Student Name: EKAL GOLAS Student NetID: EXG140230

University of Texas at Dallas
Department of Computer Science
CS6322 – Information Retrieval
Spring 2016

Instructor: Dr. Sanda Harabagiu

Take-Home Mid-Term Exam

Issued: March 2nd 2016

Due: March 9th 2016 –in class

Problem 1 :

Consider the following three short documents:

Doc #1

Hillary Clinton and Donald Trump are tightening their grips on the Democratic and Republican presidential nominations while Ted Cruz will win Super Tuesday's biggest prize: Texas.

Doc #2

Trump's Virginia win is especially disappointing to Rubio, who had hoped a win there would kick-start his effort to challenge the real estate mogul.

Doc #3

Bernie Sanders, Clinton's insurgent Democratic rival, will capture his home state of Vermont and Oklahoma.

A. (**18 points**) First remove stop words and punctuation; parse manually the documents and select the terms from the 3 documents and created the dictionary (3 points). Create the document vectors by computing three weights: (i) binary weights (3 points); (ii) raw weights (4 points); and (iii) TF-IDF weights (8 points).

For each form of weighting list the document vectors in the following format:

	Term1	Term2	Term3	Term4	Term5	Term6	Term7	Term8	….
DOC1	0	3	1	0	0	2	1	0	….
DOC2	5	0	0	0	3	0	0	2	….
DOC3	3	0	4	3	4	0	0	5	….

SOLUTION 1.A:

(I) Binary Weights:

	DOC1	DOC2	DOC3
bernie	0	0	1
biggest	1	0	0
capture	0	0	1
challenge	0	1	0
clinton	1	0	1
cruz	1	0	0
democratic	1	0	1
disappointing	0	1	0
donald	1	0	0
effort	0	1	0
especially	0	1	0
estate	0	1	0
grips	1	0	0
hillary	1	0	0
home	0	0	1
hoped	0	1	0
insurgent	0	0	1
kick	0	1	0
mogul	0	1	0
nominations	1	0	0
oklahoma	0	0	1
presidential	1	0	0
prize	1	0	0
real	0	1	0
republican	1	0	0
rival	0	0	1
rubio	0	1	0
sanders	0	0	1
start	0	1	0
state	0	0	1
super	1	0	0
ted	1	0	0
texas	1	0	0
tightening	1	0	0
trump	1	1	0
tuesday	1	0	0
vermont	0	0	1
virginia	0	1	0
win	1	1	0

(ii) Raw Weights

	DOC1	DOC2	DOC3
bernie	0	0	1
biggest	1	0	0
capture	0	0	1
challenge	0	1	0
clinton	1	0	1
cruz	1	0	0
democratic	1	0	1
disappointing	0	1	0
donald	1	0	0
effort	0	1	0
especially	0	1	0
estate	0	1	0
grips	1	0	0
hillary	1	0	0
home	0	0	1
hoped	0	1	0
insurgent	0	0	1
kick	0	1	0
mogul	0	1	0
nominations	1	0	0
oklahoma	0	0	1
presidential	1	0	0
prize	1	0	0
real	0	1	0
republican	1	0	0
rival	0	0	1
rubio	0	1	0
sanders	0	0	1
start	0	1	0
state	0	0	1
super	1	0	0
ted	1	0	0
texas	1	0	0
tightening	1	0	0
trump	1	1	0
tuesday	1	0	0
vermont	0	0	1
virginia	0	1	0
win	1	2	0

(iii) tf-idf Weights:

	DOC1	DOC2	DOC3
bernie	0	0	0.4771
biggest	0.4771	0	0
capture	0	0	0.4771
challenge	0	0.4771	0
clinton	0.1761	0	0.1761
cruz	0.4771	0	0
democratic	0.1761	0	0.1761
disappointing	0	0.4771	0
donald	0.4771	0	0
effort	0	0.4771	0
especially	0	0.4771	0
estate	0	0.4771	0
grips	0.4771	0	0
hillary	0.4771	0	0
home	0	0	0.4771
hoped	0	0.4771	0
insurgent	0	0	0.4771
kick	0	0.4771	0
mogul	0	0.4771	0
nominations	0.4771	0	0
oklahoma	0	0	0.4771
presidential	0.4771	0	0
prize	0.4771	0	0
real	0	0.4771	0
republican	0.4771	0	0
rival	0	0	0.4771
rubio	0	0.4771	0
sanders	0	0	0.4771
start	0	0.4771	0
state	0	0	0.4771
super	0.4771	0	0
ted	0.4771	0	0
texas	0.4771	0	0
tightening	0.4771	0	0
trump	0.1761	0.1761	0
tuesday	0.4771	0	0
vermont	0	0	0.4771
virginia	0	0.4771	0
win	0.1761	0.2291	0

