Sheetal Parikh
Problem Set – Module 6
605.744.81 Information Retrieval
Fall 2021

1.) *For this problem we will use Cover Density Ranking with the following document. The document is a portion of the lyrics in Tom Lehrer's song "A Liter and a Gram" which was based on the song "A Bushel and a Peck" in the musical Guys and Dolls. The numeric superscripts indicate word order in the document. For this problem we have only this one document and a query consisting of the two words "**newton crazy**".*

i1 love2 you3 a4 liter5 and6 a7 gram8
a9 liter10 and11 a12 gram13, and14 it's15 crazy16 that17 i18 am19
a20 meter21 and22 and23 a24 yard25 and26 a27 newton28 and29 a30 watt31
a32 newton33 and34 a35 watt36, and37 i38 wanna39 know40 a41 lot42
about43 you44
about45 you46

a47 meter48 and49 a50 liter51
nothin52 could53 be54 sweeter55

'cause56 i57 love58 you59 a60 liter61 and62 a63 gram64
and65 it's66 crazy67 that68 i69 am70 for71 you72

(a) Is the span (28,67) a cover for this query? Why or why not?

Even though the span (28,67) satisfies the term set of the query, it is not a cover because it contains a shorter extent that would also satisfy the query. For example, the span includes (33,67) which is a shorter extent than (28,67) which also satisfies the query.

(b) List all of the covers for this query.

Term set from query: {newton, crazy}

The query has the cover set: {(16,28), (33,67)}.

The cover set is a set of spans that are disjoint and contains all the query terms. The yellow highlighted words show the (16,28) span and the blue highlighted words show the (33,67) span.

(c) Using a window size of K=10, calculate the similarity score for this query and the document.

Cover set = {(16,28), (33,67)} = {(p_1, q_1), (p_2, q_1)}

Score for document and query = S(Query Term Set) = I(p_1, q_1) + I(p_2, q_2)

$$= \frac{K}{q_1 - p_1 + 1} + \frac{K}{q_2 - p_2 + 1}$$

$$= \frac{10}{28 - 16 + 1} + \frac{10}{67 - 33 + 1}$$

$= 0.7692 + 0.2857$
$= \mathbf{1.055}$

(d) Using a window size of K=20, calculate the similarity score for this query and the document.

Cover set = {(16,28), (33,67)} = {(p1, q1), (q2, q1)}

Score for Document and Query = S(Term Set) = $I(p_1, q_1) + I(p_2, q_2)$

$$= \frac{K}{q_1 - p_1 + 1} + \frac{K}{q_2 - p_2 + 1}$$

$$= 1 + \frac{20}{67 - 33 + 1}$$

$= 1 + 0.5714$
$= \mathbf{1.571}$

Because the length of the first span(16,28) is shorter than the K value of 20, the score of the first cover would be 1.

2.) 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.

(a) What is the purpose of the parameter λ in this model?

The purpose of λ in this model is to be a smoothing parameter. Smoothing helps to overcome data sparsity by increasing the probability of unseen words by stealing/using the probability of the seen words. Jelinek-Mercer smoothing is a linear interpolation of the document term frequency and collection frequency. The λ value can be a weight between(and including) 1 and 0. The entire λ value adjusts the relative document term frequency and (1- λ) adjusts the relative corpus frequency. Therefore, a high λ value means there is more emphasis or a higher weight on relative term frequency whereas smaller λ value means that there would be more emphasis or a higher weight on the collection frequency.

(b) What would be the effect of setting λ to a value of 0?

A high λ value means there is more emphasis on relative term frequency and therefore we would have a more connected search that tends to retrieve documents containing all the query terms and is better for smaller queries. Whereas, setting λ to a low value of 0, would place all the weight on the collection frequency. A low λ value, such as 0, means that all the emphasis is on the collection frequency and so we would tend to have a higher mean average precision for longer queries and a lower mean average precision for smaller queries. However, as per Lecture 6D, overall changing the λ parameter, does

not have such a drastic impact on the mean average precision as expected. We may have a small increase/decrease in the mean average precision depending on the λ. Keeping the λ = 0.5, for any type of query, would yield very good results.

*3.) 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 "**asiago brie brie derby**". The document collection consists of only the eight documents shown in the table below. Only these five indexing terms are found in the collection. The cells in the table below indicate the number of times a word appears in a document. Use a mixture model with parameter λ = 0.20. Assume 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 four digits of precision (e.g., 1.234x10-8).

	D1	D2	D3	D4	D5	D6	D7	D8		Query - TF	Cum. Freq. of each term
asiago					4					1	4
brie	1	2				1		2		2	6
cheddar	2	2									4
derby	2	3	3	2	3	3	2	2		1	20
edam	3		4	4			5				16

Query = *asiago brie brie derby*

Collection Frequency = 4 + 6 + 4 + 20 + 16 = 50 words
Total Terms in D1 = 1 + 2 + 2 + 3 = 8 words
Total Terms in D2 = 2 + 2 + 3 = 7 words

Model = MLE unigram from documents ; λ = 0.20

Similarity Score for D1 =
[(λ * occurrences of asiago in D1/terms in D1) + (1- λ)*(occurrences of asiago in collection/terms in collection)]
x
2 *[(λ * occurrences of brie in D1/terms in D1) + (1- λ)*(occurrences of brie in collection/terms in collection)]
x
[(λ * occurrences of derby in D1/terms in D1) + (1- λ)*(occurrences of derby in collection/terms in collection)]

$$= \left[(0.20)\left(\frac{0}{8}\right) + (1-0.20)\left(\frac{4}{50}\right)\right] \times\ 2 * \left[(0.20)\left(\frac{1}{8}\right) + (1-0.20)\left(\frac{6}{50}\right)\right] \times \left[(0.20)\left(\frac{2}{8}\right) + (1-0.20)\left(\frac{20}{50}\right)\right]$$

$$= 0.064 \times 2(0.121) \times 0.37$$

$$= \mathbf{5.731 \times 10^{-3}}$$

Similarity Score for D2 =

[(λ * occurrences of asiago in D2/terms in D2) + (1- λ)*(occurrences of asiago in collection/terms in collection)]

x

2 *[(λ * occurrences of brie in D2/terms in D2) + (1- λ)*(occurrences of brie in collection/terms in collection)]

x

[(λ * occurrences of derby in D2/terms in D2) + (1- λ)*(occurrences of derby in collection/terms in collection)]

$$= \left[(0.20)\left(\frac{0}{7}\right) + (1-0.20)\left(\frac{4}{50}\right)\right] \times\ 2 * \left[(0.20)\left(\frac{2}{7}\right) + (1-0.20)\left(\frac{6}{50}\right)\right] \times \left[(0.20)\left(\frac{3}{7}\right) + (1-0.20)\left(\frac{20}{50}\right)\right]$$

$$= 0.064 \times 2(0.153) \times 0.406$$

$$= \mathbf{7.953 \times\ 10^{-3}}$$

Ranking: D2 > D1