CIS 833 – Information Retrieval and Text Mining **Name:_______________**

Homework Assignment 4 – due October 28th

Note: Please remember that you are allowed to discuss the assigned exercises, but you should write your own solution. Identical solutions will receive 0 points.

Exercise 1 (Binary Independence Model)

Consider the following document-term matrix, where a 1 entry indicates that the term occurs in a document, and 0 means it does not:

	t1	t2	t3	t4
d1	0	1	1	1
d2	0	1	1	0
d3	0	1	0	1
d4	1	1	0	0

Assume that the number of non-relevant documents is approximated by the size of the collection and that the probability of occurrence in relevant documents is constant over all the terms in the query (specifically, p_i=0.9).

For each of the following queries, rank the documents in decreasing order of relevance.

```
q1 = {t1, t2}
q2 = {t3}
q3 = {t2, t4}
```

For each query q1, q2, q3, we need to rank documents d1, d2, d3, d4 in decreasing order of relevance. Given a query, we can rank documents by computing the RSV value for each document:

$$RSV = \sum_{x_i = q_i = 1} \log \frac{p_i(1-r_i)}{r_i(1-p_i)}$$

Under the assumption that the number of non-relevant documents is approximated by the size of the collection (which is N=4 here), we have:

$\log (1- r_i)/r_i = \log (N- n)/n \approx \log N/n$ = IDF

Furthermore, we know that $p_i = 0.9$ over all the terms in the query, which means that

$\log p_i (1- p_i) = \log 0.9/(1-0.9) \approx \log 9 \approx 3.16$

For each document, RSV is $\log (1- r_i)/r_i + \log p_i /(1- p_i)$

Consider query q1.

t2 is common to both d1 and q1. Thus, the RSV value for the first document is:

IDF_{t2} + log 9 = log 4/4 + log 9 = 3.16

Similarly, only t2 is common to both d2 and q1. Thus, the RSV value for d2 is:
IDF_{t2} + log 9 = log 1 + log 9 = 3.16

Same for d3, RSV is

IDF_{t2} + log 9 = log 1 + log 9 = 3.16

But q1 and d4 have both t1 and t2 in common, which means RSV is:

IDF_{t1} + log 9 + IDF_{t2} + log 9 = log 4 + log 9 + log 1 + log 9 = 8.33

Ranking: d4 and d1, d2, d3 (in any order, as they have the same score).

Consider query q2.

t3 is common to both d1 and q2. Thus, the RSV value for the first document is:

IDF_{t3} + log 9 = log 4/2 + log 9 = 4.16

Similarly, only t3 is common to both d2 and q1. Thus, the RSV value for d2 is:

IDF_{t3} + log 9 = log 4/2 + log 9 = 4.16

q2 and d3 don't have any common terms, so RSV = 0. Same for q2 and d4.

Ranking: d1 and d2 (in any order), possibly followed by d3 and d4 (in any order - although these might not be included in the result at all, as they don't have any terms in common with the query).

Consider query q3.

t2 and t4 are common to both d1 and q3. Thus, the RSV value for the first document is:

IDF_{t2} + log 9 + IDF_{t4} + log 9 = log 4/4 + log 9 + log 4/2 + log 9 = 7.33

Only t2 is common to both d2 and q3. Thus, the RSV value for d2 is:

IDF_{t2} + log 9 = log 4/4 + log 9 = 3.16

q2 and d3 have t2 and t4 in common, so RSV is:

IDF_{t2} + log 9 + IDF_{t4} + log 9 = log 4/4 + log 9 + log 4/2 + log 9 = 7.33

q2 and d4 have t2 in common, so RSV is:

IDF_{t2} + log 9 = log 4/4 + log 9 = 3.16

Ranking: d1, d3 (in any order), followed by d2, d4 (in any order).

Exercise 2 (Probabilistic Language Models)

Consider a query Q and a collection of documents A,B,C, represented as a document-word count matrix:

	Cat	Food	Fancy
Q	3	4	1
A	2	1	0
B	1	3	1
C	0	2	2

Determine the similarity of A, B, C to Q using language modeling. More precisely, determine the probability of generating the query from the language models associated with the documents using the simple multinomial model and the following smoothing techniques:

(a) No smoothing, i.e., maximum likelihood language model.

(b) Add-1 smoothing

(c) Mixture model smoothing (your choice of lambda)

Solution:

a) We know that:

$$\hat{p}(Q \mid M_d) = \prod_{t \in Q} P_{mle}(t \mid M_d) = \prod_{t \in Q} \frac{tf_{(t,d)}}{dl_d}$$

where:

M_d is the language model of document *d*
$tf_{(t,d)}$ is the raw term frequencies of *t* in document *d*
dl_d is the total number of terms in document *d*

The table below contains $tf_{(t,d)}$ for all terms t and documents d

	Cat	Food	Fancy
A	2/3 = 0.667	1/3 = 0.333	0/3 = 0
B	1/5= 0.2	3/5 = 0.6	1/5 = 0.2
C	0/4 = 0	2/4 = 0.5	2/4 = 0.5

$P(Q|M_A)$ = 0.667^3 * 0.33^4 * 0^1=0

$P(Q|M_B)$ = 0.2^3 * 0.6^4 * 0.2^1 = 0.0002073

$P(Q|M_C)$ = 0^3 * 0.5^4 * 0.5^1=0

b) Here we need to add 1 to each term frequency $tf_{(t,d)}$

The $tf_{(t,d)}$ table considering add-1 smoothing is:

	Cat	Food	Fancy
A	3/6 = 0.5	2/6 = 0.333	1/6 = 0.167
B	2/8 = 0.25	4/8 = 0.5	2/8 = 0.25
C	1/7 = 0.143	3/7 = 0.429	3/7 = 0.429

$P(Q|M_A)$ = 0.5^3 * 0.333^4 * 0.167^1=0.000257

$P(Q|M_B)$ = 0.25^3 * 0.5^4 * 0.25^1 = 0.000244

$P(Q|M_C)$ = 0.143^3 * 0.429^4 * 0.429^1=0.000042

c) Here, we need to use the formula:

$P(w|d) = \lambda P_{mle}(w|M_d) + (1 - \lambda)P_{mle}(w|M_c)$, where M_c is the language model of the collection.

The table below contains $tf_{(t,c)}$ for all terms t in the collection c

	Cat	Food	Fancy
Collection	3/12 = 0.25	6/12 = 0.5	3/12 = 0.25

Considering lambda = 0.5, we have the following probabilities P(w|d):

	Cat	Food	Fancy
A	0.5*(0.667 + 0.25) = 0.459	0.5*(0.333 + 0.5) = 0.417	0.5*(0 + 0.25) = 0.125
B	0.5*(0.2 + 0.25) = 0.225	0.5*(0.6 + 0.5) = 0.55	0.5*(0.2 + 0.25) = 0.225
C	0.5*(0 + 0.25) = 0.125	0.5*(0.5 + 0.5) = 0.5	0.5*(0.5 + 0.25) = 0.375

Therefore:

$P(Q|M_A)$ = 0.459^3 * 0.417^4 * 0.125^1=0.000366

$P(Q|M_B)$ = 0.225^3 * 0.55^4 * 0.225^1 = 0.000235

$P(Q|M_C)$ = 0.125^3 * 0.5^4 * 0.375^1=0.000042