Student Name: Student NetID:

University of Texas at Dallas
Department of Computer Science
CS6322 – Information Retrieval
Fall 2018

Instructor: Dr. Sanda Harabagiu

Take-Home Mid-Term Exam

Issued: October 10th 2018

Due: October 17th 2018 – before MidNight

Submit in eLEarning as PDF file

DO NOT DELETE ANY PROBLEM, Simply add your answers!!!!!

If you submit only solution with no problems, you will receive 0 points!!!!

Problem 1 :

Consider the following three short documents:

Doc #1

For a few days this summer, Alexa, the voice assistant who speaks to me through my Amazon Echo Dot, took to ending our interactions with a whisper: Sweet dreams

Doc #2

We're all falling for Alexa, unless we're falling for Google Assistant, or Siri, or some other genie in a smart speaker. When I say "smart," I mean the speakers possess artificial intelligence, can conduct basic conversations, and are hooked up to the internet, which allows them to look stuff up and do things for you.

Doc #3

Perhaps you think that talking to Alexa is just a new way to do the things you already do on a screen: shopping, catching up on the news, trying to figure out whether your dog is sick or just depressed. It's not that simple.

A. (**18 points**) First remove stop words and punctuation; parse manually the documents and select the terms which have the lexical roles **only of nouns**, **verbs, adjectives** or **adverbs.** Next, reduce all the terms to their lemmas, e.g. "interactions" becomes "interaction", "ending" becomes "end" etc. With the generation of the lemmas from the 3 documents you are ready to create the dictionary and present it (5 points if the dictionary is correct). Generate also the document vectors by computing three weights:
   (i) binary weights (3 points);
   (ii) raw weights (3 points); and
   (iii) TF-IDF weights (7 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:

The Dictionary:

Term	Document Frequency	Total Frequency
alexa	3	3
allow	1	1
already	1	1
amazon	1	1
artificial	1	1
assistant	2	2
basic	1	1
catch	1	1
conduct	1	1
conversation	1	1
day	1	1
depressed	1	1
dog	1	1
dot	1	1
dream	1	1
echo	1	1
end	1	1
fall	1	2
figure	1	1
genie	1	1
google	1	1
hook	1	1
intelligence	1	1

interaction	1	1
internet	1	1
look	1	1
mean	1	1
new	1	1
news	1	1
perhaps	1	1
possess	1	1
say	1	1
screen	1	1
shop	1	1
sick	1	1
simple	1	1
siri	1	1
smart	1	2
speak	1	1
speaker	1	2
stuff	1	1
summer	1	1
sweet	1	1
take	1	1
talk	1	1
thing	2	1
think	1	2
try	1	1
voice	1	1
way	1	1
whisper	1	1

The document vectors:
(i)

	Doc1	Doc2	Doc3	Weight
alexa	1	1	1	111
allow	0	1	0	010
already	0	0	1	001
amazon	1	0	0	100
artificial	0	1	0	010
assistant	1	1	0	110
basic	0	1	0	010
catch	0	0	1	001
conduct	0	1	0	010
conversation	0	1	0	010
day	1	0	0	100
depressed	0	0	1	001
dog	0	0	1	001
dot	1	0	0	100
dream	1	0	0	100
echo	1	0	0	100
end	1	0	0	100
fall	0	1	0	010
figure	0	0	1	001
genie	0	1	0	010
google	0	1	0	010
hook	0	1	0	010
intelligence	0	1	0	010
interaction	1	0	0	100
internet	0	1	0	010
look	0	1	0	010
mean	0	1	0	010
new	0	0	1	001
news	0	0	1	001
perhaps	0	0	1	001
possess	0	1	0	010
say	0	1	0	010
screen	0	0	1	001
shop	0	0	1	001
sick	0	0	1	001
simple	0	0	1	001
siri	0	1	0	010

smart	0	1	0	010
speak	1	0	0	100
speaker	0	1	0	010
stuff	0	1	0	010
summer	1	0	0	100
sweet	1	0	0	100
take	1	0	0	100
talk	0	0	1	001
thing	0	1	1	011
think	0	0	1	001
try	0	0	1	001
voice	1	0	0	100
way	0	1	0	010
whisper	1	0	0	100

(ii)

