

(Sinh viên được sử dụng tài liệu giấy)

HỌ VÀ TÊN SV:	CÁN BỘ COI THI
MSSV:	
STT:	
PHÒNG THI:.....	

CÂU HỎI TỰ LUẬN

Câu 1 (2 điểm) (G1, G2, G3)

- a) Vì sao có nhiều danh sách stopword trong cùng một ngôn ngữ? (G1, 0.5 điểm)
- b) Vì sao không thực hiện stemming khi tiền xử lý văn bản tiếng Việt? (G2, 0.5 điểm)
- c) Nêu mục đích của việc lập chỉ mục? (G3, 1 điểm)

Dữ liệu bên dưới được dùng cho Câu 2 và Câu 3

Cho ma trận Term-Document được xây dựng từ tập tài liệu $D = \{d_1, d_2, d_3, d_4, d_5, d_6\}$. Các tài liệu đã được tiền xử lý với các bước tách token theo khoảng trắng, lọc stopword và stemming với Porter stemmer. Biết các kết quả stemming như sau:

- Kết quả stemming các từ networks, networking, network là network.
- Kết quả stemming các từ computer, computers, compute, computes, computation là comput.
- Kết quả stemming từ neural là neural.
- Kết quả stemming từ predict, predicts, predicted, prediction là predict.

Ma trận Term-Document

	d ₁	d ₂	d ₃	d ₄	d ₅	d ₆
network	10	0	30	1	23	0
comput	2	20	18	32	16	6
neural	10	0	1	0	22	4
predict	1	8	3	6	9	7

Câu 2 (3 điểm) (G1, G2, G3)

- a) Vẽ chỉ mục đảo ngược của mô hình Boolean tương ứng với ma trận Term-Document trên? (G3, 1 điểm)
- b) Trình bày quá trình xử lý câu truy vấn $q = \text{"neural AND networks"}$ theo mô hình Boolean với chỉ mục đã xác định ở câu a) (G2, G3, 1 điểm)

- c) Giả sử trọng số $tf_{td} = \log_2(\text{count}(td))$ và $idf_{td} = N/N_{td}$. Tính trọng tf và idf của term **network** trong tài liệu d_5 theo dữ liệu đã cho ở ma trận Term-Document (G1, 1 điểm)

Câu 3 (4 điểm) (G2, G4)

Cho các ma trận S , Σ , U^T là kết quả của phép tính SVD (Singular Value Decomposition) trên ma trận Term-Document theo mô hình LSI (Latent Semantic Index) như sau:

Ma trận S	Ma trận Σ	Ma trận U^T
0.58	0.65	0.18
0.72	-0.63	0.29
0.28	0.39	0.55
0.23	-0.12	0.44
	56.25 0	0.59 0.12
	30.28	-0.67 0.41 0.1

- a) Tính các vector từ khóa theo LSI? (G4, 1 điểm)
- b) Tính các vector tài liệu theo LSI? (G4, 1 điểm)
- c) Trình bày quá trình xử lý câu truy vấn $q = \text{"neural computation"}$ theo LSI? (G2, G4, 1 điểm).
- d) Trình bày quá trình gom cụm các tài liệu trong D thành hai nhóm theo phương pháp K-Means từ kết quả của mô hình LSI? (G4, 1 điểm)

Câu 4 (1 điểm) (G1)

Cho tập kết quả truy vấn trên toàn bộ 50 tài liệu như Bảng GOLD bên dưới.

Bảng GOLD

Truy vấn	Danh sách tài liệu liên quan
1	1, 29, 32, 40, 45, 48
2	2, 8, 16, 44

Cho tập kết quả xử lý truy vấn của hệ thống A như Bảng SYS bên dưới.

Bảng SYS

Truy vấn	Hệ thống A
1	1, 2, 9, 20, <u>29</u> , 31, <u>40</u> , 44, <u>45</u> , <u>48</u>
2	1, <u>2</u> , 17, 22, 28, 37, 41, <u>44</u> , 49

Hãy tính F_1 của hệ thống A?