B. (**12 points**) Create and inverted (sorted) list index of the three documents, including the dictionary and the postings. The dictionary should also contain (for each term) statistics such as the document frequency of each term. The postings for each term should contain the document ids and the term frequency in that document. List the index for the first 5 terms from the dictionary only. You do not need to list the entire index.

SOLUTION 1.B:

	TF	DF	Doc	Freq
bernie	1	1	3	1
biggest	1	1	1	1
capture	1	1	3	1
challenge	1	1	2	1
clinton	2	2	1	1
			3	1

C. (**12 points**) What are the hit lists for the following Boolean queries (in each case explain how you obtained them from the inverted index):

1. Clinton AND Trump (3 points)
2. (Clinton AND Democratic) OR (Trump AND Republican) (3 points)
3. (Clinton AND Democratic AND Texas) OR source (3 points)
4. (Clinton OR Trump) AND (Cruz OR Rubio) (3 points)

SOLUTION 1.C:

Query 1: Clinton AND Trump
clinton: {1=1, 3=1}
trump: {1=1, 2=1}

Clinton is in Doc 1 and Doc 3
Trump is in Doc 1 and Doc 2

Therefore, only Doc 1 satisfies the query

Query 2: (Clinton AND Democratic) OR (Trump AND Republican)
clinton: {1=1, 3=1}
democratic: {1=1, 3=1}
trump: {1=1, 2=1}
republican: {1=1}

First, divide the query into 2 sub-queries:

1. (Clinton AND Democratic)
    Clinton is in Doc 1 and Doc 3
    Democratic is in Doc 1 and Doc 3
    So, Doc 1 and Doc 3 satisfy the sub-query 1

2. (Trump AND Republican)
    Trump is in Doc 1 and Doc 2
    Republican is in Doc 1
    So, only Doc 1 satisfies the sub-query 2

Therefore, Doc 1 and Doc 3 satisfy the query

Query 3: (Clinton AND Democratic AND Texas) OR source
clinton: {1=1, 3=1}
democratic: {1=1, 3=1}
texas: {1=1}

1. (Clinton AND Democratic AND Texas)
    Clinton is in Doc 1 and Doc 3
    Democratic is in Doc 1 and Doc 3
    Texas is in Doc 1
    So, only Doc 1 satisfies the sub-query

2. source
    No document satisfies this sub-query

Therefore, only Doc 1 satisfies the query

Query 4: **(Clinton OR Trump) AND (Cruz OR Rubio)**
clinton: {1=1, 3=1}
trump: {1=1, 2=1}
cruz: {1=1}
rubio: {2=1}

First, divide the query into 2 sub-queries:

1. (Clinton OR Trump)
    Clinton is in Doc 1 and Doc 3
    Trump is in Doc 1 and Doc 2
    So, Doc 1, Doc 2 and Doc 3 satisfy the sub-query 1

2. (Cruz OR Rubio)
    Cruz is in Doc 1
    Rubio is in Doc 2
    So, Doc 1 and Doc 2 satisfy the sub-query 2

Therefore, Doc 1 and Doc 2 satisfy the query.

D. (**24 points**) Compute the similarity coefficients for each of the four queries and each of the three documents using: (i) the cosine similarity (4 points for each query); (ii) the Jaccard similarity (2 points for each query).

