Báo cáo Tổng hợp Câu hỏi và Trả lời

# Câu 1: Mục tiêu của bài toán là gì? Xác định rõ input và output là gì?

Mục tiêu của bài toán là thực hiện tóm tắt văn bản bằng phương pháp rút trích, rút gọn nội dung văn bản dài thành đoạn tóm tắt ngắn gọn nhưng vẫn giữ được ý chính và thông tin quan trọng.  
- Input: Văn bản gốc dưới dạng file XML chứa các câu được đánh dấu.  
- Output: Bản tóm tắt gồm một số câu quan trọng được chọn lọc từ văn bản gốc.

# Câu 2: Phương pháp tiếp cận bài toán là gì? Mô tả ý tưởng chính của phương pháp tiếp cận?

Phương pháp tiếp cận sử dụng kỹ thuật tóm tắt trích xuất dựa trên đồ thị và thuật toán PageRank để đánh giá mức độ quan trọng của từng câu trong văn bản.  
Ý tưởng chính: Tính toán độ tương đồng giữa các câu dựa trên đặc trưng TF-IDF, xây dựng đồ thị câu trong đó các câu là nút, các cạnh là độ tương đồng trên ngưỡng, sau đó dùng PageRank để xếp hạng câu và chọn ra các câu quan trọng.

# Câu 3: Mô tả chi tiết các bước thực hiện của phương pháp đã tiếp cận ở trên

Trả lời:

1. Đọc và trích xuất câu từ file XML:
   * Sử dụng biểu thức chính quy để tìm và lấy các câu nằm trong thẻ <s>...</s> từ file XML đầu vào.
   * Lưu các câu này vào danh sách để xử lý tiếp.
2. Tiền xử lý câu:
   * Tokenize từng câu, tức là tách câu thành các từ đơn (tokens), loại bỏ dấu câu và chuyển về chữ thường để chuẩn hóa dữ liệu.
3. Tính TF (Term Frequency) cho mỗi câu:
   * Đếm số lần xuất hiện của mỗi từ trong câu.
   * Tính tỷ lệ xuất hiện của từ trên tổng số từ trong câu, tạo thành vector TF cho câu đó.
4. Tính IDF (Inverse Document Frequency) cho toàn bộ câu:
   * Đếm số câu chứa từ đó.
   * Tính IDF dựa trên công thức: IDF(term) = log(N / (1 + số câu chứa term)) với N là tổng số câu.
5. Tính TF-IDF vectors:
   * Nhân TF của từng từ với IDF tương ứng để có vector đặc trưng cho mỗi câu.
6. Tính ma trận cosine similarity giữa các câu:
   * Dùng cosine similarity để đo độ giống nhau giữa từng cặp câu dựa trên TF-IDF vectors.
7. Xây dựng đồ thị câu:
   * Mỗi câu là một nút trong đồ thị.
   * Tạo cạnh giữa hai câu nếu cosine similarity của chúng lớn hơn hoặc bằng ngưỡng threshold.
8. Áp dụng thuật toán PageRank trên đồ thị:
   * Khởi tạo điểm PageRank đều cho tất cả câu.
   * Lặp cập nhật điểm dựa trên điểm của các câu liên kết, trọng số cạnh và hệ số giảm dần (damping factor).
   * Dừng khi điểm hội tụ hoặc đạt số vòng lặp tối đa.
9. Chọn các câu có điểm PageRank cao nhất:
   * Sắp xếp các câu theo điểm PageRank giảm dần.
   * Lấy ra số câu cần tóm tắt dựa trên số câu của bản tóm tắt mẫu (do giáo viên cung cấp) hoặc theo yêu cầu.
10. Tạo bản tóm tắt:
    * Ghép các câu đã chọn theo thứ tự xuất hiện trong văn bản gốc để tạo bản tóm tắt cuối cùng.

# Câu 4: Code được >= 5 đặc trưng biểu diễn dữ liệu/ biểu diễn được văn bản đồ thị

Chương trình biểu diễn dữ liệu văn bản theo:  
- Đặc trưng TF-IDF của từng câu.  
- Ma trận độ tương đồng cosine giữa các câu.  
- Đồ thị câu, trong đó mỗi nút là một câu và các cạnh có trọng số dựa trên độ tương đồng cosine.  
- Thuật toán PageRank sử dụng trọng số cạnh để tính điểm xếp hạng câu.  
Đây là cách biểu diễn văn bản dưới dạng đồ thị trọng số phục vụ cho việc xếp hạng câu.

# Câu 5: Áp dụng được bất kỳ phương pháp phân lớp dữ liệu trong thư viện máy học/ Xếp hạng được từ trong đồ thị theo mức độ quan trọng

Chương trình sử dụng thuật toán PageRank - một thuật toán dựa trên đồ thị để xếp hạng các câu theo mức độ quan trọng.  
PageRank không phải là mô hình học máy truyền thống, nhưng là phương pháp phân loại phi giám sát dựa trên cấu trúc liên kết.

# Câu 6: Lấy được tóm tắt văn bản

Chương trình thực hiện tóm tắt văn bản trích xuất bằng cách chọn các câu có điểm PageRank cao nhất.  
Quy trình:  
- Tính điểm PageRank cho từng câu trên đồ thị.  
- Sắp xếp các câu theo điểm giảm dần.  
- Chọn số câu cần tóm tắt theo yêu cầu hoặc theo bản tóm tắt mẫu.  
- Ghép các câu đã chọn theo thứ tự ban đầu thành bản tóm tắt hoàn chỉnh.  
Kết quả tóm tắt được hiển thị trên giao diện người dùng.

# Câu 7: Nhận xét về kết quả đạt được: độ chính xác, ưu nhược điểm của phương pháp đang áp dụng

Độ chính xác:  
- Phương pháp PageRank giúp chọn các câu đại diện tốt, giữ được nội dung chính.  
Ưu điểm:  
- Dễ triển khai, không cần dữ liệu huấn luyện.  
- Xây dựng dựa trên quan hệ nội bộ câu nên tóm tắt nhất quán.  
- Phù hợp với nhiều loại văn bản.  
Nhược điểm:  
- Chỉ chọn câu nguyên bản, bản tóm tắt có thể dài và chưa cô đọng.  
- Phụ thuộc ngưỡng threshold và số câu chọn.  
- Không xử lý ngữ cảnh phức tạp hoặc diễn giải sâu.  
Tóm lại, phù hợp với tóm tắt nhanh, nếu cần tóm tắt tự nhiên hơn có thể dùng kỹ thuật học sâu.

# Câu 8: Cải tiến phương pháp đang áp dụng, chẳng hạn bổ sung thêm đặc trưng biểu diễn dữ liệu hoặc thêm trọng số vào đồ thị

Các cải tiến có thể áp dụng:  
- Sử dụng embedding hiện đại như Word2Vec, GloVe hoặc BERT để biểu diễn câu, giúp nắm bắt ngữ cảnh sâu hơn.  
- Kết hợp thêm đặc trưng ngữ pháp hoặc vị trí câu trong đoạn văn.  
- Điều chỉnh trọng số đồ thị kết hợp các yếu tố khác như tần suất từ khóa hoặc vị trí câu.  
- Áp dụng threshold động dựa trên phân bố similarity.  
- Kết hợp PageRank với các thuật toán học máy hoặc clustering để tránh lặp thông tin.  
- Cải thiện tiền xử lý như loại bỏ stopwords, chuẩn hóa từ, phân tích cú pháp.