BÀI TẬP ÔN TẬP KHAI PHÁ WEB 20221

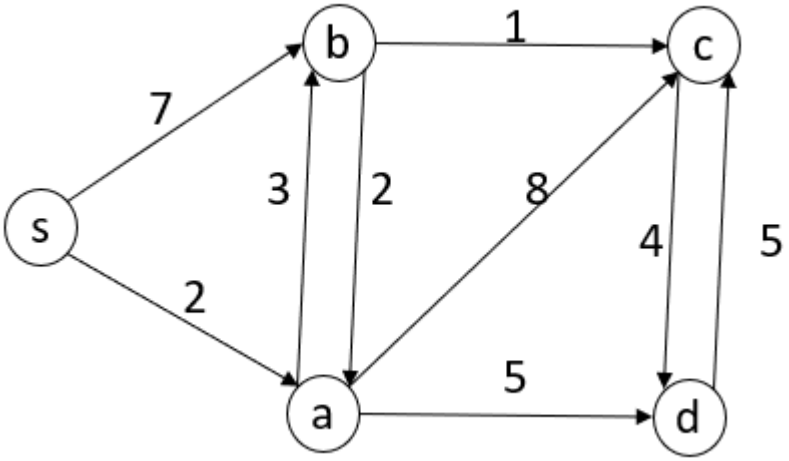
Ôn tập các nội dung lý thuyết dựa trên slide, tài liệu tham khảo, bài báo, các bài trắc nghiệm trong học kỳ.

Ôn tập các bài tập đã làm trong học kỳ.

Làm các bài tập bổ sung sau:

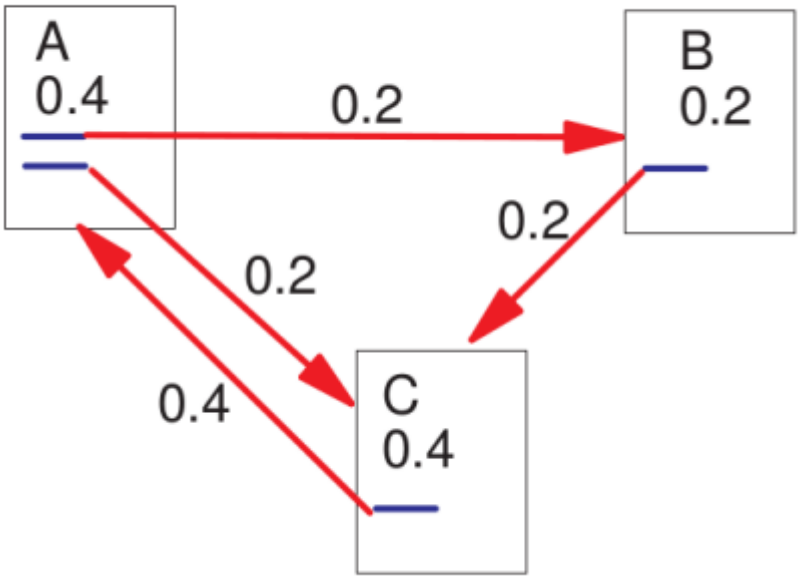
Bài tập 1:

Dijkstra: Khoảng cách ngắn nhất từ a tới c theo thuật toán Dijkstra



Bài tập 2:

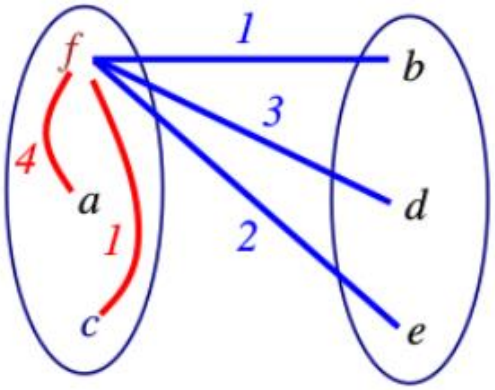
Pagerank: Giải theo hệ phương trình và theo phương pháp lặp (tới vòng lặp 4) với d = 0.8



Bài tập 3:

Kerninghan-Lin: Thực hiện một vòng lặp của thuật toán với đồ thị và khởi tạo như sau:

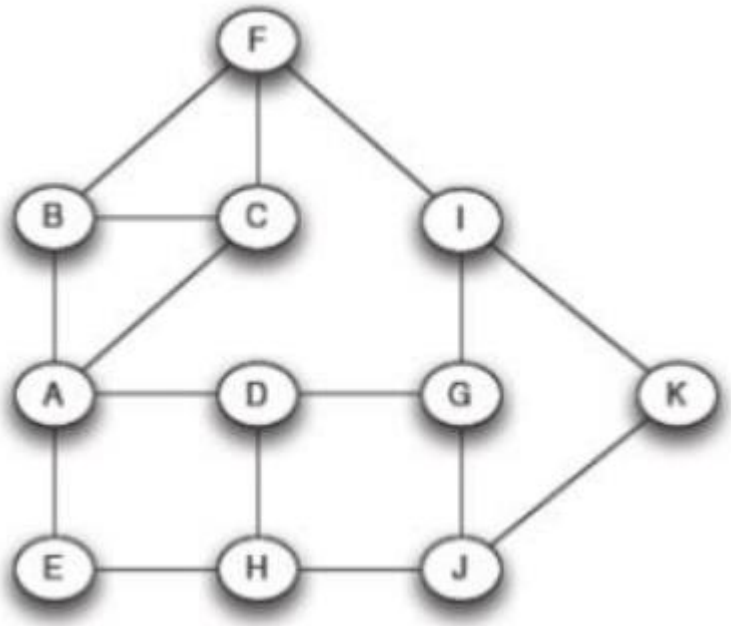
	a	b	c	d	e	f
a	0	1	2	3	2	4
b	1	0	1	4	2	1
c	2	1	0	3	2	1
d	3	4	3	0	4	3
e	2	2	2	4	0	2
f	4	1	1	3	2	0



