BÀI TẬP ÔN TẬP LÝ THUYẾT

1. KHAI THÁC NỘI DUNG WEB

2. KHAI THÁC CẤU TRÚC WEB

1. KHAI THÁC NỘI DUNG WEB

Hệ thống truy vấn thông tin

Cho nội dung 3 tài liệu và một câu truy vấn như sau:

D1: "*An information retrieval model governs how a document and a query are represented and how the relevance of a document to a user query is defined.*"

D2: "*Information retrieval is the area of study concerned with searching for documents, for information within documents, and for metadata about documents, as well as that of searching structured storage, relational databases, and the World Wide Web.*"

D3: "*Web search has become very important in the information age. Increased exposure of pages on the Web can result in significant financial gains and/or fames for organizations and individuals.*"

Q: "*information retrieval on web*"

Bài tập 1. Thực hiện các bước tiền xử lý tài liệu trên: rút trích từ để xây dựng tập ngữ vựng V, loại bỏ stopword, stemming.

Danh sách stopword: how, and, or, an, a, the, there, that, of, for, to, is, are, can, has, with, within, in, on, about, as, well, very.

Thực hiện stemming theo các quy tắc sau:

- Chuyển danh từ từ số nhiều thành số ít: -s, -es. Ví dụ: books → book, classes → class, ponies → pony.
- Chuyển động từ từ các thể về thể gốc: -ing, -ed, bất quy tắc. Ví dụ: running → run, used → use, written → write.

Bài tập 2. Lập bảng chỉ mục đảo với mỗi posting chứa id tài liệu, tần số xuất hiện, và các vị trí xuất hiện.

Bài tập 3. Sử dụng bảng chỉ mục đảo trên để tìm kiếm các tài liệu chứa các câu truy vấn sau (liệt kê các posting):

a) Information AND retrieval
b) Web OR (Document AND Retrieval)
c) Search AND NOT Document

Bài tập 4. Hình thành vector trọng số TF-IDF cho từng tài liệu và câu truy vấn trên. *Lưu ý: câu truy vấn sử dụng giá trị TF-IDF tính được từ tập tài liệu (D1, D2, D3).*

Bài tập 5. Tính toán độ đo cosin và cho biết tài liệu nào sẽ được trả về tương ứng với câu truy vấn Q.

Đánh giá hệ thống truy vấn thông tin

Cho tập kết quả trả về của một công cụ tìm kiếm dựa trên nội dung như sau:

Rank i	+/-	Rank i	+/-	Rank i	+/-	Rank i	+/-
1	+	8	+	15	-	22	-
2	+	9	-	16	-	23	+
3	+	10	+	17	+	24	+
4	+	11	+	18	+	25	-
5	-	12	-	19	+	26	-
6	-	13	-	20	-	27	-
7	-	14	+	21	-	28	-

Bài tập 6. Tính toán độ chính xác, độ phủ và độ F cho từng hạng. Giả sử có 13 tài liệu thực sự liên quan đến truy vấn (kí hiệu dấu '+')

Bài tập 7. Tính độ chính xác trung bình.

Bài tập 8. Vẽ đường cong tương quan giữa độ phủ và độ chính xác.

Các bài toán khác

Bài tập 9. Cho bảng dữ liệu phân lớp các tài liệu từ d_1 – d_{19} (mỗi tài liệu đã được biểu diễn dưới dạng vector TF-IDF) như sau:

	history	science	research	offers	students	hall	Class
d1	0	0.537	0.477	0	0.673	0.177	A
d2	0	0	0	0.961	0.195	0.196	B
d3	0	0.347	0.924	0	0.111	0.112	A
	0	0.975	0	0	0.155	0.158	A
	0	0	0	0.780	0.626	0	B
	0	0.989	0	0	0.130	0.067	A
	0	0	0	0	1	0	B
.	0	0	1	0	0	0	A
.	0	0	0	0.980	0	0.199	B
.	0	0.849	0	0	0.528	0	A
	0.991	0	0	0.135	0	0	B
	0	0.616	0.549	0.490	0.198	0.201	A
	0	0	0	0.928	0	0.373	B
	0.970	0	0	0	0.170	0.172	B
	0.741	0	0	0.658	0	0.136	B
	0	0	0.894	0	0.315	0.318	A
	0	0.933	0.348	0	0.062	0.063	A
	0	0	0.852	0.387	0.313	0.162	A
	0	0	0.639	0.570	0.459	0.237	A
d_{20}	0	0	0	0	0.967	0.254	? (B)

Sử dụng thuật toán *k* láng giềng gần nhất để xác định xem tài liệu d_{20} thuộc về lớp nào với k = 5.

