KHAI THÁC DỮ LIỆU VĂN BẢN VÀ ỨNG DỤNG

HƯỚNG DẪN THỰC HIỆN ĐỒ ÁN ĐÁNH GIÁ MÔN HỌC

1. PHẦN 1: ĐỒ ÁN LÝ THUYẾT (30%)

Thời gian thực hiện: 4 tuần.

Dựa trên các bài báo (paper) được giao về chủ đề của đề tài thuộc lĩnh vực Khai thác dữ liệu văn bản, được phân cho mỗi nhóm. Tiến hành tìm hiểu về bài toán của đề tài được giao theo các bước:

- **B1 (10%): Tìm kiếm tài liệu (Reference):**
  - Dựa trên đề tài, tìm kiếm các bài báo liên quan của đề tài.
  - Ưu tiên các paper/journal từ hội nghị tạp chí (conference) có uy tín, luận văn từ thạc sĩ/tiến sĩ trở lên hoặc từ các tác giả có uy tín có chỉ số H-index cao, hoặc các bài báo có số citation cao (được trích dẫn lại nhiều) trong ngành.
  - Là bài báo có ảnh hưởng lớn giúp giải quyết bài toán.
  - Ưu tiên các phương pháp đạt hiệu suất state-of-the-art tại thời điểm công bố.
  - Các trang, nguồn gợi ý tìm kiếm:
    - Google Scholar (https://scholar.google.com/)
    - Paper with code (https://paperswithcode.com): chứa paper và link source code (nếu có)
    - ACM Digital Library (https://dl.acm.org/)
    - arXiv (http://arxiv.org/)
    - https://dblp.org/
    - Science Direct.
    - IEEE .
    - .....
  - Các bài báo tổng hợp bài toán có thể là các bài Survey, Overview, Study, Review.
  - Cần tìm kiếm ít nhất 5-10 bài báo có liên quan đề tài, trong thời gian gần đây. Càng nhiều bài báo, bài báo càng chất lượng thì càng tốt.
  - Lưu lại danh sách tên bài báo/tác giả/hội nghị/năm xuất bản/link bài báo.

- **B2 (50%): Khảo sát bài toán. Tìm hiểu về các nội dung sau:**
  - Định nghĩa bài toán (Introduction):
    - Bài toán giải quyết vấn đề gì?
  - Động lực bài toán (Motivation)
    - Tại sao cần giải quyết bài toán? Tính cấp thiết?
    - Ý nghĩa về mặt khoa học?
    - Ý nghĩa về mặt thực tiễn?
    - Các Ứng dụng trong thực tế.
  - Phát biểu bài toán (Problem Definition)
    - Input/Output bài toán?
    - Thể hiện toán học của bài toán? ẩn số bài toán?
    - Sơ đồ bài toán.
    - Loại bài toán.
  - Thách thức bài toán (Challenges).
    - Bài toán có các thách thức/vấn đề/khó khăn gì?
    - Vấn đề nào là quan trọng? (được nhiều bài báo tập trung giải quyết)

        - Nêu ví dụ minh hoạ.
    - Công trình liên quan (Related works):
        - Cách giải quyết thách thức bài toán?
        - Các phân nhóm bài toán con.
        - Các Phương pháp tiếp cận? Tổng quan theo 2 hướng/giai đoạn:
            - Máy học truyền thống (machine learning, từ 2013 về trước)
            - Học sâu (deep learning, từ 2013 về sau)
        - Liệt kê và tóm tắt một vài (từ 3 trở lên) phương pháp của các bài báo:
            - Phương pháp giải quyết?
            - Giải quyết thách thức gì
            - Ít nhất có từ 1-2 phương pháp của bài báo 3 năm trở lại đây.
            - Hiệu suất của phương pháp? Có đạt state-of-the-art?
            - Ưu điểm, nhược điểm của phương pháp.
        - Tông quan nghiên cứu bài toán trong nước (Việt Nam).
        - Tổng hợp thành dạng bảng các phương pháp, gồm các cột:

Phương pháp (tên/tóm tắt)	Năm công bố (tăng dần)	Thang đo đánh giá	Tập dữ liệu	Hiệu năng