SOLUTION 1.D:

(I) Cosine similarity:

Query 1: Clinton AND Trump

Cosine similarity with doc1: 0.3333

	Query						Document			
	TF	tf-wt	DF	iDF	tf-idf	Normalized	TF	tf-wt	Normalized	Product
grips	0	0	1	0.4771	0	0	1	1	0.2357	0
donald	0	0	1	0.4771	0	0	1	1	0.2357	0
trump	1	1	2	0.1761	0.1761	0.7071	1	1	0.2357	0.1667
republican	0	0	1	0.4771	0	0	1	1	0.2357	0
clinton	1	1	2	0.1761	0.1761	0.7071	1	1	0.2357	0.1667
texas	0	0	1	0.4771	0	0	1	1	0.2357	0
nominations	0	0	1	0.4771	0	0	1	1	0.2357	0
prize	0	0	1	0.4771	0	0	1	1	0.2357	0
super	0	0	1	0.4771	0	0	1	1	0.2357	0
ted	0	0	1	0.4771	0	0	1	1	0.2357	0
tuesday	0	0	1	0.4771	0	0	1	1	0.2357	0
biggest	0	0	1	0.4771	0	0	1	1	0.2357	0
tightening	0	0	1	0.4771	0	0	1	1	0.2357	0
democratic	0	0	2	0.1761	0	0	1	1	0.2357	0
cruz	0	0	1	0.4771	0	0	1	1	0.2357	0
presidential	0	0	1	0.4771	0	0	1	1	0.2357	0
win	0	0	2	0.1761	0	0	1	1	0.2357	0
hillary	0	0	1	0.4771	0	0	1	1	0.2357	0

Cosine similarity with doc2: 0.1845

	Query						Document			
	TF	tf-wt	DF	iDF	tf-idf	Normalized	TF	tf-wt	Normalized	Product
trump	1	1	2	0.1761	0.1761	0.7071	1	1	0.2609	0.1845
hoped	0	0	1	0.4771	0	0	1	1	0.2609	0
start	0	0	1	0.4771	0	0	1	1	0.2609	0
estate	0	0	1	0.4771	0	0	1	1	0.2609	0
mogul	0	0	1	0.4771	0	0	1	1	0.2609	0
effort	0	0	1	0.4771	0	0	1	1	0.2609	0
clinton	1	1	2	0.1761	0.1761	0.7071	0	0	0	0
real	0	0	1	0.4771	0	0	1	1	0.2609	0
especially	0	0	1	0.4771	0	0	1	1	0.2609	0
kick	0	0	1	0.4771	0	0	1	1	0.2609	0
rubio	0	0	1	0.4771	0	0	1	1	0.2609	0
challenge	0	0	1	0.4771	0	0	1	1	0.2609	0
virginia	0	0	1	0.4771	0	0	1	1	0.2609	0
disappointing	0	0	1	0.4771	0	0	1	1	0.2609	0
win	0	0	2	0.1761	0	0	2	1.301	0.3394	0

Cosine similarity with doc3: 0.2132

	Query						Document			
	TF	tf-wt	DF	iDF	tf-idf	Normalized	TF	tf-wt	Normalized	Product
insurgent	0	0	1	0.4771	0	0	1	1	0.3015	0
trump	1	1	2	0.1761	0.1761	0.7071	0	0	0	0
bernie	0	0	1	0.4771	0	0	1	1	0.3015	0
democratic	0	0	2	0.1761	0	0	1	1	0.3015	0
capture	0	0	1	0.4771	0	0	1	1	0.3015	0
vermont	0	0	1	0.4771	0	0	1	1	0.3015	0
clinton	1	1	2	0.1761	0.1761	0.7071	1	1	0.3015	0.2132
state	0	0	1	0.4771	0	0	1	1	0.3015	0
oklahoma	0	0	1	0.4771	0	0	1	1	0.3015	0
sanders	0	0	1	0.4771	0	0	1	1	0.3015	0
rival	0	0	1	0.4771	0	0	1	1	0.3015	0
home	0	0	1	0.4771	0	0	1	1	0.3015	0

Query 2:(Clinton AND Democratic) OR (Trump AND Republican)

