Date...../...../.....

Page.....

Ans2 @ Binary independence model is a language model (LM) and shares following similarity with vector space model

(i) uses term frequencies

(ii) probabilities are normalised as to length, similar to cosine normalisation in vector space model.

Ans2 (b) We use smoothening as even if a term doesn't occur in the particular document, it can be a possibility to consider.

Date...../...../.....

Page.....

Ans 3

(a) (i) Content selection to choose sentences to extract from document.

(ii) Information ordering to choose order to place them in the summary.

(iii) Sentence realization to simplify sentences

(iv) Removing redundancy.

summarized text.

Ans3

(b) Luhn uses weights for words that are salient & informative. In this way it finds weights of sentences & choose top ones.

LexRank is a graph based approach that uses pagerank algorithm to find similarity b/w sentences & score them

Ans 3

(c) Rouge-1 =

Ref: The quick brown fox jumped over the lazy dog

Algo summary: The quick brown dog jumped over the lazy fox ...

$$\text{Rouge1} = \frac{9}{9} = 1$$

$$\text{Rouge2} = \frac{5}{8} = 0.625$$

Rouge1 doesn't reflect the correctness of algo summary as it is clear that the algo summary means very different from above ref summary. This happens bcz interchanging two words doesn't affect Rouge1 as the relative order is lost.

Rouge2 does better than Rouge1 as it still requires correct order for a better score. However, it still isn't perfect.

Ans (a) Likelihood ratio:

$P(\text{string}|M_{d_1}) = 0.2 * 0.005 * 0.02 * 0.04 * 0.001$

$= 2 \times 5 \times 2 \times 4 \times 10^{-1-3-2-2-3}$

$= 80 \times 10^{-11}$

$= 8 \times 10^{-10}$

$P(\text{string}|M_{d_2}) = 0.13 * 0.01 * 0.04 * 0.04 * 0.002$

$= 480 \times 10^{-11}$

$= 48 \times 10^{-10}$

$\therefore$ Likelihood ratio $= \frac{8}{48} = \frac{1}{6}$.

(b) yes. Now $P(M_1) = 8 \times 10^{-10} \times 0.1$

$P(M_2) = 48 \times 10^{-10} \times 0.1$

$\therefore$ Likelihood ratio $= \frac{1}{6}$.

(c) $P(q|d_1) = \left(\frac{1}{4} \times \frac{1}{4} + \frac{3}{4} \times \frac{2}{14}\right) \times \left(\frac{1}{4} \times 0 + \frac{3}{56}\right)$

$= 0.1349 \times 0.0535 = 7.2 \times 10^{-3}$

$P(q|d_2) = \left(\frac{1}{4} \times \frac{1}{5} + \frac{3}{4} \times \frac{2}{14}\right) \times \left(\frac{1}{4} \times \frac{1}{5} + \frac{3}{4} \times \frac{1}{14}\right)$

$= \frac{11}{70} \times \frac{29}{280} = 0.016$

$\therefore$ Doc 2 has more probability