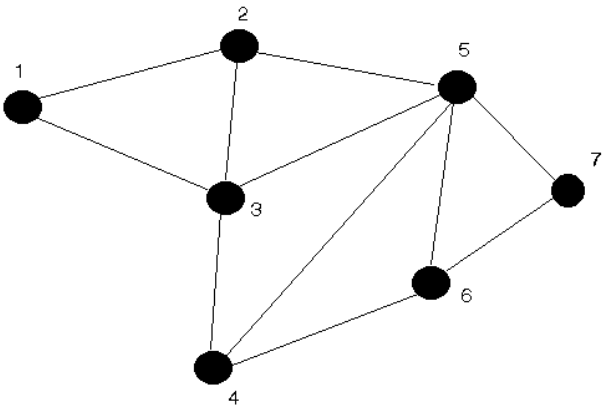
Bài tập 10. Cho tập dữ liệu với các vector thành phần TF-IDF sau:

Tài liệu	Tên Lửa	Cầu Thủ	Hoa Hồng
d_1	0.35	0.03	0.67
d_2	0.05	0.12	0.33
d_3	0.19	0.06	0
d_4	0.58	0.41	0.33
d_5	0	0	0.67
d_6	0.33	0.15	0.33
d_7	1	1	0
d_8	0.81	0.65	0.33
d_9	0.91	0.82	0
d_{10}	0.12	0.06	1

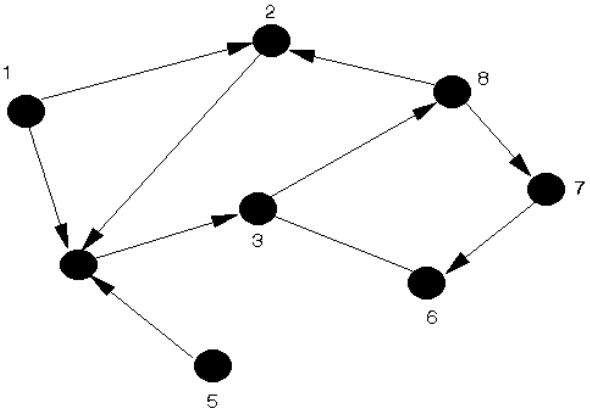
Sử dụng k-means với k = 2 để thực hiện gom nhóm các tài liệu trên. Khoảng cách giữa các tài liệu là độ đo cosin.

2. KHAI THÁC CẤU TRÚC WEB

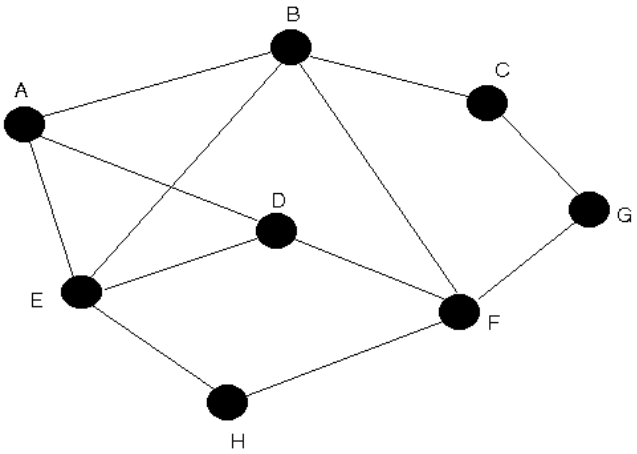
CÁC ĐỘ ĐO TÍNH TRUNG TÂM



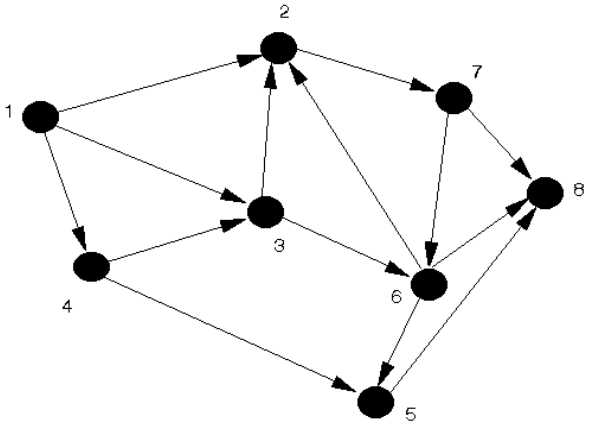
(a)



