Homework2

106703014 Chen Chia Chun

April 2020

1 Probabilistic Information Retrieval

1.[5/20 points] What are the differences between standard VSM with tf-idf weightings and the RSJ probabilistic retrieval model (in the case where no relevance information is provided)?

Vector Space Model use a set of vectors to represent the appearance of each word, using F-IDF weight to caculate. Then use the query vector to compare with the ducuments veectors. In order to find the similarity of words and documents. In contrast, RSJ is focused on the appearance of the specific words. RSJ need to judge whether the document is relevant for a query keyword, and TF-IDF does not need.

2.[5/20 points] Describe the differences of relevance feedback used in VSM and probabilistic retrieval models.

Vector Space Model calculating ranking method is to calculate the similarity between the query and the document. It's focused on the frequency of the same word appeared in different documents. But the probabilistic retrieval models additionally calculates relation and meanings between the words.

3.[10/20 points]Please show that based on the model above, documents can be ranked via the following formula:

$$\sum_{t:x_t=q_t=1} log\frac{P_t(1-u_t)}{u_t(1-P_t)} = \sum_{t:x_t=q_t=1} log\frac{P(A_t=1|Q,R=1)(1-P(A_t=1|Q,R=0))}{P(A_t=1|Q,R=0)(1-P(A_t=1|Q,R=1))}$$

$$= \sum_{t:x_t=q_t=1} log\frac{P(A_t=1|Q,R=1)(1-P(A_t=0|Q,R=0))}{P(A_t=1|Q,R=0)(1-P(A_t=0|Q,R=1))}$$

$$= log\frac{(\frac{s}{S})(\frac{(N-df_t)-(S-s)}{N-s})}{\frac{df_t-s}{N-S}\frac{S-s}{S}} = log\frac{s((N-df_t-(S-s)))}{(df_t-s)(S-s)}$$

2 Language Model

[5/10 points] Under a MLE-estimated unigram probability model, what are P(the) and P(martian)?

$$P(the) = \frac{2}{11}, P(martian) = \frac{1}{11}$$

[5/10 points] Under a MLE-estimated bigram model, what are P(sensation—pop) and P(pop—the)?

$$P(sensation|pop) = \frac{1}{1}, P(pop|the) = \frac{0}{2} = 0$$

3 Language Model

[10/30 points] (i) click

$$P(q|D_1) = \frac{1}{2} * \frac{1}{2} + \frac{1}{2} * \frac{7}{16} = \frac{15}{32}$$

$$(q|D_2) = \frac{1}{2} * 1 + \frac{1}{2} * \frac{7}{16} = \frac{23}{32}$$

$$P(q|D_3) = \frac{1}{2} * 0 + \frac{1}{2} * \frac{7}{16} = \frac{7}{32}$$

$$P(q|D_4) = \frac{1}{2} * \frac{1}{4} + \frac{1}{2} * \frac{7}{16} = \frac{11}{32}$$

$$D_2 > D_1 > D_4 > D_3$$

[10/30 points] (ii) shears

$$P(q|D_1) = \frac{1}{2} * \frac{1}{8} + \frac{1}{2} * \frac{2}{16} = \frac{2}{16}$$

$$P(q|D_2) = \frac{1}{2} * 0 + \frac{1}{2} * \frac{2}{16} = \frac{1}{16}$$

$$P(q|D_3) = \frac{1}{2} * 0 + \frac{1}{2} * \frac{2}{16} = \frac{1}{16}$$

$$P(q|D_4) = \frac{1}{2} * \frac{1}{4} + \frac{1}{2} * \frac{2}{16} = \frac{3}{16}$$
$$D_4 > D_1 > D_2 = D_3$$

[10/30 points] (iii) click shears

$$P(q|D_1) = \frac{15}{32} * \frac{2}{16}$$

$$P(q|D_2) = \frac{23}{32} * \frac{1}{16}$$
$$P(q|D_3) = \frac{7}{32} * \frac{1}{16}$$

$$P(q|D_4) = \frac{11}{32} * \frac{3}{16}$$
$$D_4 > D_1 > D_2 > D_3$$

4 Mixture model

$$D_1 = (England)(Lordon) = (\frac{1}{2} * \frac{1}{8} + \frac{1}{2} * \frac{3}{23}) * (\frac{1}{2} * \frac{2}{8} + \frac{1}{2} * \frac{5}{23})$$
$$D_2 = (England)(Lordon) = (\frac{1}{2} * \frac{1}{8} + \frac{1}{2} * \frac{3}{23}) * (\frac{1}{2} * \frac{2}{8} + \frac{1}{2} * \frac{5}{23})$$
$$D_3 = (England)(Lordon) = (\frac{1}{2} * \frac{1}{7} + \frac{1}{2} * \frac{3}{23}) * (\frac{1}{2} * \frac{1}{7} + \frac{1}{2} * \frac{5}{23})$$
$$D_1 = D_2 < D_3$$

5 Classic Probabilistic Retrieval Model

[15/35 points] The retrieval function above won't work unless we can estimate all the parameters. Suppose we use the entire collection C=D1,...,Dn as an approximation of the examples of non-relevant documents. Propose an estimate of p(w—Q,R=0). (Hint: study the slide about how to do this for the RSJ model.)

The total number of parameters is $2|V|$,because there are $p(w|Q, R = 1)$ and $p(w|Q, R = 0)$for each word w.
$p(w|Q, R = 1)$ sum has to be equal to 1 ,so the parameters in $p(w|Q, R = 1)$ will

be $|V|-1$, and $p(w|Q,R=0)$ is the same ,so we have $2|V|-2$ free parameters in total.

[5/35 points] Suppose we use the query as the only example of a relevant document. Propose an estimate of p(w—Q,R=1).

Suppose we use the query Q as the only example of a relevant document.

$$p(w|Q,R=1)=\frac{c(w,Q)}{|Q|}$$

[5/35 points] With the two estimates you proposed, we should now have a retrieval function that can be used to compute a score for any document D and any query Q. Write down your retrieval function by plugging in the two estimates. Can your retrieval function capture the three major retrieval heuristics (i.e., TF, IDF, and document length normalization)? How?

$$\frac{P(R=1|Q,D)}{P(R=0|Q,D)}\propto\frac{P(D,Q|R=1)}{P(Q,D|R=0)}\frac{P(D|Q,R=1)P(Q|R=1)}{P(D|Q,R=0)P(Q|R=0)}\propto\frac{P(D|Q,R=1)}{P(D|Q,R=0)}$$

[5/35 points] Do you believe your formula would work well as compared with a state of the art formula such as BM25? Can you propose a way to further improve your formula? (While it's the best if you could improve your formula through using an improved estimate of p(w—Q,R=1), it would also be fine to propose any reasonable heuristic modification of the formula.)

No, I think that BM25 work better than my formula. Because BM25 will condiser the length of the documents, if the documents are too long, it will punish them. Thus, these documents won't be chosen. However, if the document length is short or fixed, I think my formula can work as well as BM25.

As the above said, we get the estimation like these:

$$p(D|Q,R=1)=\prod_{w\in V}p(w|Q,R=1)^{c(w,D)}$$

$$p(D|Q,R=0)=\prod_{w\in V}p(w|Q,R=0)^{c(w,D)}$$

and return them to the original derivation formula, so my formula will become

$$\frac{\prod_{w\in V}p(w|Q,R=1)^{c(w,D)}}{\prod_{w\in V}p(w|Q,R=0)^{c(w,D)}}$$