605.744: Information Retrieval
Problem Set (Module 6)

Sabbir Ahmed

October 6, 2022

1. (30%) For this problem we will use Cover Density Ranking for the following document which is two verses from an 18th century English nursery rhyme. The numeric superscripts indicate the word order in the document. For this problem we have this one document and a query consisting of the two words **"sing pie"**.

   sing[1] a[2] song[3] of[4] sixpence[5]
   a[6] pocket[7] full[8] of[9] rye[10]
   four[11] and[12] twenty[13] blackbirds[14]
   baked[15] in[16] a[17] pie[18]

   when[19] the[20] pie[21] was[22] opened[23]
   the[24] birds[25] began[26] to[27] sing[28]
   wasn't[29] that[30] a[31] dainty[32] dish[33]
   to[34] set[35] before[36] the[37] king[38]

   Cover Density Ranking is not discussed in the course text. However, there is an example in the lecture slides, and you can find the paper *"Relevance ranking for one to three term queries"* by Clarke et al., in the EReserves in Blackboard. That paper formally defines a cover and gives several examples that you may find useful.

   (a) Is (1, 21) a cover for this query? Explain why or why not?
   Answer: No, because (1, 18) is already a cover for the query that contains the shortest interval between the terms.

   (b) List the covers for this query.
   Answer: {(1, 18), (21, 28)}

   (c) Using a window size of K=8, calculate the similarity score for the document.
   Answer: The similarity score using the cover density ranking is given by the following:

$$S(\ell) = \sum_{j=1}^{n} I(p_j, q_j),\ I(p,q) = \begin{cases} \frac{K}{q-p+1} & \text{if } q-p+1 > K, \\ 1 & \text{otherwise} \end{cases}$$

Therefore, with $K = 8$,

$$\begin{aligned} S(\ell) &= \sum_{j=1}^{n} I(p_j, q_j) = I(1,18) + I(21,28) \\ I(1,18) &= \frac{8}{18-1+1} \text{ (since } 18-1+1 = 18 > 8) \\ &= \frac{4}{9} \\ I(21,28) &= 1 \text{ (since } 28-21+1 = 8 \not> 8) \\ \implies S(\ell) &= \frac{4}{9} + 1 = 1.44 \end{aligned}$$

2. (20%) In the statistical language model presented in the lecture and in Chapter 12 of the text we use linear interpolation (also called a "mixture model" or "Jelinek-Mercer smoothing") to make a probability estimate of a term. This estimate is based both on the term frequency in a document, and on the collection frequency of the term. See Equation 12.12 in IIR.

$$P(d|q) \propto P(d) \prod_{t \in q} ((1-\lambda)P(t|M_c) + \lambda P(t|M_d))$$

   (a) What is the purpose of the parameter λ?

   Answer: It is a non-constant smoothing parameter between (0, 1) where the smaller its value means more smoothing. Its values can be a function of the query size since a small amount of smoothing is suitable for short queries while longer queries perform better with more smoothing.

   (b) What would be the effect of setting λ to a value of 1?

   Answer: Setting the parameter to 1 would yield,

$$P(d) \prod_{t \in q} P(t|M_d)$$

   where the language model built from the entire document collection, M_c, gets discarded. The probability gets reduced to the model created using the single document and the frequency of the term in that document. This introduces a problem, where if a single term of the query is not present in the document, i.e. $P(t_i|M_d) = 0$, then the probability gets reduced to zero as well.

3. (50%) Compute similarity scores for and rank documents D1 and D2 using query Q with a unigram statistical language model. Query Q contains the four words "aardvark bird dog dog". The corpus consists of only these eight documents and only these six indexing terms are found in the collection. The cells in the table below indicate the number of times a word appears in a document. You should use a mixture model with parameter $\lambda = 0.40$. Assume that the prior probability of relevance is equal for all documents.

   Plainly show your work. It is fine to check your work with a program or spreadsheet, but I expect you to show how you derive probability estimates and to see the equations that you use to calculate document similarity.

   Report scores using scientific notation with three digits after the decimal point (e.g., 1.234 $\times 10^{-8}$).

