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}
```

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)