Cosine similarity with doc1: 0.4185

	Query						Document			
	TF	tf-wt	DF	iDF	tf-idf	Normalized	TF	tf-wt	Normalized	Product
grips	0	0	1	0.4771	0	0	1	1	0.2357	0
donald	0	0	1	0.4771	0	0	1	1	0.2357	0
trump	1	1	2	0.1761	0.1761	0.311	1	1	0.2357	0.0733
republican	1	1	1	0.4771	0.4771	0.8426	1	1	0.2357	0.1986
clinton	1	1	2	0.1761	0.1761	0.311	1	1	0.2357	0.0733
texas	0	0	1	0.4771	0	0	1	1	0.2357	0
nominations	0	0	1	0.4771	0	0	1	1	0.2357	0
prize	0	0	1	0.4771	0	0	1	1	0.2357	0
super	0	0	1	0.4771	0	0	1	1	0.2357	0
ted	0	0	1	0.4771	0	0	1	1	0.2357	0
tuesday	0	0	1	0.4771	0	0	1	1	0.2357	0
biggest	0	0	1	0.4771	0	0	1	1	0.2357	0
tightening	0	0	1	0.4771	0	0	1	1	0.2357	0
democratic	1	1	2	0.1761	0.1761	0.311	1	1	0.2357	0.0733
cruz	0	0	1	0.4771	0	0	1	1	0.2357	0
presidential	0	0	1	0.4771	0	0	1	1	0.2357	0
win	0	0	2	0.1761	0	0	1	1	0.2357	0
hillary	0	0	1	0.4771	0	0	1	1	0.2357	0

Cosine similarity with doc2: 0.0811

	Query						Document			
	TF	tf-wt	DF	iDF	tf-idf	Normalized	TF	tf-wt	Normalized	Product
trump	1	1	2	0.1761	0.1761	0.311	1	1	0.2609	0.0811
hoped	0	0	1	0.4771	0	0	1	1	0.2609	0
republican	1	1	1	0.4771	0.4771	0.8426	0	0	0	0
start	0	0	1	0.4771	0	0	1	1	0.2609	0
estate	0	0	1	0.4771	0	0	1	1	0.2609	0
mogul	0	0	1	0.4771	0	0	1	1	0.2609	0
effort	0	0	1	0.4771	0	0	1	1	0.2609	0
clinton	1	1	2	0.1761	0.1761	0.311	0	0	0	0
real	0	0	1	0.4771	0	0	1	1	0.2609	0
especially	0	0	1	0.4771	0	0	1	1	0.2609	0
kick	0	0	1	0.4771	0	0	1	1	0.2609	0
rubio	0	0	1	0.4771	0	0	1	1	0.2609	0
democratic	1	1	2	0.1761	0.1761	0.311	0	0	0	0
challenge	0	0	1	0.4771	0	0	1	1	0.2609	0
virginia	0	0	1	0.4771	0	0	1	1	0.2609	0
disappointing	0	0	1	0.4771	0	0	1	1	0.2609	0
win	0	0	2	0.1761	0	0	2	1.301	0.3394	0

Cosine similarity with doc3: 0.1875

	Query						Document			
	TF	tf-wt	DF	iDF	tf-idf	Normalized	TF	tf-wt	Normalized	Product
trump	1	1	2	0.1761	0.1761	0.311	0	0	0	0
bernie	0	0	1	0.4771	0	0	1	1	0.3015	0
republican	1	1	1	0.4771	0.4771	0.8426	0	0	0	0
capture	0	0	1	0.4771	0	0	1	1	0.3015	0
clinton	1	1	2	0.1761	0.1761	0.311	1	1	0.3015	0.0938
sanders	0	0	1	0.4771	0	0	1	1	0.3015	0
home	0	0	1	0.4771	0	0	1	1	0.3015	0
insurgent	0	0	1	0.4771	0	0	1	1	0.3015	0
democratic	1	1	2	0.1761	0.1761	0.311	1	1	0.3015	0.0938
vermont	0	0	1	0.4771	0	0	1	1	0.3015	0
state	0	0	1	0.4771	0	0	1	1	0.3015	0
oklahoma	0	0	1	0.4771	0	0	1	1	0.3015	0
rival	0	0	1	0.4771	0	0	1	1	0.3015	0

