

V2V2V2V2V2 V2V2V2V2V2 V2V2V2V2V2 V2V2V2V2V2V2V2

National University of Computer & Emerging Sciences

FAST-Karachi Campus

CS4051- Information Retrieval

Quiz#3

Dated: April 22, 2024

Marks: 20

Time: 20 min.

Std-ID: _____ Sol _____

Question No. 1

- a. What are the assumptions in building language model for IR? List them.

Assumptions for Language Model

- Both documents and query are objects of the same type.
- Documents are relevant to the query, if the same generative process is used to generate query that generated the document.
- Probability that the query is generated from the same document is used as a relevance to the query. (It is a generative model for each document).
-

- b. Consider making a language model from the following training text:

Humpty Dumpty sat on a wall. Humpty Dumpty had a great fall

- (i) Under a MLE-estimated unigram probability model, what are $P(\text{humpty})$ and $P(\text{sat})$?

$$P(\text{humpty}) = 2/12 = 1/6$$

$$P(\text{sat}) = 1/12$$

- (ii) Under a MLE-estimated bigram model, what are $P(\text{dumpty}|\text{humpty})$ and $P(\text{humpty}/\text{dumpty})$?

$$P(\text{dumpty}|\text{humpty}) = P(\text{dumpty}, \text{humpty}) / P(\text{humpty}) = 2/2 = 1$$

$$P(\text{humpty}/\text{dumpty}) = P(\text{humpty}, \text{dumpty}) / P(\text{dumpty}) = 0/2 = 0$$

V2V2V2V2V2 V2V2V2V2V2 V2V2V2V2V2 V2V2V2V2V2V2V2

Question No. 2

Suppose we have a collection that consists of the 3 documents given in the below table.

DocID	Doc-content
1	w3 w4 w2 w5
2	w3 w1 w3 w2
3	w2 w3 w4 w3

Build a query likelihood language model for this document collection. Assume a mixture model between the documents and the collection, with $\lambda = 0.6$ for document. Give the document ranking against the query = "w3 w4". You can use Laplace Smoothing. [10]

Language Model

Doc-Model	w1	w2	w3	w4	w5
D1	0	1/4	1/4	1/4	1/4
D2	1/4	1/4	2/4	0	0
D3	0	1/4	2/4	1/4	0
Collection Model	1/12	1/4	5/12	1/6	1/12

Model Probabilities for query = "w3 w4"

$$P(MD1/q) = [0.6 * 1/4 + 0.4 * 5/12] + [0.6 * 1/4 + 0.4 * 1/6] = 0.15 + 0.16 + 0.15 + 0.06 = 0.526$$

$$P(MD2/q) = [0.6 * 2/4 + 0.4 * 5/12] + [0.4 * 1/4 + 0.4 * 0] = 0.25 + 0.1 = 0.35$$

$$P(MD3/q) = [0.6 * 2/4 + 0.4 * 5/12] + [0.4 * 1/4 + 0.4 * 1/6] = 0.46 + 0.46 = 0.91$$

Ranking will be D3, D1 and D2.