First Sem 20-21

Comprehensive Exam solution

CS F469 Information Retrieval

Q1. A) $q_{23} = \frac{q_{21}}{2} + \frac{q_{22}}{2}$

B)

	1	2	3	4
1	0	1/2	0	0
2	1/2	0	0	0
3	1/2	1/2	0	0
4	0	0	1	0

C) The matrix generated is not suitable for pagerank since the node 4 has a dead end and there is no way a user can exit once landed on this page. Hence the matrix is not column stochastic.

D)

$$0.8 \begin{bmatrix} 0 & 1/2 & 0 & 0 \\ 1/2 & 0 & 0 & 0 \\ 1/2 & 1/2 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} + 0.2 \begin{bmatrix} 1/4 & 1/4 & 1/4 & 1/4 \\ 1/4 & 1/4 & 1/4 & 1/4 \\ 1/4 & 1/4 & 1/4 & 1/4 \\ 1/4 & 1/4 & 1/4 & 1/4 \end{bmatrix}$$

$$\begin{bmatrix} 0.05 & 0.45 & 0.05 & 0.05 \\ 0.45 & 0.05 & 0.05 & 0.05 \\ 0.45 & 0.45 & 0.05 & 0.05 \\ 0.05 & 0.05 & 0.85 & 0.05 \end{bmatrix}$$

E) $R_{t+1} = M R_t$

$$\begin{bmatrix} r_1 \\ r_2 \\ r_3 \\ r_4 \end{bmatrix} = \begin{bmatrix} 0.05 & 0.45 & 0.05 & 0.05 \\ 0.45 & 0.05 & 0.05 & 0.05 \\ 0.45 & 0.45 & 0.05 & 0.05 \\ 0.05 & 0.05 & 0.85 & 0.05 \end{bmatrix} \begin{bmatrix} 1/4 \\ 1/4 \\ 1/4 \\ 1/4 \end{bmatrix}$$

$$\begin{bmatrix} r_1 \\ r_2 \\ r_3 \\ r_4 \end{bmatrix} = \begin{bmatrix} 1/4 \\ 1/4 \\ 1/4 \\ 1/4 \end{bmatrix} \begin{bmatrix} 0.15 \\ 0.15 \\ 025 \\ 0.25 \end{bmatrix}$$

F) Tweet graph Construction – 3 M

Teleportation / dead ends etc – 2M

There is no single solution hence the Evaluation is based on the above two Explanations.

Q2.

A) $Cd(:,i) = A(:,j)/\sqrt{CP(j)}$

$C_3 = \frac{1}{\sqrt{76/171}} [1\ 0\ 5\ 5\ 5]^T$

$= [1.5\ 0\ 7.6\ 7.6\ 7.6]^T$ } 2M

$R_4 = \frac{1}{\sqrt{51/171}} [0\ 1\ 5\ 5]$

$= [0\ 1.85\ 9.3\ 9.3]$ } 2M

B) Intersection of C & R is 5. → 1M

C) If $v = \frac{1}{25}$ rating is computed using $r = C \cup R$

$$\begin{bmatrix} 1.5 \\ 0 \\ 7.6 \\ 7.6 \\ 76 \end{bmatrix} \left[\frac{1}{25}\right] \begin{bmatrix} 0 & 1.85 & 9.3 & 9.3 \end{bmatrix}$$

2M

Rating for Ria for the Movie Newton is 2.8

D.
i) Since the algorithms use different approaches for predictions they are bound to give different results and considering the strength & weakness of Each of the algorithms we can take a linear combination of the all predictions using the following approaches. (4M)

1. Take the avg prediction of all algorithms.
2. learn weights for Each algorithm using a log linear model & then computed a weighted avg of all predictions.
3. Apply an Ensemble methodlogy.

ii) The RMSE of the final rating using any of the above shall be lower than a single algorithm. (2M)

E) Shyam's model with 5 latent factors would be superior since the no of latent factors are higher. (3M)

(Q3)

A) $9 - 4 - 1 = 4$

The reordering distance for the phrase "for hunting" is 4.