Query 3: (Clinton AND Democratic AND Texas) OR source

Cosine similarity with doc1: 0.3632

	Query						Document			
	TF	tf-wt	DF	iDF	tf-idf	Normalized	TF	tf-wt	Normalized	Product
grips	0	0	1	0.4771	0	0	1	1	0.2357	0
donald	0	0	1	0.4771	0	0	1	1	0.2357	0
trump	0	0	2	0.1761	0	0	1	1	0.2357	0
republican	0	0	1	0.4771	0	0	1	1	0.2357	0
clinton	1	1	2	0.1761	0.1761	0.3272	1	1	0.2357	0.0771
texas	1	1	1	0.4771	0.4771	0.8865	1	1	0.2357	0.209
nominations	0	0	1	0.4771	0	0	1	1	0.2357	0
prize	0	0	1	0.4771	0	0	1	1	0.2357	0
super	0	0	1	0.4771	0	0	1	1	0.2357	0
ted	0	0	1	0.4771	0	0	1	1	0.2357	0
tuesday	0	0	1	0.4771	0	0	1	1	0.2357	0
biggest	0	0	1	0.4771	0	0	1	1	0.2357	0
tightening	0	0	1	0.4771	0	0	1	1	0.2357	0
democratic	1	1	2	0.1761	0.1761	0.3272	1	1	0.2357	0.0771
cruz	0	0	1	0.4771	0	0	1	1	0.2357	0
presidential	0	0	1	0.4771	0	0	1	1	0.2357	0
win	0	0	2	0.1761	0	0	1	1	0.2357	0
hillary	0	0	1	0.4771	0	0	1	1	0.2357	0

Cosine similarity with doc2: 0

	Query						Document			
	TF	tf-wt	DF	iDF	tf-idf	Normalized	TF	tf-wt	Normalized	Product
trump	0	0	2	0.1761	0	0	1	1	0.2609	0
hoped	0	0	1	0.4771	0	0	1	1	0.2609	0
republican	0	0	1	0.4771	0	0	0	0	0	0
start	0	0	1	0.4771	0	0	1	1	0.2609	0
estate	0	0	1	0.4771	0	0	1	1	0.2609	0
mogul	0	0	1	0.4771	0	0	1	1	0.2609	0
effort	0	0	1	0.4771	0	0	1	1	0.2609	0
clinton	1	1	2	0.1761	0.1761	0.3272	0	0	0	0
real	0	0	1	0.4771	0	0	1	1	0.2609	0
texas	1	1	1	0.4771	0.4771	0.8865	0	0	0	0
especially	0	0	1	0.4771	0	0	1	1	0.2609	0
kick	0	0	1	0.4771	0	0	1	1	0.2609	0
rubio	0	0	1	0.4771	0	0	1	1	0.2609	0
democratic	1	1	2	0.1761	0.1761	0.3272	0	0	0	0
challenge	0	0	1	0.4771	0	0	1	1	0.2609	0
virginia	0	0	1	0.4771	0	0	1	1	0.2609	0
disappointing	0	0	1	0.4771	0	0	1	1	0.2609	0
win	0	0	2	0.1761	0	0	2	1.301	0.3394	0

Cosine similarity with doc3: 0.1973

	Query						Document			
	TF	tf-wt	DF	iDF	tf-idf	Normalized	TF	tf-wt	Normalized	Product
trump	0	0	2	0.1761	0	0	0	0	0	0
bernie	0	0	1	0.4771	0	0	1	1	0.3015	0
republican	0	0	1	0.4771	0	0	0	0	0	0
capture	0	0	1	0.4771	0	0	1	1	0.3015	0
clinton	1	1	2	0.1761	0.1761	0.3272	1	1	0.3015	0.0986
texas	1	1	1	0.4771	0.4771	0.8865	0	0	0	0
sanders	0	0	1	0.4771	0	0	1	1	0.3015	0
home	0	0	1	0.4771	0	0	1	1	0.3015	0
insurgent	0	0	1	0.4771	0	0	1	1	0.3015	0
democratic	1	1	2	0.1761	0.1761	0.3272	1	1	0.3015	0.0986
vermont	0	0	1	0.4771	0	0	1	1	0.3015	0
state	0	0	1	0.4771	0	0	1	1	0.3015	0
oklahoma	0	0	1	0.4771	0	0	1	1	0.3015	0
rival	0	0	1	0.4771	0	0	1	1	0.3015	0

Query 4: (Clinton OR Trump) AND (Cruz OR Rubio)