	Doc1	dOC2	DOC3	Weight
alexa	1	1	1	3
allow	0	1	0	1
already	0	0	1	1
amazon	1	0	0	1
artificial	0	1	0	1
assistant	1	1	0	1
basic	0	1	0	1
catch	0	0	1	1
conduct	0	1	0	1
conversation	0	1	0	1
day	1	0	0	1
depressed	0	0	1	1
dog	0	0	1	1
dot	1	0	0	1
dream	1	0	0	1
echo	1	0	0	1
end	1	0	0	1
fall	0	2	0	2
figure	0	0	1	1
genie	0	1	0	1
google	0	1	0	1
hook	0	1	0	1
intelligence	0	1	0	1
interaction	1	0	0	1
internet	0	1	0	1
look	0	1	0	1
mean	0	1	0	1
new	0	0	1	1
news	0	0	1	1

perhaps	0	0	1	1
possess	0	1	0	1
say	0	1	0	1
screen	0	0	1	1
shop	0	0	1	1
sick	0	0	1	1
simple	0	0	1	1
siri	0	1	0	1
smart	0	2	0	2
speak	1	0	0	1
speaker	0	2	0	2
stuff	0	1	0	1
summer	1	0	0	1
sweet	1	0	0	1
take	1	0	0	1
talk	0	0	1	1
think	0	0	1	1
thing	0	1	1	1
try	0	0	1	1
voice	1	0	0	1
way	0	1	0	1
whisper	1	0	0	1

(iii)

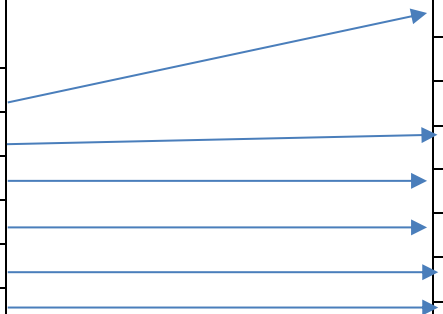
	tf-idf1	tf-id2	tf-idf3
alexa	0	0	0
allow	0	0.47712125	0
already	0	0	0.47712125
amazon	0.47712125	0	0
artificial	0	0.47712125	0
assistant	0.17609126	0.17609126	0
basic	0	0.47712125	0
catch	0	0	0.47712125
conduct	0	0.47712125	0
conversation	0	0.47712125	0
day	0.47712125	0	0
depressed	0	0	0.47712125
dog	0	0	0.47712125
dot	0.47712125	0	0
dream	0.47712125	0	0
echo	0.47712125	0	0
end	0.47712125	0	0
fall	0	0.62074906	0
figure	0	0	0.47712125
genie	0	0.47712125	0
google	0	0.47712125	0
hook	0	0.47712125	0
intelligence	0	0.47712125	0
interaction	0.47712125	0	0
internet	0	0.47712125	0
look	0	0.47712125	0
mean	0	0.47712125	0
new	0	0	0.47712125
news	0	0	0.47712125
perhaps	0	0	0.47712125
possess	0	0.47712125	0
say	0	0.47712125	0
screen	0	0	0.47712125
shop	0	0	0.47712125
sick	0	0	0.47712125
simple	0	0	0.47712125
siri	0	0.47712125	0
smart	0	0.62074906	0

speak	0.47712125	0	0
speaker	0	0.62074906	0
stuff	0	0.47712125	0
summer	0.47712125	0	0
sweet	0.47712125	0	0
take	0.47712125	0	0
talk	0	0	0.47712125
thing	0	0.17609126	0.17609126
think	0	0	0.47712125
try	0	0	0.47712125
voice	0.47712125	0	0
way	0	0	0.47712125
whisper	0.47712125	0	0

B. (**12 points**) Create an 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 and the total number of occurrences of the term in the collection (collection frequency). The postings for each term should contain the document id and the term frequencies in the respective document. **List the index for the first 6 terms from the dictionary only**. You do not need to list the entire index. You will get credit only for the first 6 terms from the dictionary and their respective inverted lists.

SOLUTION 1.B:

Term	Document Frequency	Total Frequency
alexa	3	3
allow	1	1
already	1	1
amazon	1	1
artificial	1	1
assistant	2	2



DOC#	Frequency
1	1
2	1
3	1
1	1
1	1
1	1
1	1
1	1
2	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):

Q1. Alexa AND simple AND screen (3 points)
Q2. (Alexa AND whisper) OR (speaker AND conversation) (3 points)
Q3. (Alexa AND smart AND intelligence) OR talk (3 points)
Q4. (fall OR think) AND (artificial OR screen) (3 points)

SOLUTION 1.C:

Q1:

Terms	Doc1	Doc2	Doc3
Alexa	1	1	1
Simple	0	0	1
Screen	0	0	1
Result of AND operation	0	0	1

Only Doc3 contains all 3 terms.
Thus, answer is Doc3

Q2.

