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

Homework Assignment 2 (20 points) – due September 25th (by midnight)

Note: Please remember that you are allowed to discuss the assigned exercises, but you should write your own solution.

Exercise 1 (Boolean Model) Which of the documents in the table below will be retrieved given the Boolean query: (chaucer AND (NOT swift)) OR ((NOT chaucer) AND (swift OR shakespeare))

	Chaucer	Milton	Shakespeare	Swift
D1	0	0	0	0
D2	0	0	0	1
D3	0	0	1	0
D4	0	0	1	1
D5	0	1	0	0
D6	0	1	0	1
D7	0	1	1	0
D8	0	1	1	1
D9	1	0	0	0
D10	1	0	0	1
D11	1	0	1	0
D12	1	0	1	1
D13	1	1	0	0
D14	1	1	0	1
D15	1	1	1	0
D16	1	1	1	1

Exercise 2 (Vector Space Model) Given 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

1. Compute the tf .idf representation of the query and of the documents (use log base 2).
2. Compute the cosine similarity of each document to the query Q.

Exercise 3 (Retrieval using an Inverted Index)

i. Construct an inverted index for the following collection of documents. Show the index graphically with linked lists. Your index terms should be stemmed and should not be in the stop words list.

```
Doc 1. John gives a book to Mary
Doc 2. John who reads a book loves Mary
Doc 3. Who does John think Mary love ?
Doc 4. John thinks a book is a good gift
```

ii. What other information can be pre-computed offline, in addition to the actual index?

iii. Show how you can compute document lengths incrementally by parsing the index. Remember to make use of a hashtable.

iv. Consider the query "`love Mary`" and simulate the retrieval of documents in response to this query. Show how the inverted index is used to identify relevant documents and how the cosine similarity between the query and the relevant documents is calculated incrementally using a hashtable.

Note: you don't need to calculate the exact number for this problem; instead, you should explain how each step is done, using the sample documents.