(b)



(c)



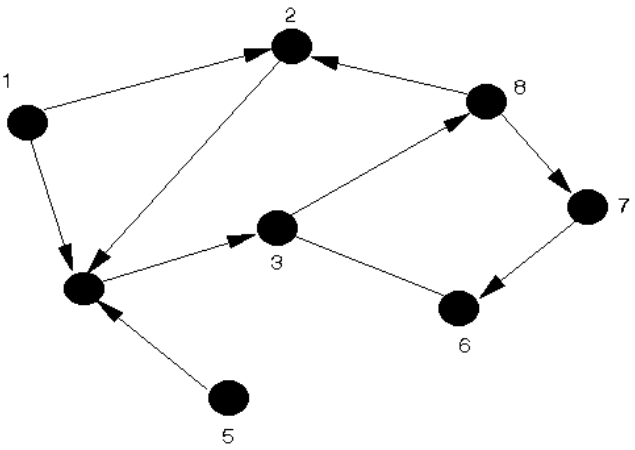
(d)

Bài tập 1. Hãy tính giá trị tính trung tâm bậc (degree centrality) cho các đỉnh trong đồ thị (a), (b), (c), (d)

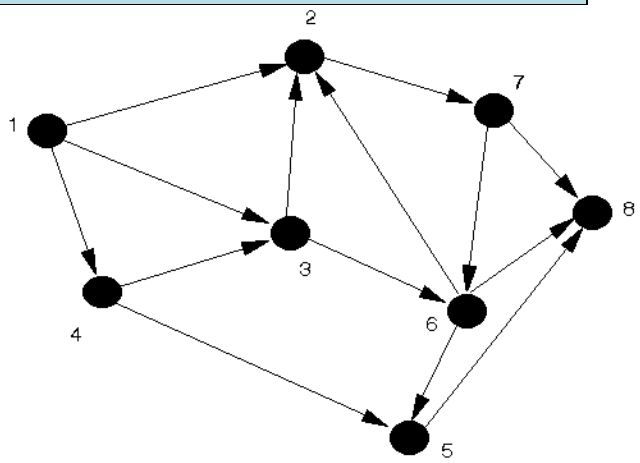
Bài tập 2. Hãy tính giá trị tính trung tâm gần (closeness centrality) cho các đỉnh trong đồ thị (a), (b), (c), (d)

Bài tập 3. Hãy tính giá trị tính trung tâm trung gian (betweeness centrality) cho các đỉnh trong đồ thị (a), (b), (c), (d)

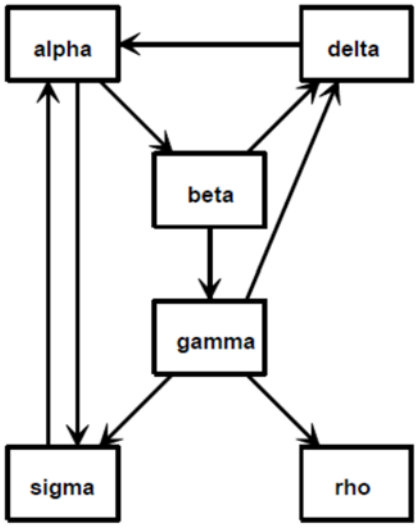
CÁC ĐỘ ĐO TÍNH TRUNG TÂM



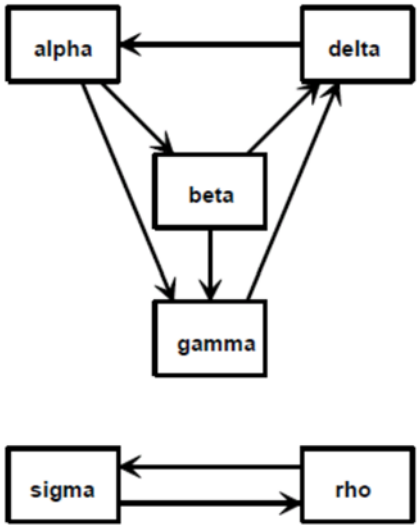
(e)