Cosine similarity with doc1: 0.2718

	Query						Document			
	TF	tf-wt	DF	iDF	tf-idf	Normalized	TF	tf-wt	Normalized	Product
grips	0	0	1	0.4771	0	0	1	1	0.2357	0
donald	0	0	1	0.4771	0	0	1	1	0.2357	0
trump	1	1	2	0.1761	0.1761	0.2448	1	1	0.2357	0.0577
republican	0	0	1	0.4771	0	0	1	1	0.2357	0
clinton	1	1	2	0.1761	0.1761	0.2448	1	1	0.2357	0.0577
texas	0	0	1	0.4771	0	0	1	1	0.2357	0
nominations	0	0	1	0.4771	0	0	1	1	0.2357	0
prize	0	0	1	0.4771	0	0	1	1	0.2357	0
super	0	0	1	0.4771	0	0	1	1	0.2357	0
ted	0	0	1	0.4771	0	0	1	1	0.2357	0
tuesday	0	0	1	0.4771	0	0	1	1	0.2357	0
biggest	0	0	1	0.4771	0	0	1	1	0.2357	0
tightening	0	0	1	0.4771	0	0	1	1	0.2357	0
democratic	0	0	2	0.1761	0	0	1	1	0.2357	0
rubio	1	1	1	0.4771	0.4771	0.6634	0	0	0	0
cruz	1	1	1	0.4771	0.4771	0.6634	1	1	0.2357	0.1564
presidential	0	0	1	0.4771	0	0	1	1	0.2357	0
win	0	0	2	0.1761	0	0	1	1	0.2357	0
hillary	0	0	1	0.4771	0	0	1	1	0.2357	0

Cosine similarity with doc2: 0.2369

	Query						Document			
	TF	tf-wt	DF	iDF	tf-idf	Normalized	TF	tf-wt	Normalized	Product
trump	1	1	2	0.1761	0.1761	0.2448	1	1	0.2609	0.0639
hoped	0	0	1	0.4771	0	0	1	1	0.2609	0
republican	0	0	1	0.4771	0	0	0	0	0	0
start	0	0	1	0.4771	0	0	1	1	0.2609	0
estate	0	0	1	0.4771	0	0	1	1	0.2609	0
mogul	0	0	1	0.4771	0	0	1	1	0.2609	0
effort	0	0	1	0.4771	0	0	1	1	0.2609	0
clinton	1	1	2	0.1761	0.1761	0.2448	0	0	0	0
real	0	0	1	0.4771	0	0	1	1	0.2609	0
texas	0	0	1	0.4771	0	0	0	0	0	0
especially	0	0	1	0.4771	0	0	1	1	0.2609	0
kick	0	0	1	0.4771	0	0	1	1	0.2609	0
rubio	1	1	1	0.4771	0.4771	0.6634	1	1	0.2609	0.1731
democratic	0	0	2	0.1761	0	0	0	0	0	0
challenge	0	0	1	0.4771	0	0	1	1	0.2609	0
cruz	1	1	1	0.4771	0.4771	0.6634	0	0	0	0
virginia	0	0	1	0.4771	0	0	1	1	0.2609	0
disappointing	0	0	1	0.4771	0	0	1	1	0.2609	0
win	0	0	2	0.1761	0	0	2	1.301	0.3394	0

Cosine similarity with doc3: 0.0738

	Query						Document			
	TF	tf-wt	DF	iDF	tf-idf	Normalized	TF	tf-wt	Normalized	Product
trump	1	1	2	0.1761	0.1761	0.2448	0	0	0	0
bernie	0	0	1	0.4771	0	0	1	1	0.3015	0
republican	0	0	1	0.4771	0	0	0	0	0	0
capture	0	0	1	0.4771	0	0	1	1	0.3015	0
clinton	1	1	2	0.1761	0.1761	0.2448	1	1	0.3015	0.0738
texas	0	0	1	0.4771	0	0	0	0	0	0
sanders	0	0	1	0.4771	0	0	1	1	0.3015	0
home	0	0	1	0.4771	0	0	1	1	0.3015	0
insurgent	0	0	1	0.4771	0	0	1	1	0.3015	0
democratic	0	0	2	0.1761	0	0	1	1	0.3015	0
rubio	1	1	1	0.4771	0.4771	0.6634	0	0	0	0
vermont	0	0	1	0.4771	0	0	1	1	0.3015	0
cruz	1	1	1	0.4771	0.4771	0.6634	0	0	0	0
state	0	0	1	0.4771	0	0	1	1	0.3015	0
oklahoma	0	0	1	0.4771	0	0	1	1	0.3015	0
rival	0	0	1	0.4771	0	0	1	1	0.3015	0