Alexa AND whisper

Terms	Doc1	Doc2	Doc3
Alexa	1	1	1
Whisper	1	0	0
Ans1	1	0	0

Terms	Doc1	Doc2	Doc3
Speaker	0	1	0
conversation	0	1	0
Ans2	0	1	0

Terms	Doc1	Doc2	Doc3
Ans1	1	0	0
Ans2	0	1	0
Result	1	1	0

Answer:
- Doc1
- Doc2

Q3:

Terms	Doc1	Doc2	Doc3
Alexa	1	1	1
Smart	0	1	0
Intelligence	0	1	0
ANS1	0	1	0

Terms	Doc1	Doc2	Doc3
ANS1	0	1	0
Talk	0	0	1
Result	0	1	1

Answer:
- Doc2
- Doc3

Q4]

Terms	Doc1	Doc2	Doc3
Fall	0	1	0
Think	0	0	1
Ans1	0	1	1

Terms	Doc1	Doc2	Doc3
Artificial	0	1	0
Screen	0	0	1
Ans2	0	1	1

Terms	Doc1	Doc2	Doc3
Ans1	0	1	1
Ans2	0	1	1
Result	0	1	1

Ans:
- Doc2
- Doc3

D. (**24 points**) Compute the similarity coefficients for each of the four queries from 1.C and each of the three documents using: (i) the cosine similarity with the lnc.ltc scheme – which stands for the SMART notation: ddd.qqq (4 points for each query); (ii) the Jaccard similarity (2 points for each query).

SOLUTION 1.D: (i)

Cos(Q1,D1)=

		q1							d1		
	tf-raw	tf-wt	DF	IDF	TF-IDF	N'LIZE	tf-raw	tf-wt	WT	N'LIZE	Product
alexa	1	1	3	0	0	0	1	1	1	0.2581	0
amazon	0	0	1	0.4771	0	0	1	1	1	0.2581	0
assistant	0	0	2	0.1760	0	0	1	1	1	0.2581	0
day	0	0	1	0.4771	0	0	1	1	1	0.2581	0
dot	0	0	1	0.4771	0	0	1	1	1	0.2581	0
dream	0	0	1	0.4771	0	0	1	1	1	0.2581	0
echo	0	0	1	0.4771	0	0	1	1	1	0.2581	0
end	0	0	1	0.4771	0	0	1	1	1	0.2581	0
interaction	0	0	1	0.4771	0	0	1	1	1	0.2581	0
screen	1	1	1	0.4771	0.4771	0.7071	0	0	0	0	0
simple	1	1	1	0.4771	0.4771	0.7071	0	0	0	0	0
speak	0	0	1	0.4771	0	0	1	1	1	0.2581	0
summer	0	0	1	0.4771	0	0	1	1	1	0.2581	0
sweet	0	0	1	0.4771	0	0	1	1	1	0.2581	0
take	0	0	1	0.4771	0	0	1	1	1	0.2581	0
voice	0	0	1	0.4771	0	0	1	1	1	0.2581	0
whisper	0	0	1	0.4771	0	0	1	1	1	0.2581	0
										Score	0

Cos(Q1,D1)=0

Cos(Q1,D2)=0

		q1							D2		
	tf-raw	tf-wt	DF	idf	TF IDF	nlize	Tf-raw	tf-wt	WT	nLIZE	pRODUCT
alexa	1	1	3	0	0	0	1	1	1	0.2236068	0
allow	0	0	1	0.4771	0	0	1	1	1	0.2236068	0

artificial	0	0	1	0.4771	0	0	1	1	1	0.2132	0
assistant	0	0	2	0.1760	0	0	1	1	1	0.2132	0
basic	0	0	1	0.4771	0	0	1	1	1	0.2132	0
conduct	0	0	1	0.4771	0	0	1	1	1	0.2132	0
conversation	0	0	1	0.4771	0	0	1	1	1	0.2132	0
fall	0	0	1	0.4771	0	0	1	1	1	0.2132	0
genie	0	0	1	0.4771	0	0	1	1	1	0.2132	0
google	0	0	1	0.4771	0	0	1	1	1	0.2132	0
hook	0	0	1	0.4771	0	0	1	1	1	0.2132	0
intelligence	0	0	1	0.4771	0	0	1	1	1	0.2132	0
internet	0	0	1	0.4771	0	0	1	1	1	0.2132	0
look	0	0	1	0.4771	0	0	1	1	1	0.2132	0
mean	0	0	1	0.4771	0	0	1	1	1	0.2132	0
possess	0	0	1	0.4771	0	0	1	1	1	0.2132	0
say	0	0	1	0.4771	0	0	1	1	1	0.2132	0
screen	1	1	1	0.4771	0.4771	0.7071	0	0	0	0	0
simple	1	1	1	0.4771	0.4771	0.7071	0	0	0	0	0
siri	0	0	1	0.4771	0	0	1	1	1	0.2132	0
smart	0	0	1	0.4771	0	0	1	1	1	0.2132	0
speaker	0	0	1	0.4771	0	0	1	1	1	0.2132	0
stuff	0	0	1	0.4771	0	0	1	1	1	0.2132	0
thing	0	0	2	0.1760	0	0	1	1	1	0.2132	0
										SCORE	0

