experimental design and evaluation
- Use of public datasets and/or tools
- Statistical or comparative analysis

c) New Technique or Model

Propose a new method or variant in the IR pipeline, such as:

- A new ranking or re-ranking model
- A hybrid sparse–dense retrieval technique
- Improved embedding generation or similarity computation
- Integration of LLMs in retrieval (e.g., prompt-driven re-ranking, LLM-as-index)
- A new evaluation framework or metric

Your contribution should be:

- Clearly motivated and grounded in existing work
- Technically detailed and implemented (prototype or proof-of-concept)
- Evaluated against appropriate baselines or metrics

Note: This option is encouraged for students interested in research publication or further academic work. Ambitious ideas are welcome but must be feasible within the semester timeframe.

Datasets

Suggested datasets: TREC, MS MARCO, Wikipedia Dumps, Common Crawl, arXiv, PubMed, Stack Overflow, Reddit

Evaluation Criteria

Your final project will be graded based on the following criteria:

1. **Problem Choice and Relevance**: The significance of the problem you selected and how well IR is applied to it.
2. **Related work:** Is the related work discussed and evaluated adequately? Is the paper's contribution compared and contrasted with prior work in the field?

3. **Algorithm Implementation (if applicable)**: How well you have implemented the chosen IR algorithm(s), including any modifications or original contributions.
4. **Experimental Design and Results**: Quality of your experiments, evaluation metrics, and the clarity of the results and insights drawn.
5. **Clarity and Completeness of Report**: How well you communicate your findings and the depth of your analysis in the report. Adherence to IEEE format, clarity, and coherence of the writing
6. **Presentation**: Effectiveness in communicating your work during the presentation.

Deliverables:

- A **technical report** written in IEEE conference format.
  Template: https://www.ieee.org/conferences/publishing/templates.html

  It should typically include:

  1. Abstract
  2. Introduction and motivation
  3. Related work
  4. System design or methodology
  5. Experiments and results
  6. Discussion and future work
  7. References

- A **10–15 Minute recorded presentation** summarizing your final technical report. All team members must participate and present a portion of the material.

  A simple way to do this is by setting up a Zoom meeting with all team members, recording the session while presenting and sharing your screen to display your presentation slide deck. Once complete, save the video as an .mp4 file and submit the video file.

- A jupyter/colab notebook (.ipynb)) or .py file containing the code used for implementation (if applicable). Ensure your code runs with no errors and the results match what is listed in your paper.

Submission Guidelines:

Only one team member should submit all project deliverables to the designated folder on D2L. This helps prevent duplicate submissions and avoids issues related to self-plagiarism.

Academic Integrity

Please remember that academic integrity is paramount. Any form of plagiarism, including reusing code from online repositories or other students' work without attribution, will result in penalties according to the academic integrity policy.

I look forward to seeing what you create!
Please feel free to reach out for guidance and support throughout the research/implementation and writing process.

Suggested Project Topics

1. Classical and Neural Retrieval Models

- Comparison of BM25, TF-IDF, and dense embedding-based models
- Fine-tuning transformer-based retrievers (e.g., BERT, DPR, ColBERT) for domain-specific corpora
- Evaluating hybrid retrieval systems (e.g., sparse + dense fusion)
- Learning-to-rank approaches using LambdaMART, RankNet, or GBDT

2. Evaluation and Benchmarks

- Reproducing and extending results from the BEIR benchmark
- Designing a new evaluation metric tailored to a specific application (e.g., legal IR, medical search)
- Comparative analysis of evaluation tools (e.g., TREC_eval vs. pytrec_eval)
- Error analysis of ranking results in neural IR models

3. IR for Specialized Domains

- Building a search engine for legal documents, scientific papers, or historical archives
- Domain adaptation techniques for biomedical or low-resource domain retrieval
- IR in multilingual or cross-lingual settings
- Semantic search over programming Q&A sites like Stack Overflow or GitHub

4. IR + LLMs / Generative Retrieval

- Integrating LLMs (e.g., GPT-4, Claude) with retrieval pipelines (e.g., RAG)
- Prompt engineering for retrieval-augmented generation
- LLM-as-index: using large models to directly answer questions without a retriever
- Evaluating hallucination and factual consistency in generative IR systems

5. Efficiency and Scalability

- Index compression techniques and memory-efficient retrieval
- Vector search optimization using tools like FAISS, ScaNN, or Annoy
- Approximate nearest neighbor (ANN) retrieval trade-offs
- Streaming or real-time search over dynamic datasets

6. Fairness, Ethics, and Interpretability

- Investigating bias in search results and ranking algorithms
- Explainable IR: making retrieval decisions interpretable
- Privacy-preserving IR using federated learning or anonymization
- Detecting and mitigating toxic or harmful content in retrieved results

7. Personalization and Recommendation

- Context-aware or session-based retrieval models
- Query understanding and user intent prediction
- Building personalized document or product retrieval systems
- Evaluation challenges in personalized IR

You may also propose **your own topic** that extends beyond this list, as long as it aligns with the course themes.