Information Retrieval (TDT4117)

Assignment 2

Submitted by

Name : Md Anwarul Hasan
Student ID : 583233

Task 1 – Language Model

1) Answer:

Language models are probabilistic and statistical distribution on words or sequence of words to determine whether the certain sequence of words is valid or not. Using this concept of language model researchers try to determine that after a certain sequence of words a certain is will be next. Language models is very useful in natural language processing, information retrieval, automated language generation etc.

The strength of language models are that they provide strong base of learning and training dataset and also it can predict more efficiently what user may likely are searching or more precisely which document it is searching. Simplicity of these models are another advantage.

On the other hand the weakness of a language model is that is do not understands the syntax of a language. Rather it interacts by using semantics of the language that present in the documents. Moreover, language models are also called “bag of words” models as the order of words are irrelevant to this models. The other basic problem of such model is it not precisely defined what model should be used as it is very dependant on smoothing parameter.

2) Answer:

Provided that,

d1 = failure is the opportunity to begin again more intelligently

d2 = intelligence is the ability to adapt to change

d3 = lack of will power leads to more failure than lack of intelligence or ability

and,

q1 = failure

q2 = intelligence opportunity

q3 = intelligence failure

Using the Jelinek-Mercer smoothing

$$\hat{P}(t|M_d) = (1-\lambda)\hat{p}_{mle}(t|M_d) + \lambda\hat{p}_{mle}(t|_C), \lambda = 0.5.$$

From the provided, documents and given query terms we see below information,

Documents	Length of the document	Frequency of the term "failure" in the document	Frequency of the term "intelligence" in the document	Frequency of the term "opportunity" in the document
d1	9	1	0	1
d2	8	0	1	0
d3	14	1	1	0
Total	31	2	2	1

So for query q1,

$P(q1|d1) = (1 - 0.5) * 1/9 + 0.5 * 2/31 \approx 0.0878$

$P(q1|d2) = (1 – 0.5) * 0/8 + 0.5 * 2/31 \approx 0.0323$

$P(q1|d3) = (1 – 0.5) * 1/14 + 0.5 * 2/31 \approx 0.068$

So the ranking of the documents for q1 is d1 > d3 > d2

So for query q2,

$P(q2|d1) = [(1 - 0.5) * 0/9 + 0.5 * 2/31] * [(1 - 0.5) * 1/9 + 0.5 * 1/31] \approx 0.0023$

$P(q2|d2) = [(1 – 0.5) * 1/8 + 0.5 * 2/31] * [(1 – 0.5) * 0/8 + 0.5 * 1/31] \approx 0.0015$

$P(q2|d3) = [(1 – 0.5) * 1/14 + 0.5 * 2/31] * [(1 – 0.5) * 0/14 + 0.5 * 1/31] \approx 0.0011$

So the ranking of the documents for q2 is d1 > d2 > d3

So for query q3,

$P(q3|d1) = [(1 - 0.5) * 0/9 + 0.5 * 2/31] * [(1 - 0.5) * 1/9 + 0.5 * 2/31] \approx 0.0028$

$P(q3|d2) = [(1 – 0.5) * 1/8 + 0.5 * 2/31] * [(1 – 0.5) * 0/8 + 0.5 * 2/31] \approx 0.0031$

$P(q3|d3) = [(1 – 0.5) * 1/14 + 0.5 * 2/31] * [(1 – 0.5) * 1/14 + 0.5 * 2/31] \approx 0.0046$

So the ranking of the documents for q3 is d3 > d2 > d1

3) Answer:

When a term is not present in the document there then the probability of that term is zero according to the language model. If the smoothing is not present then the whole query may get zero probability. Smoothing do not let the whole query become zero rather it provides a

system so that even though the term is not present in the document the probability generates.

In the previous question, the term “failure” was not present in d2. Which made this part of the equation, $(1-\lambda)\hat{p}_{mle}(t|M_d)$, to zero. But the smoothing part , $\lambda\hat{p}_{mle}(t|_C)$, did not let the total probability be zero.

Task 2

1) Answer:

MAP: The elaboration of MAP is Mean Average Precision. After calculating the precision of all the documents for a query and providing zero for no retrieved document for a query, then summing up all the results from all the query like previously and then dividing the result by total no of queries is the method of calculating MAP.

The pros: For an IR system, MAP makes the rank of all the relevant items towards high so that the top results displays more desired or relevant results in connection with information retrieval. This concept is particularly useful for search engines in the web as it will show first few documents with high precision and user will have better experience.

The cons: As it is a mean, a specific bad query may make the MAP bring down significantly irrespective of other good query performance. The mean average system not always shows the real scenario of a information retrieval system and some bad query could bring the value of the mean drastically.

MRR: The elaboration of MRR is Mean Reciprocal Rank. It is reciprocal ranking of a document at which it first retrieved. Such as, if the document is retrieved at 1st, then the reciprocal ranking is 1 and 0.5 if it is retrieved at 2. MRR is the mean of all the retrieved documents.

The pros: For system answering system the MRR will be very suitable. In the web search, when searches for URL or homepage the method will provide most relevant response.

The cons: The method may not display several relevant documents which limits the scope of the user which searching for multiple relevant documents.

2) Answer:

Provided relevant documents rel = {23, 10, 33, 500, 70, 59, 82, 47, 72, 9}

And provided retrieved documents ret = {5, 500, 2, 23, 72, 79, 82, 215}

So the calculated Recall and Precision is,

Serial	Recall	Precision
1	0	0
2	0.1	0.5
3	0.1	0.33
4	0.2	0.5
5	0.3	0.6
6	0.3	0.5
7	0.4	0.57
8	0.4	0.5

Task 3

2) Answer:

For the given example in Task 2.2, the interpolated precision is provided below,

Serial	Recall	Interpolated Precision
1	0	0.6
2	0	0.6
3	0.1	0.6
4	0.1	0.6
5	0.2	0.6
6	0.3	0.6
7	0.3	0.57
8	0.4	0.57
9	0.4	0.5