Cos(Q1,D2)=0

Cos(Q1,D3)=

				Q1				D3			
	tf-raw	tf-wt	DF	idf	TF IDF	nlize	Tf-raw	tf-wt	WT	nLIZE	pRODUCT
alexa	1	1	3	0	0	0	1	1	1	0.2357	0
already	0	0	1	0.4771	0	0	1	1	1	0.2357	0

catch	0	0	1	0.4771	0	0	1	1	1	0.2357	0
depressed	0	0	1	0.4771	0	0	1	1	1	0.2357	0
dog	0	0	1	0.4771	0	0	1	1	1	0.2357	0
figure	0	0	1	0.4771	0	0	1	1	1	0.2357	0
new	0	0	1	0.4771	0	0	1	1	1	0.2357	0
news	0	0	1	0.4771	0	0	1	1	1	0.2357	0
perhaps	0	0	1	0.4771	0	0	1	1	1	0.2357	0
screen	1	1	1	0.4771	0.4771	0.7071	1	1	1	0.2357	0.16666347
shop	0	0	1	0.4771	0	0	1	1	1	0.2357	0
sick	0	0	1	0.4771	0	0	1	1	1	0.2357	0
simple	1	1	1	0.4771	0.4771	0.7071	1	1	1	0.2357	0.16666347
talk	0	0	1	0.4771	0	0	1	1	1	0.2357	0
thing	0	0	2	0.4771	0	0	1	1	1	0.2357	0
think	0	0	1	0.4771	0	0	1	1	1	0.2357	0
try	0	0	1	0.4771	0	0	1	1	1	0.2357	0
way	0	0	1	0.4771	0	0	1	1	1	0.2357	0
										sCORE	0.33332694

Cos(Q1,D3)= 0.3333269

Cos(Q2,D1)=

		Q2							D1		
	tf-raw	tf-wt	DF	idf	TF IDF	nlize	Tf-raw	tf-wt	WT	nLIZE	pRODUCT
alexa	1	1	3	0	0	0	1	1	1	0.258	
amazon	0	0	1	0.4771	0	0	1	1	1	0.258	0
conversation	1	1	1	0.4771	0.4771	0.5773	0	0	0	0	0
day	0	0	1	0.4771	0	0	1	1	1	0.258	0
dot	0	0	1	0.4771	0	0	1	1	1	0.258	0
dream	0	0	1	0.4771	0	0	1	1	1	0.258	0
echo	0	0	1	0.4771	0	0	1	1	1	0.258	0
end	0	0	1	0.4771	0	0	1	1	1	0.258	0
interaction	0	0	1	0.4771	0	0	1	1	1	0.258	0
speak	0	0	1	0.4771	0	0	1	1	1	0.258	0
speaker	1	1	1	0.4771	0.4771	0.5773	0	0	0	0	0
summer	0	0	1	0.4771	0	0	1	1	1	0.258	0
sweet	0	0	1	0.4771	0	0	1	1	1	0.258	0
take	0	0	1	0.4771	0	0	1	1	1	0.258	0
voice	0	0	1	0.4771	0	0	1	1	1	0.258	0
whisper	1	1	1	0.4771	0.4771	0.5773	1	1	1	0.258	0.1489
										score	0.1489

Cos(Q2,D1)=0.1489

Cos(Q2,D2)=

		Q2							D2		
	tf-raw	tf-wt	DF	idf	TF IDF	nlize	Tf-raw	tf-wt	WT	nLIZE	pRODUCT
alexa	1	1	3	0	0	0	1	1	1	0.2038	0
allow	0	0	1	0.4771	0	0	1	1	1	0.2038	0
artificial	0	0	2	0.1761	0	0	1	1	1	0.2038	0
assistant	0	0	1	0.4771	0	0	1	1	1	0.2038	0
basic	0	0	1	0.4771	0	0	1	1	1	0.2038	0

conduct	0	0	1	0.4771	0	0	1	1	1	0.2038	0
conversation	1	1	1	0.4771	0.4771	0.577	1	1	1	0.2038	0.1176
fall	0	0	1	0.4771	0	0	2	1.301	1.301	0.2652	0
genie	0	0	1	0.4771	0	0	1	1	1	0.2038	0
google	0	0	1	0.4771	0	0	1	1	1	0.2038	0
hook	0	0	1	0.4771	0	0	1	1	1	0.2038	0
intelligence	0	0	1	0.4771	0	0	1	1	1	0.2038	0
internet	0	0	1	0.4771	0	0	1	1	1	0.2038	0
look	0	0	1	0.4771	0	0	1	1	1	0.2038	0
mean	0	0	1	0.4771	0	0	1	1	1	0.2038	0
possess	0	0	1	0.4771	0	0	1	1	1	0.2038	0
say	0	0	1	0.4771	0	0	1	1	1	0.2038	0
siri	0	0	1	0.4771	0	0	1	1	1	0.2038	0
smart	0	0	1	0.4771	0	0	2	1.301	1.301	0.2652	0
speaker	1	1	1	0.4771	0.4771	0.577	2	1.301	1.301	0.2652	0.1530
stuff	0	0	1	0.4771	0	0	1	1	1	0.2038	0
thing	0	0	1	0.4771	0	0	1	1	1	0.2038	0
whisper	1	1	1	0.4771	0.4771	0.577	0	0	0	0.0000	0
										sCORE =	0.27062311

