Assignment #4 Language Models for Retrieval

yuweic3

2017/9/26

#1 Classic Probabilistic Retrieval Model

1. When using multinomial language model, the query likelihood would be

then we have

With a such model, we need to estimate all

1. With the MLE and collection , we have
2. We interpolate the MLE with the collection language model with a fixed coefficient :
3. The scoring formula can be:

**#2 Language Models**

1. For the query likelihood scoring method, we have

then rewrite the ranking function with Jelinek-Mercer smoothing method with a fixed ,

we find that the parameter in ranking function will be the same for all documents, which makes the length normalization term be a constant.

We need not to care about and because they are all the same in all documents. The only one we care is .

Therefore,

1. 1) Vector for document :

2) Treat the query as a “short document” in the same vector space of document, so the vector for query is:

3) Similarity of two vectors defined by cosine similarity: . In another words, this similarity function can be used as a measure of the score of the document for that query.

4) Yes.

Although we do not see any specific weighting of terms in the document vector, in practice there are a number of weighting schemes are possible for query and document vectors. For example, we can simply use as the weight of a term in the query (and zero for a term not in the query). For documents, weighting can be used with no use of with Euclidean normalization to turn all document vectors into unit vectors and eliminate all information on the length of the original documents

In fact, in the second part of the retrieval function we get above:

the role of in the smoothing distribution is similar to the IDF, capturing the TF-IDF weighting.

On the other hand, plays the role of document length normalization.

1. For Jelinek-Mercer smoothing with ,

will not over penalize a long document

For Dirichelet prior smoothing with , we notice that is document-dependent. It is smaller for long document, so it can be treated as a penalty for long document

So , will over penalize the long document.

**#3 KL-divergence Retrieval Function**

When we set the query language model to the empirical word distribution in the query , the estimate will be

Then, the model covers the query likelihood scoring function. In fact, when we use empirical word distribution in estimation, it would zero all words except for the query words. It will make easier to score the document with .