COMP 5790/ 6790/ 6796

Special Topics: Information Retrieval

Instructor: Shubhra ("Santu") Karmaker

Assignment 3: Probabilistic Retrieval Models + Smoothing [100 points]

⚠ **Notice:** This assignment is due **Friday, February 14, 2020 at 11:59pm**.

Please submit your solutions via Canvas (https://auburn.instructure.com/). You should submit your assignment as a **typeset PDF**. Please do not include scanned or photographed equations as they are difficult for us to grade.

1. Smoothing [15 pts]

a. **[10 pts]** Write down the formula for Dirichlet Prior Smoothing. Then, mathematically prove the following two lemmas:
   - Show, in the limit where document length tends to infinity, that a unigram language model smoothed with a Dirichlet prior becomes equivalent to one estimated using the maximum likelihood estimate.
   - Show, in the limit where the parameter μ tends to infinity, that a unigram language model smoothed with a Dirichlet prior becomes equivalent to the background language model used in the smoothing.

b. **[5 pts]** Point out one advantage of Jelinek-Mercer smoothing over Katz-Backoff smoothing. Explain why.

2. Application of Smoothing [40 pts]

Again, Consider the document d: **"the sun rises in the east and sets in the west"**. This time, assume that we have a background word distribution (pre-computed somehow) denoted by REF which is characterized as follows:

- $P_{REF}(a)$=0.18
- $P_{REF}(the)$=0.17
- $P_{REF}(from)$=0.13
- $P_{REF}(retrieval)$=0.02
- $P_{REF}(sun)$=0.05
- $P_{REF}(rises)$=0.04
- $P_{REF}(in)$=0.16
- $P_{REF}(BM25)$=0.01
- $P_{REF}(east)$=0.02
- $P_{REF}(sets)$=0.04
- $P_{REF}(west)$=0.02
- $P_{REF}(and)$=0.16

a. **[10 pts]** Assume document d is generated by a Unigram Language Model. Estimate the parameters of the Unigram Language Model using Dirichlet Prior Smoothing assuming μ=4. Now, compare this result against the results obtained from 2(b).
b. **[10 pts]** Repeat problem 5(a) assuming μ=0.01 and μ=100. Compare these results with results from problem 5(a). Do the results match with your intuition? explain why.
c. **[20 pts]** Repeat problem 5(a) with Jelinek-Mercer smoothing instead of Dirichlet Prior Smoothing assuming λ={0.01,0.5,0.9} and compare the results obtained for different λ's. Also, compare these results with results from problem 5(a) and 5(b). What similarities or differences do you observe?

3. Classic Probabilistic Retrieval Model [45 points]

a. **[20 pts]** In the derivation of the Robertson-Sparck-Jones (RSJ) model, a multi-variate Bernoulli model was used to model term presence/absence in a relevant document and a non-relevant document. Suppose, we change the model to a multinomial model (see the slide that covers both models for computing query likelihood). Using a similar independence assumption as we used in deriving RSJ, show that ranking based on probability that a document is relevant to a query Q, i.e., $p(R=1|D, Q)$, is equivalent to ranking based on the following formula:

$$score(Q,D) = \sum_{w \in V} c(w,D) \log \frac{p(w \mid Q, R=1)}{p(w \mid Q, R=0)}$$

where the sum is taken over all the words in our vocabulary (denoted by V), and $c(w,D)$ is the count of word w in document D (i.e., how many times w occurs in D). How many parameters are there in such a retrieval model that we have to estimate?

(Hint: ranking based on probability of relevance is equivalent to ranking based on the odds ratio of relevance, just like we did in RSJ.)

b. **[5 pts]** The retrieval function above won't work unless we can estimate all the parameters. Suppose we use the entire collection $C=D_1,\ldots,D_n$ as an approximation of the examples of non-relevant documents. Give the formula for the Maximum Likelihood estimate of $p(w|Q,R=0)$.
c. **[5 pts]** Suppose we use the query as the only example of a relevant document. Give the formula for the Maximum Likelihood estimate of $p(w|Q,R=1)$ based on this single example.
d. **[5 pts]** One problem with the maximum likelihood estimate of $p(w|Q,R=1)$ is that many words would have zero probability, which limits its accuracy of modeling words in relevant documents. Give the formula for smoothing this maximum likelihood estimate using fixed coefficient linear interpolation (i.e., Jelinek-Mercer) with a collection language model.
e. **[10 pts]** With the two estimates you proposed, i.e., the estimate of $p(w|Q,R=0)$ based on the collection and the estimate of $p(w|Q,R=1)$ based on the query with smoothing, you 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?