Initial cut cost = (1+3+2)+(1+3+2)+(1+3+2) = 18 (22−4)

Bài tập 4:

Tính khả năng thông qua của các cạnh dựa trên số đường đi ngắn nhất từ A tới các đỉnh còn lại trong đồ thị



Bài tập 5:

Tìm lớp của văn bản 5 dựa trên Multinomial Naïve Bayes với kĩ thuật làm mịn thêm 1

	docID	words in document	in $c = China$?
training set	1	Chinese Beijing Chinese	yes
	2	Chinese Chinese Shanghai	yes
	3	Chinese Macao	yes
	4	Tokyo Japan Chinese	no
test set	5	Chinese Chinese Chinese Tokyo Japan	?

Bài tập 6:

Tìm lớp của văn bản 5 dựa trên Multinomial NB với kĩ thuật làm mịn thêm 1

	docID	words in document	in $c = China$?
training set	1	Taipei Taiwan	yes
	2	Macao Taiwan Shanghai	yes
	3	Japan Sapporo	no
	4	Sapporo Osaka Taiwan	no
test set	5	Taiwan Taiwan Sapporo	?

Bài tập 7:

Dự đoán đánh giá của người dùng u_3 với sản phẩm i_4 theo phương pháp CF-knn dựa trên người dùng với k = 2. Độ tương đồng pearson, công thức tính dự đoán như trong slide bài giảng.

	i_1	i_2	i_3	i_4
u_1	5	4	4	1
u_2	2	1		
u_3	5	4	4	?
u_4		1	2	5

Bài tập 8:

Một mô hình RNN được thiết kế cho bài toán NER có kiến trúc như sau: Nhúng từ có kích thước N = 300. Nhúng từ dựa trên kí tự có kích thước H_C = 100. Hai biểu diễn nhúng trên được ghép nối (concat) và đưa vào mạng RNN có số nơ-ron ở tầng ẩn là H = 150. Tầng ẩn được ghép nối đầy đủ với tầng đầu ra.

Nhúng từ dựa trên kí tự được thiết kế như sau: Kí tự được qua một tầng nhúng có kích thước C = 100. Nhúng kí tự được đưa vào một mạng RNN có số nơ-ron tầng ẩn là H_C = 100. Giá trị đầu ra của tầng ẩn của kí tự cuối cùng trong từ được sử dụng làm biểu diễn của từ.

Cho biết kích thước từ điển V = 10000. Số kí tự khác nhau là V_C = 150. Tập thực thể có tên gồm {PER, ORG, LOC) và sử dụng cơ chế BIO để gán nhãn. Hãy tính số lượng tham số của mô hình nói trên.

Bài tập 9:

Tính rating của user 3 với item 1 và item 6 theo user-based và item-based KNN-CF sử dụng các công thức tính độ tương đồng và dự đoán như sau (k = 2):

	1	2	3	4	5	6
1	7	6	7	4	5	4
2	6	7	?	4	3	4
3	?	3	3	1	1	?
4	1	2	2	3	3	4
5	1	?	1	2	3	3

User-based:

$$\mu_u = \frac{\sum_{k \in I_u} r_{uk}}{|I_u|} \quad \forall u \in \{1 \dots m\}$$

$$\text{Sim}(u,v) = \text{Pearson}(u,v) = \frac{\sum_{k \in I_u \cap I_v} (r_{uk} - \mu_u) \cdot (r_{vk} - \mu_v)}{\sqrt{\sum_{k \in I_u \cap I_v} (r_{uk} - \mu_u)^2} \cdot \sqrt{\sum_{k \in I_u \cap I_v} (r_{vk} - \mu_v)^2}}$$

$$s_{uj} = r_{uj} - \mu_u \quad \forall u \in \{1 \dots m\}$$

$$\hat{r}_{uj} = \mu_u + \frac{\sum_{v \in P_u(j)} \text{Sim}(u,v) \cdot s_{vj}}{\sum_{v \in P_u(j)} |\text{Sim}(u,v)|} = \mu_u + \frac{\sum_{v \in P_u(j)} \text{Sim}(u,v) \cdot (r_{vj} - \mu_v)}{\sum_{v \in P_u(j)} |\text{Sim}(u,v)|}$$

trong đó I_u là tập sản phẩm mà người dùng u đánh giá

$P_u(J)$ là tập top k người dùng tương đồng với người dùng u mà có đánh giá với sản phẩm j

Item-based

$$\text{AdjustedCosine}(i,j) = \frac{\sum_{u \in U_i \cap U_j} s_{ui} \cdot s_{uj}}{\sqrt{\sum_{u \in U_i \cap U_j} s_{ui}^2} \cdot \sqrt{\sum_{u \in U_i \cap U_j} s_{uj}^2}}$$

$$\hat{r}_{ut} = \frac{\sum_{j \in Q_t(u)} \text{AdjustedCosine}(j,t) \cdot r_{uj}}{\sum_{j \in Q_t(u)} |\text{AdjustedCosine}(j,t)|}$$

trong đó U_i là tập người dung đánh giá sản phẩm i

$Q_t(u)$ là top k sản phẩm tương đồng với sản phẩm t được người dùng u đánh giá