    - Tài nguyên bài toán (Resources):
        - Các tập dữ liệu (dataset) phổ biến?
        - Cuộc thi (competition) liên quan.
        - Framework xây dựng
    - Đánh giá bài toán (Metrics/Evaluation):
        - Các thang đo đánh giá hiệu quả mô hình.
    - Hướng phát triển tương lai (Future works/trends).
    - Các nội dung khác nhóm tìm hiểu được (cộng điểm).
- **B3 (30%): Trình bày và vấn đáp.**
    - Trình bày các nội dung đã tìm hiểu ở bước 2.
    - Hình thức: trình chiếu (powerpoint, ...).
        - Không dùng word/PDF để trình bày -> (0d)
        - Không quay trước video.
    - Thời gian: 15-20 phút (tối đa).
    - Tất cả các thành viên đều trình bày.
    - Vấn đáp: 10 phút:
        - Nhóm: liên quan Nội dung đã tìm hiểu của bài toán.
        - Cá nhân: nội dung đã tìm hiểu, vấn đáp kiến thức cơ bản liên quan đã học (máy học, xử lí văn bản, khai thác văn bản...)
    - Các nhóm đặt câu hỏi.
    - Demo bài toán hoặc demo một số ứng dụng của bài toán (nếu có).
- **B4 (10%): Báo cáo**
    - Viết thành báo cáo nội dung đã tìm hiểu.
    - Nội dung theo các mục lớn của bước 2.
    - Trích dẫn tài liệu liên quan đã tìm hiểu.
    - Nộp file PDF + Powerpoint trình chiếu.
    - Không nộp báo cáo → Trừ 50% số điểm.

2. PHẦN 2: ĐỒ ÁN THỰC HÀNH (50%)

Dựa trên đề tài đã tìm hiểu từ Phần 1.

Thời gian thực hiện: 2-3 tuần.

- **B1 (5%): Lựa chọn phương pháp.**
  - **Lựa chọn ít nhất 2 trong số các phương pháp đã tìm hiểu từ Phần 1.**
  - Nên chọn một trong 2 hướng:
    - Hướng 1: Từ 2 phương pháp học sâu trong khoảng 5 năm gần đây.
    - Hướng 2: Một phương pháp truyền thống và 1-2 phương pháp học sâu.
  - Ưu tiên các phương pháp đạt hiệu suất state-of-the-art tại thời điểm công bố.
  - Ưu tiên các phương pháp có source code mẫu, data mẫu, hướng dẫn chạy mẫu.
  - Nếu là mô hình học sâu, Ưu tiên các mô hình có công bố mô hình đã huấn luyện trước (pretrained).
  - Ưu tiên được xây dựng trên Framework là thế mạnh bản thân.

- **B2 (25%): Tìm hiểu về phương pháp.** Dựa trên các phương pháp đã lựa chọn B1, tìm hiểu:
  - Động lực giải quyết bài toán của phương pháp.
  - Các vấn đề phương pháp giải quyết
    - Suy ra đóng góp của tác giả.
  - Nội dung cụ thể phương pháp:
    - Mô hình bài toán.
    - Các phương pháp mới được đề xuất (Methods).
    - Sự cải tiến so với các hướng tiếp cận trước đây.
    - Phương pháp thực nghiệm (Experiments):
      - Tập dữ liệu.
      - Cách huấn luyện.
      - Thang đo đánh giá.
      - Kết quả thực nghiệm.
  - So sánh các phương pháp được lựa chọn:
    - Mô hình (kiến trúc, trọng số,...)
    - Kết quả hiệu năng.
    - Ưu điểm, nhược điểm.

- **B3 (45%): Thực nghiệm phương pháp.**
  - Lựa chọn tập trung ít nhất 1 phương pháp từ B2.
  - Chạy code thực nghiệm.
  - Hướng thực nghiệm:
    - *Hướng 1:* Tinh chỉnh một phần kiến trúc mô hình và huấn luyện lại trên tập dữ liệu trong bài báo để đánh giá sự cải tiến.
      - Ý tưởng tinh chỉnh kiến trúc (vì sao?).
      - Giải thích sự thay đổi kết quả (tăng / giảm).
      - Đánh giá kết quả.
    - *Hướng 2:* Giữ nguyên kiến trúc, chạy lại mô hình trên một tập dữ liệu mới (mà bài báo chưa chạy) và đánh giá.
      - Thông tin tập dữ liệu mới (cộng điểm nếu là dữ liệu tiếng Việt).
      - Kết quả huấn luyện (minh chứng, biểu đồ huấn luyện, hiệu năng,...).
      - So sánh kết quả với các phương pháp khác cùng tập dữ liệu.
      - Đánh giá kết quả.

- *Hướng 3:* Kết hợp hướng 1 và 2 → điểm cộng.

- Nếu mô hình không có code được công bố: nhóm tự viết code, tự huấn luyện → cộng rất nhiều điểm.

- **B4 (15%): Vấn đáp và trình bày.**
  - Trình bày trực tiếp.
  - Hình thức: Powerpoint + Demo code + Demo chạy thử.
  - GV yêu cầu chạy thử trên một dữ liệu bất kì.
  - Minh chứng về các hướng thực nghiệm.
  - Vấn đáp: giải thích code, kết quả thực nghiệm.
  - Không tham gia vấn đáp → 0 điểm.

- **B5 (10%)**: **Báo cáo**
  - Viết thành báo cáo nội dung đã tìm hiểu.
  - Nội dung theo các mục của bước 2 và bước 3.
  - Nộp file PDF báo cáo + Powerpoint + source code.
  - Không nộp báo cáo → Trừ 50% số điểm.