|  |  |
| --- | --- |
| TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN  **KHOA KHOA HỌC MÁY TÍNH** | **ĐỀ THI GIỮA/CUỐI KỲ**  **HỌC KỲ 1 – NĂM HỌC 2019-2020**    Môn thi:Truy vấn thông tin đa phương tiện  Mã lớp: CS336.K11  Thời gian làm bài: 90 phút  (*Sinh viên không được sử dụng tài liệu)* |

**Câu hỏi 1:** (2 điểm, 1+1)

1. Mô tả ngắn gọn các bước chính để xây dựng một hệ thống tìm kiếm văn bản.
2. Theo Anh/Chị, khuyết điểm chung lớn nhất của mô hình Vector Space và Boolean Space là gì? Hãy phân tích khuyết điểm này và cho biết giải pháp khắc phục.

**Câu hỏi 2:** (3 điểm, 1+2)

Anh/Chị hãy ***trả lời ngắn gọn*** các câu hỏi sau:

1. Hãy nêu ngắn gọn các bước để tiền xử lý văn bản trên tiếng Việt trước khi đánh chỉ mục.
2. Trình bày phương pháp hậu xử lý mở rộng câu truy vấn (query expansion). Hãy cho biết phương pháp này thường dùng để cải tiến hiệu quả của hệ thống truy vấn ở độ chính xác (precision) hay độ phủ (recall)?

**Câu hỏi 3:** (3 điểm, 2+1)

Cho hai hệ thống tìm kiếm khác nhau, A và B. Kết quả tìm kiếm của A và B với một câu truy vấn Q được thể hiện trong 2 danh sách sau:

- Kết quả của hệ thống A:

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Thứ tự | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 |
| Phù hợp | X | -- | -- | -- | X | X | X | X | -- | X | -- | X | X | X | X |

- Kết quả của hệ thống B:

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Thứ tự | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 |
| Phù hợp | X | X | -- | X | X | -- | -- | -- | -- | X | X | X | X | X | X |

- Các văn bản được đánh dấu ‘X’ được xem là phù hợp với câu truy vấn Q. Ngược lại, các văn bản đánh dấu ‘--‘ là không phù hợp với câu truy vấn Q. Biết rằng tổng số văn bản có liên quan đến truy vấn là 10.

Anh/Chị hãy:

1. Đánh giá hiệu quả của các hệ thống A và B bằng phương pháp Average Precision.
2. Theo Anh/Chị hệ thống tìm kiếm nào tốt hơn? Tại sao ?

**Câu hỏi 4**: (2 điểm)

Trình bày các bước để tính vector biểu diễn TF-IDF cho các văn bản dưới đây. Biết rằng tập hợp các term bao gồm {A, B, C, D, E}.

* Doc 1: A, B, D, B, A
* Doc 2: C, E, D, C
* Doc 3: B, C, A, B, A
* Doc 4: A, B, E

**HẾT**

**Duyệt đề Giảng viên ra đề**