(f)



(g)



(h)

Bài tập 4. Hãy tính giá trị tính uy tín bậc (degree prestige) cho các đỉnh trong đồ thị (e), (f), (g), (h)

Bài tập 5. Hãy tính giá trị tính uy tín lân cận (proximity prestige) cho các đỉnh trong đồ thị (e), (f), (g), (h)

Bài tập 6. Hãy tính giá trị tính uy tín hạng (rank prestige) cho các đỉnh trong đồ thị (e), (f), (g), (h)

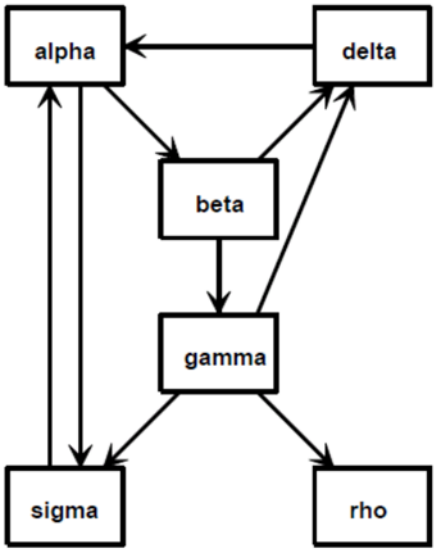
ĐỒNG TRÍCH DẪN VÀ LIÊN KẾT DANH MỤC

Bài tập 7. Hãy lập ma trận đồng trích dẫn C_{ij} cho đồ thị (e), (f), (g), (h) trong phần trên.

Bài tập 8. Hãy lập ma trận liên kết danh mục B_{ij} cho đồ thị (e), (f), (g), (h) trong phần trên.

CÁC THUẬT TOÁN XẾP HẠNG

Bài tập 9. Cho đồ thị như bên dưới



a. Thuật toán PageRank

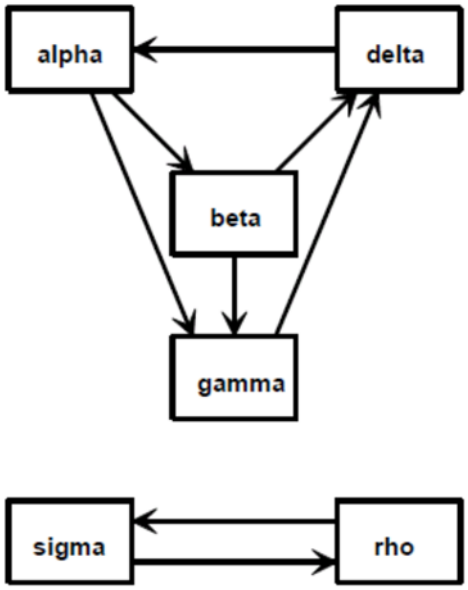
- Hãy thiết lập công thức PageRank cho các đỉnh của đồ thị, biết công thức tổng quát là: $PR(P) = \frac{d}{N} + (1-d)\sum_{u \in In(P)} \frac{PR(u)}{OutDeg(u)}$
  - *N là số nút trong đồ thị hay tổng số trang trên web*
  - *d gọi là hệ số tắt dần ∈ [0,1]*
  - *In(P) là tập đỉnh có liên kết đến P*
  - *OutDeg (u) là số liên kết ngoài của trang u*

- Tính giá trị của các trang qua 5 vòng lặp, biết rằng giá trị PageRank khởi tạo tại vòng lặp thứ 0 cho các trang đều bằng 1.

b. Thuật toán HITS

- Hãy thiết lập công thức tính Hub và Authority cho các đỉnh của đồ thị, biết công thức tổng quát là $a(i) = \sum_{(j,\ i)\in E} h(j)$ và $h(i) = \sum_{(i,j)\in E} a(j)$
- Tính giá trị của các trang qua 5 vòng lặp, biết rằng giá trị Hub và Authority khởi tạo tại vòng lặp thứ 0 cho các trang đều bằng 1.

Bài tập 10. Cho đồ thị như bên dưới



Hãy thực hiện các yêu cầu như Bài tập 9