Cos(Q2,D2)=0.270

Cos(Q2,D3)=

		Q2							D3		
	tf-raw	tf-wt	DF	idf	TF IDF	nlize	Tf-raw	tf-wt	WT	nLIZE	pRODUCT
alexa	1	1	3	0	0	0	1	1	1	0.236	0
already	0	0	1	0.4772	0	0	1	1	1	0.236	0
catch	0	0	1	0.4772	0	0	1	1	1	0.236	0
conversation	1	1	1	0.4772	0.4772	0.577	0	0	0	0	0
depressed	0	0	1	0.4772	0	0	1	1	1	0.236	0
dog	0	0	1	0.4772	0	0	1	1	1	0.236	0
figure	0	0	1	0.4772	0	0	1	1	1	0.236	0
new	0	0	1	0.4772	0	0	1	1	1	0.236	0
news	0	0	1	0.4772	0	0	1	1	1	0.236	0
perhaps	0	0	1	0.4772	0	0	1	1	1	0.236	0
screen	0	0	1	0.4772	0	0	1	1	1	0.236	0
shop	0	0	1	0.4772	0	0	0	0	0	0.236	0
sick	0	0	1	0.4772	0	0	1	1	1	0.236	0
simple	0	0	1	0.4772	0	0	1	1	1	0.236	0
speaker	1	1	1	0.4772	0.4772	0.577	0	0	0	0	0
talk	0	0	1	0.4772	0	0	1	1	1	0.236	0
thing	0	0	2	0.176	0	0	1	1	1	0.236	0
think	0	0	1	0.4772	0	0	1	1	1	0.236	0
try	0	0	1	0.4772	0	0	1	1	1	0.236	0
way	0	0	1	0.4772	0	0	1	1	1	0.236	0
whisper	1	1	1	0.4772	0.4772	0.577	0	0	0	0	0
										sCORE	0

Cos(q2,d3)=0

Cos(Q3,D1)=

		q3							D1		
	tf-raw	tf-wt	DF	idf	TF IDF	nlize	Tf-raw	tf-wt	WT	nLIZE	pRODUCT
alexa	1	1	3	0	0	0	1	1	1	0.2425	0
amazon	0	0	1	0.4771	0	0	1	1	1	0.2425	0
assistant	0	0	2	0.1761	0	0	1	1	1	0.2425	0
day	0	0	1	0.4771	0	0	1	1	1	0.2425	0
dot	0	0	1	0.4771	0	0	1	1	1	0.2425	0
dream	0	0	1	0.4771	0	0	1	1	1	0.2425	0
echo	0	0	1	0.4771	0	0	1	1	1	0.2425	0
end	0	0	1	0.4771	0	0	1	1	1	0.2425	0
intelligence	1	1	1	0.4771	0.4771	0.577	0	0	0	0	0
interaction	0	0	1	0.4771	0	0	1	1	1	0.2425	0
screen	0	0	1	0.4771	0	0	1	1	1	0.2425	0
simple	0	0	1	0.4771	0	0	1	1	1	0.2425	0
smart	1	1	1	0.4771	0.4771	0.577	0	0	0	0	0
speak	0	0	1	0.4771	0	0	1	1	1	0.2425	0

summer	0	0	1	0.4771	0	0	1	1	1	0.2425	0
sweet	0	0	1	0.4771	0	0	1	1	1	0.2425	0
take	0	0	1	0.4771	0	0	1	1	1	0.2425	0
talk	1	1	1	0.4771	0.4771	0.577	0	0	0	0	0
voice	0	0	1	0.4771	0	0	1	1	1	0.2425	0
whisper	0	0	1	0.4771	0	0	1	1	1	0.2425	0
										SCORE	0

Cos(Q3,D1)=0

Cos(Q3,D2)=

		q3							D2		
	tf-raw	tf-wt	DF	idf	TF IDF	nlize	Tf-raw	tf-wt	WT	nLIZE	pRODUCT
alexa	1	1	3	0	0	0	1	1	1	0.204	0
allow	0	0	1	0.4771	0	0	1	1	1	0.204	0
artificial	0	0	1	0.4771	0	0	1	1	1	0.204	0
assistant	0	0	2	0.1761	0	0	1	1	0	0.204	0
basic	0	0	1	0.4771	0	0	1	1	1	0.204	0
conduct	0	0	1	0.4771	0	0	1	1	1	0.204	0
conversation	0	0	1	0.4771	0	0	1	1	1	0.204	0
fall	0	0	1	0.4771	0	0	2	1.301	1.301	0.265	0
genie	0	0	1	0.4771	0	0	1	1	1	0.204	0
google	0	0	1	0.4771	0	0	1	1	1	0.204	0
hook	0	0	1	0.4771	0	0	1	1	1	0.204	0
intelligence	1	1	1	0.4771	0.4771	0.577	1	0	0	0.204	0.118
internet	0	0	1	0.4771	0	0	1	1	1	0.204	0
look	0	0	1	0.4771	0	0	1	1	1	0.204	0
mean	0	0	1	0.4771	0	0	0	0	0	0.204	0
possess	0	0	1	0.4771	0	0	1	1	1	0.204	0
say	0	0	1	0.4771	0	0	1	1	1	0.204	0
siri	0	0	1	0.4771	0	0	1	1	1	0.204	0
smart	1	1	1	0.4771	0.4771	0.577	2	1.301	1.301	0.265	0.153
speaker	0	0	1	0.4771	0	0	2	1.301	1.301	0.265	0
stuff	0	0	1	0.4771	0	0	1	1	1	0.204	0
talk	1	1	1	0.4771	0.4771	0.577	0	0	0	0	0
thing	0	0	2	0.1761	0	0	1	1	1	0.204	
										Score	0.271