(ii) Jaccard Similarity:

Query 1: Clinton AND Trump
Intersection with doc1: 2
Union with doc1: 18
Jaccard similiarity: 0.1111

Intersection with doc2: 1
Union with doc2: 15
Jaccard similiarity: 0.0667

Intersection with doc3: 1
Union with doc3: 12
Jaccard similiarity: 0.0833

Query 2:(Clinton AND Democratic) OR (Trump AND Republican)
Intersection with doc1: 4
Union with doc1: 18
Jaccard similiarity: 0.2222

Intersection with doc2: 1
Union with doc2: 17
Jaccard similiarity: 0.0588

Intersection with doc3: 2
Union with doc3: 13
Jaccard similiarity: 0.1538

Query 3: (Clinton AND Democratic AND Texas) OR source
Intersection with doc1: 3
Union with doc1: 18
Jaccard similiarity: 0.1667

Intersection with doc2: 0
Union with doc2: 17
Jaccard similiarity: 0

Intersection with doc3: 2
Union with doc3: 12
Jaccard similiarity: 0.1667

Query 4: (Clinton OR Trump) AND (Cruz OR Rubio)
Intersection with doc1: 3
Union with doc1: 19
Jaccard similiarity: 0.1579

Intersection with doc2: 2
Union with doc2: 16
Jaccard similiarity: 0.125

Intersection with doc3: 1
Union with doc3: 14
Jaccard similiarity: 0.0714

Problem 2 :

Suppose you have a collection of 5 documents, and only 10 terms are used. The following table represents the incidence matrix for the collection:

	Term1	Term2	Term3	Term4	Term5	Term6	Term7	Term8	Term9	Term10
DOC1	0	3	5	0	2	3	1	0	1	2
DOC2	5	0	2	5	3	0	0	3	1	4
DOC3	1	0	2	6	4	0	1	5	0	3
DOC4	3	5	0	3	0	1	6	0	2	1
DOC5	2	4	0	0	0	3	4	2	3	0

List the values of the gaps for the first four terms in your index computed for this collection. Encode these gaps with (i) unary codes (**8 points**); (ii) Elias gamma codes (**16 points**); and (iii) Elias delta codes (**10 points**). You are allowed to write a program to enable you computing the codes. Please add to the exam the code of the program if you chose to use one.

SOLUTION 2:

Gamma (1): Binary (1) = 1, Len (-) = 0, Unary (0) = 0, Gamma (1) = 0
Gamma (2): Binary (2) = 10, Len (0) = 1, Unary (1) = 10, Gamma (2) = 100
Gamma (3): Binary (3) = 11, Len (1) = 1, Unary (1) = 10, Gamma (3) = 101

Delta (1): Binary (1) = 1, Len (1) = 1, Gamma (1) = 0, Delta (1) = 0
Delta (2): Binary (2) = 10, Len (10) = 2, Gamma (2) = 100, Delta (2) = 1000
Delta (3): Binary (3) = 11, Len (11) = 2, Gamma (2) = 100, Delta (3) = 1001

Term 1

Documents	**Gap**	**Unary code**	**Elias Gamma Code**	**Elias Delta Code**
2	2	110	100	1000
3	1	10	0	0
4	1	10	0	0
5	1	10	0	0

Term 2

Documents	**Gap**	**Unary code**	**Elias Gamma Code**	**Elias Delta Code**
1	1	10	0	0
4	3	1110	101	1001
5	1	10	0	0

Term 3

Documents	**Gap**	**Unary code**	**Elias Gamma Code**	**Elias Delta Code**
1	1	10	0	0
2	1	10	0	0
3	1	10	0	0

Term 4

Documents	Gap	Unary code	Elias Gamma Code	Elias Delta Code
2	2	110	100	1000
3	1	10	0	0
4	1	10	0	0