	D1	**D2**	D3	D4	D5	D6	D7	D8
aardvark	5	2		1				2
bird	1	1	1	2	1	1	5	8
cat				2		3		
dog		1	3		2		2	
egret	1			1				
fish	3	2						

Answer: Computing the $tf_{t,d}$ (the (raw) term frequency of term t in document d) and L_d (the number of tokens in document d): Using $\lambda = 0.4$ and Q=*aardvark bird dog dog*, and

	tf_{D1}	L_{D1}	tf_{D2}	L_{D2}	cf_t	cs
aardvark	5	10	2	6	10	50
bird	1	10	1	6	20	50
dog	0	10	1	6	8	50

expanding the probability:

$$\begin{aligned} P(d|q) &\propto P(d)\prod_{t\in q}((1-\lambda)P(t|M_c)+\lambda P(t|M_d)) \\ &\propto \prod_{t\in q}\left((1-\lambda)\frac{cf_t}{cs}+\lambda\frac{tf_{t,d}}{L_d}\right) \end{aligned}$$

Computing the probability for each documents:

$$\begin{aligned} P(Q|D1) &= \prod\left((1-\lambda)\frac{cf_t}{cs}+\lambda\frac{tf_{t,D1}}{L_{D1}}\right) \\ &= \prod\left(0.6\times\frac{cf_t}{50}+0.4\times\frac{tf_{t,D1}}{10}\right) \\ &= \left(0.6\times\frac{cf_{aardvark}}{50}+0.4\times\frac{tf_{aardvark,D1}}{10}\right)\times\left(0.6\times\frac{cf_{bird}}{50}+0.4\times\frac{tf_{bird,D1}}{10}\right) \\ &\quad\times\left(0.6\times\frac{cf_{dog}}{50}+0.4\times\frac{tf_{dog,D1}}{10}\right)\times\left(0.6\times\frac{cf_{dog}}{50}+0.4\times\frac{tf_{dog,D1}}{10}\right) \\ &= \left(0.6\times\frac{10}{50}+0.4\times\frac{5}{10}\right)\times\left(0.6\times\frac{20}{50}+0.4\times\frac{1}{10}\right)\times\left(0.6\times\frac{8}{50}+0.4\times\frac{0}{10}\right)^2 \\ &= (0.12+0.20)(0.24+0.04)(0.096+0)^2 \\ &= (0.32)(0.28)(0.096)^2 \\ &= 8.258\times10^{-4} \end{aligned}$$

$$
\begin{aligned}
P(Q|D2) &= \prod \left((1-\lambda)\frac{cf_t}{cs} + \lambda \frac{tf_{t,D2}}{L_{D2}} \right) \\
&= \prod \left(0.6 \times \frac{cf_t}{50} + 0.4 \times \frac{tf_{t,D2}}{6} \right) \\
&= \left(0.6 \times \frac{cf_{aardvark}}{50} + 0.4 \times \frac{tf_{aardvark,D2}}{6} \right) \times \left(0.6 \times \frac{cf_{bird}}{50} + 0.4 \times \frac{tf_{bird,D2}}{6} \right) \\
&\quad \times \left(0.6 \times \frac{cf_{dog}}{50} + 0.4 \times \frac{tf_{dog,D2}}{6} \right) \times \left(0.6 \times \frac{cf_{dog}}{50} + 0.4 \times \frac{tf_{dog,D2}}{6} \right) \\
&= \left(0.6 \times \frac{10}{50} + 0.4 \times \frac{2}{6} \right) \times \left(0.6 \times \frac{20}{50} + 0.4 \times \frac{1}{6} \right) \times \left(0.6 \times \frac{8}{50} + 0.4 \times \frac{1}{6} \right)^2 \\
&= (0.12 + 0.133)(0.24 + 0.067)(0.096 + 0.067)^2 \\
&= (0.253)(0.307)(0.163)^2 \\
&= 2.056 \times 10^{-3}
\end{aligned}
$$

Therefore, the scores are: $D1 = 2.056 \times 10^{-3}$, $D2 = 8.258 \times 10^{-4}$ with the ranking D2 > D1.