Cos(Q3,D2)=0.271

Cos(Q3,D3)=

		q3							d3		
	tf-raw	tf-wt	DF	idf	TF IDF	nlize	Tf-raw	tf-wt	WT	nLIZE	pRODUCT
alexa	1	1	3	0	0	0	1	1	1	0.2357	0
already	0	0	1	0.4771	0	0	1	1	1	0.2357	0
catch	0	0	1	0.4771	0	0	1	1	1	0.2357	0
depressed	0	0	1	0.4771	0	0	1	1	1	0.2357	0
dog	0	0	1	0.4771	0	0	1	1	1	0.2357	0
figure	0	0	1	0.4771	0	0	1	1	1	0.2357	0
intelligence	1	1	1	0.4771	0.4771	0.577	0	0	0	0	0
new	0	0	1	0.4771	0	0	1	1	1	0.2357	0
news	0	0	1	0.4771	0	0	1	1	1	0.2357	0
perhaps	0	0	1	0.4771	0	0	1	1	1	0.2357	0
screen	0	0	1	0.4771	0	0	1	1	1	0.2357	0
shop	0	0	1	0.4771	0	0	1	1	1	0.2357	0
sick	0	0	1	0.4771	0	0	1	1	1	0.2357	0
simple	0	0	1	0.4771	0	0	1	1	1	0.2357	0
smart	1	1	1	0.4771	0.4771	0.577	0	0	0	0	0
talk	1	1	1	0.4771	0.4771	0.577	1	1	1	0.2357	0.1359
thing	0	0	2	0.1761	0	0	1	1	1	0.2357	0
think	0	0	1	0.4771	0	0	1	1	1	0.2357	0
try	0	0	1	0.4771	0	0	1	1	1	0.2357	0
way	0	0	1	0.4771	0	0	1	1	1	0.2357	0
										Score	0.1359989

Cos(Q3,D3)=0.1359989

Cos(Q4,D1)=

				Q4				D1			
	tf-raw	tf-wt	DF	idf	TF IDF	nlize	Tf-raw	tf-wt	WT	nLIZE	pRODUCT
alexa	0	0	3	0	0	0	1	1	1	0.258	0
amazon	0	0	1	0.4772	0	0	1	1	1	0.258	0
artificial	1	1	1	0.4772	0.4772	0.5	0	0	0	0	0
assistant	0	0	2	0.176	0	0	1	1	1	0.258	0
day	0	0	1	0.4772	0	0	1	1	1	0.258	0
dot	0	0	1	0.4772	0	0	1	1	1	0.258	0
dream	0	0	1	0.4772	0	0	1	1	1	0.258	0
echo	0	0	1	0.4772	0	0	1	1	1	0.258	0
end	0	0	1	0.4772	0	0	1	1	1	0.258	0
fall	1	1	1	0.4772	0.4772	0.5	0	0	0	0	0
interaction	0	0	1	0.4772	0	0	1	1	1	0.258	0
screen	1	1	1	0.4772	0.4772	0.5	0	0	0	0	0
speaker	0	0	1	0.4772	0	0	1	1	1	0.258	0
summer	0	0	1	0.4772	0	0	1	1	1	0.258	0
sweet	0	0	1	0.4772	0	0	1	1	1	0.258	0
take	0	0	1	0.4772	0	0	1	1	1	0.258	0
think	1	1	1	0.4772	0.4772	0.5	0	0	0	0	0
voice	0	0	1	0.4772	0	0	1	1	1	0.258	0
whisper	0	0	1	0.4772	0	0	1	1	1	0.258	0
										score	0