B) Since the order of the languages are different if the phrases do not move the sentences generated may not be semantically correct hence movement of the phrases should not be penalised. This can be controlled by a factor α which is generally in a range of 0-1. Since the reordering is modeled as $\alpha^{|d|}$ an α-value close to 1 will not penalise the movement.

C) Phrase based SMT models are better than word based models since it reduces lexical ambiguities. We can also control the movement of the phrases using the parameter α. Since we translate phrases the words are seen together in both languages which is an advantage for shorter sentences since we are learning the whole sentence translation. in a few cases.

Q4. A) Since the odds ratio is a monotonic function which behaves in the same way as computing the probability it is used to derive the ranking function. (5M)

B) Since Step considers that there are M terms in a document the products runs from 1 to M. Each term may be present or absent in the document and if the document does not contain a term this parameter computation can be avoided hence we split it into two factors (5M)

$$O(R|\vec{d}) = \prod_{t:x_t=1} \frac{P(x_t=1|R=1,\vec{q})}{P(x_t=1|R=0,\vec{q})} \cdot \prod_{t:x_t=0} \frac{P(x_t=0|R=1,\vec{q})}{P(x_t=0|R=0,\vec{q})}$$

c)

$$O(R|\vec{x},\vec{q}) = \prod_{\substack{t:x_{td}=1 \\ t:x_{tq}=1}} \frac{P(x_t=1|R=1,\vec{q})}{P(x_t=1|R=0,\vec{q})} * \prod_{\substack{t:x_{td}=0 \\ t:x_{tq}=1}} \frac{P(x_t=1|R=1,\vec{q})}{P(x_t=1|R=0,\vec{q})} *$$

$$\prod_{\substack{t:x_{td}=1 \\ t:x_{tq}=0}} \frac{P(x_t=1|R=1,\vec{q})}{P(x_t=1|R=0,\vec{q})} * \prod_{\substack{t:x_{td}=0 \\ t:x_{tq}=0}} \frac{P(x_t=1|R=1,\vec{q})}{P(x_t=1|R=0,\vec{q})}$$

Basically the idea is to have all combinations of the presents/absence of the words in the document & the query From these four combinations we know that if the word is not present in the query & the doc and also not present in the query & present in the doc they need not be estimated hence we only consider the combinations where by default $t:x_{tq}=1$ and find their presence absence ($t:x_{td}=1$ or $t:x_{td}=0$) in the document.

D) Use the BM25 formula & add the scores for "health" and "insurance" terms:

$$RSV(D_1) = \log_{10}\left(\frac{2000}{20}\right)\frac{(0.8+1)\times 6}{0.8\left((1-0.7)+0.7\times\frac{52}{156}\right)+6} + \log_{10}\left(\frac{2000}{200}\right)\frac{(0.8+1)\times 6}{0.8\left((1-0.7)+0.7\times\frac{52}{156}\right)+6}$$

$$RSV_1(D_1) = 5.04$$

$$RSV(D_2) = \log_{10}\left(\frac{2000}{20}\right)\frac{(0.8+1)\times 8}{0.8\left((1-0.7)+0.7\times\frac{78}{156}\right)+8} + \log_{10}\left(\frac{2000}{200}\right)\frac{(0.8+1)\times 4}{0.8\left((1-0.7)+0.7\times\frac{78}{156}\right)+4}$$

$$RSV_1(D_2) = 4.97$$

Since $RSV(D_1) > RSV(D_2)$, D_1 is considered more relevant to the query.

ii) When b=0 the document length is not taken into account and as b increases the relative document length matters more. The higher the document length, the term frequency is discounted.

iii) As k1 goes higher, the effect of tf is cancelled and hence RSV approximates to scaled version of IDF.