Cos(Q4,D1)=0

Cos(Q4,D2)=

		q4							d2		
	tf-raw	tf-wt	DF	idf	TF IDF	nlize	Tf-raw	tf-wt	WT	nLIZE	pRODUCT
alexa	0	0	3	0	0	0	1	1	1	0.2038	0
allow	0	0	1	0.4771	0	0	1	1	1	0.2038	0
artificial	1	1	1	0.4771	0.4771	0.499	1	1	1	0.2038	0.1017
assistant	0	0	1	0.4771	0	0	1	1	1	0.2038	0

basic	0	0	1	0.4771	0	0	1	1	1	0.2038	0
conduct	0	0	1	0.4771	0	0	1	1	1	0.2038	0
conversation	0	0	1	0.4771	0	0	1	1	1	0.2038	0
fall	1	1	1	0.4771	0.4771	0.499	2	1.3010	1.3010	0.2652	0.1323
genie	0	0	1	0.4771	0	0	1	1	1	0.2038	0
google	0	0	1	0.4771	0	0	1	1	1	0.2038	0
hook	0	0	1	0.4771	0	0	1	1	1	0.2038	0
intelligence	0	0	1	0.4771	0	0	1	1	1	0.2038	0
internet	0	0	1	0.4771	0	0	1	1	1	0.2038	0
look	0	0	1	0.4771	0	0	1	1	1	0.2038	0
mean	0	0	1	0.4771	0	0	1	1	1	0.2038	0
possess	0	0	1	0.4771	0	0	1	1	1	0.2038	0
say	0	0	1	0.4771	0	0	1	1	1	0.2038	0
screen	1	1	1	0.4771	0.4771	0.499	0	0	0	0.0000	0
siri	0	0	1	0.4771	0	0	1	1	1	0.2038	0
smart	0	0	1	0.4771	0	0	2	1.3010	1.3010	0.2652	0
speaker	0	0	1	0.4771	0	0	2	1.3010	1.3010	0.2652	0
stuff	0	0	1	0.4771	0	0	1	1	1	0.2038	0
thing	0	0	2	0.1761	0	0	1	1	1	0.2038	0
think	1	1	1	0.4771	0.4771	0.499	0	0	0	0.0000	0
										scORE	0.2340

Cos(Q4,D2)=0.2340

Cos(Q4,D3)=

		q4							d2		
	tf-raw	tf-wt	DF	idf	TF IDF	nlize	Tf-raw	tf-wt	WT	nLIZE	pRODUCT
alexa	0	0	3	0	0	0	1	1	1	0.236	0
already	0	0	1	0.4771	0	0	1	1	1	0.236	0
artificial	1	1	1	0.4771	0.4771	0.5	0	0	0	0	0
catch	0	0	1	0.4771	0	0	1	1	1	0.236	0
depress	0	0	1	0.4771	0	0	1	1	1	0.236	0
dog	0	0	1	0.4771	0	0	1	1	1	0.236	0
fall	1	1	1	0.4771	0.4771	0.5	0	0	0	0	0
figure	0	0	1	0.4771	0	0	1	1	1	0.236	0
new	0	0	1	0.4771	0	0	1	1	1	0.236	0
news	0	0	1	0.4771	0	0	1	1	1	0.236	0
perhaps	0	0	1	0.4771	0	0	1	1	1	0.236	0
screen	1	1	1	0.4771	0.4771	0.5	1	1	1	0.236	0.118
shop	0	0	1	0.4771	0	0	1	1	1	0.236	0
sick	0	0	1	0.4771	0	0	1	1	1	0.236	0
simple	0	0	1	0.4771	0	0	1	1	1	0.236	0
talk	0	0	1	0.4771	0	0	1	1	1	0.236	0
thing	0	0	2	0.1761	0.1761	0	1	1	1	0.236	0
think	1	1	1	0.4771	0.4771	0.5	1	1	1	0.236	0.118
try	0	0	1	0.4771	0	0	1	1	1	0.236	0
way	0	0	1	0.4771	0	0	1	1	1	0.236	0
										Score:	0.236

Cos(Q4,D3)=0.236

SrNo	Q1	Q2	Q3	Q4

1.	alexa	alexa	alexa	fall
2.	simple	whisper	smart	think
3.	screen	speaker	intelligence	artificial
4.		conversation	talk	screen

D1	D2	D3
1. alexa	alexa	alexa
2. amazon	allow	already
3. assistant	artificial	catch
4. day	assistant	depressed
5. dot	basic	dog
6. dream	conduct	figure
7. echo	conversation	new
8. end	fall	news
9. interaction	genie	perhaps
10. speak	google	screen
11. summer	hook	shop
12. sweet	intelligence	sick
13. take	internet	simple
14. voice	look	talk
15. whisper	mean	thing
16.	possess	think
17.	say	try
18.	siri	way
19.	smart	
20.	speaker	
21.	stuff	
22.	thing	

(ii) Jaccard Coefficients:

JC(Q1,D1)=

n(Q1 ∩ D1) / n(Q1 U D1) = 1 / 17 = **0.059**

JC(Q1,D2)=

n(Q1 ∩ D2) / n(Q1 U D2) = 1 / 24 = **0.042**

JC(Q1,D3)=

n(Q1 ∩ D3) / n(Q1 U D3) = 3 / 18 = **0.167**

JC(Q2,D1)=

n(Q2 ∩ D1) / n(Q2 U D1) = 2 / 17 = **0.118**

JC(Q2,D2)=

n(Q2 ∩ D2) / n(Q2 U D2) = 3 / 23 = **0.130**

JC(Q2,D3)=

n(Q2 ∩ D3) / n(Q2 U D3) = 1 / 21 = **0.048**

JC(Q3,D1)=

n(Q3 ∩ D1) / n(Q3 U D1) = 1 / 18 = **0.055**

JC(Q3,D2)=

n(Q3 ∩ D2) / n(Q3 U D2) = 3 / 23 = **0.130**

JC(Q3,D3)=

(Q3 ∩ D3) / n(Q3 U D3) = 2 / 20 = **0.1**

JC(Q4,D1)=

n(Q4 ∩ D1) / n(Q4 U D1) = 0 / 19 = **0**

JC(Q4,D2)=

n(Q4 ∩ D2) / n(Q4 U D2) = 2 / 24 = **0.083**

JC(Q4,D3)=

n(Q4 ∩ D3) / n(Q4 U D3) = 2 / 20 = **0.1**

Problem 2 (34 points):
Suppose you have a collection of 5 documents, and only 10 terms are used in them:

	Term1	Term2	Term3	Term4	Term5	Term6	Term7	Term8	Term9	Term10
DOC1	0	3	5	4	0	5	0	0	4	2
DOC2	3	0	1	4	3	0	0	5	1	6
DOC3	5	0	5	1	2	0	2	5	0	7
DOC4	1	8	0	2	0	1	6	0	2	1
DOC5	2	7	0	0	0	3	0	2	3	0

List the values of the gaps for the last four terms in your index computed for this collection. Encode these gaps with (i) unary codes (**8 points**); (ii) Gamma codes (**16 points**); and (iii) 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.I:
POSTING FILE:
Term7->3,4
Term 8-> 2,3,5
Term 9->1,2,4,5
Term 10->1,2,3,4

Gaps for Term 1:

Gaps=3,1

Unary:

Docid	Gaps	Unary
3	3	1110
4	1	10

Gaps for Term 2:

Term8 is present in Doc2,Doc3 and Doc5

Gaps=2,3,5

Docid	Gaps	unARY
2	2	110
3	1	10
5	2	110

Gaps for Term 3:

Term 9 is present in Doc1,Doc2,Doc4 and Doc5

Gaps: 1,1,2,1

DociD	GAP	Unary
1	1	10
2	1	10
4	2	110
5	1	10

Gaps for Term 4:

Term 10 is present in Doc1,Doc2,Doc3 and Doc4
Gaps:1,1,1,1

DociD	GAP	Unary
1	1	10
2	1	10
3	1	10
4	1	10

SOLUTION 2.II:

Gamma codes for Term 1:

Term 7:

Gaps=3,1

DociD	GAP	bINARY	lENGTH(OFFSET)	Unary offset	Gamma
3	3	11	1	10	101
1	1	1	0	0	00

Gamma codes for Term 2:

DociD	GAP	bINARY	lENGTH(OFFSET)	Unary offset	Gamma
2	2	10	1	10	100
3	1	1	0	0	0
5	2	10	1	10	100

Gamma codes for Term 3:

DociD	GAP	bINARY	lENGTH(OFFSET)	Unary offset	Gamma
1	1	1	0	0	0
2	1	1	0	0	0
4	2	10	1	10	100
5	1	1	0	0	0

Gamma codes for Term 4:

DociD	GAP	bINARY	lENGTH(OFFSET)	Unary offset	Gamma
1	1	1	0	0	0
2	1	1	0	0	0
3	1	1	0	0	0
4	1	1	0	0	0

SOLUTION 2.III:

Delta codes for Term 1:

DociD	GAP	bINARY	OFFSET	lENGTH(binary gap)	Gamma of length	Delta
3	3	11	1	2	100	1001
4	1	1	NULL	1	0	0

Delta codes for Term 2:

DociD	GAP	bINARY	OFFSET	lENGTH(binary gap)	Gamma of length	Delta
2	2	10	0	2	100	1000
3	1	1	NULL	1	0	0
5	2	10	0	2	100	1000

Delta codes for Term 3:

DociD	GAP	bINARY	OFFSET	lENGTH(binary gap)	Gamma of length	Delta
1	1	1	NULL	1	0	0
2	1	1	NULL	1	0	0
4	2	10	0	2	100	1000
5	1	1	NULL	1	0	0

Delta codes for Term 4:

DociD	GAP	bINARY	OFFSET	lENGTH(binary gap)	Gamma of length	Delta
1	1	1	NULL	1	0	0
2	1	1	NULL	1	0	0
3	1	1	NULL	1	0	0
4	1	1	NULL	1	0	0