Homework #5

Assign date: 2015-08-02

Due date: 2015-08-10, 6pm

Submission:

1. **Please submit your results in email to the grader: 130301039@svuca.edu Chi Zhang**
2. **Please separate the written answers from the python code: you should submit 1 file in your email – cs596-29-hw5_yourID#.doc**
3. **30 pts per day will be deducted for late submission**

Problem 1) Boolean Model (10 pts.)

We are given 6 documents with labels "ham" and "spam" as below. Answer the following questions about Boolean model text retrieval:

a) If the query is "great OR free NOT hurry", which documents will be retrieved?
Answer: D1, D2, D3, D4

b) If the query is "vocation OR food AND experience", which documents will be retrieved?
Answer: D1, D2, D4, D6

ID	Document	Class
1	Enjoy great food experience	Ham
2	Enjoy free vocation experience	Ham
3	Congratulation great vocation reward	Ham
4	Reward great vocation experience	Ham
5	Congratulation great reward hurry	Spam
6	Experience free vocation hurry	Spam

Problem 2) Naïve Bayes spam filter (20 pts.)

We are building a Naïve Bayes classifier based spam filter using the same training data as in Problem-1. Please answer the following questions.

a) First build the vocabulary set from the training data.
Answer:
V = {enjoy, great, food, experience, free, vocation, congratulation, reward, hurry}

b) Build "bag of words" representation for the training set. Please sort the words in alphabetical order from 'a' to 'z'.
Answer:
Sorted V = {congratulation, enjoy, experience, food, free, great, hurry, reward, vocation}

ID	Class	congratulation	enjoy	experience	food	free	great	hurry	reward	vocation
1	Ham	0	1	1	1	0	1	0	0	0
2	Ham	0	1	1	0	1	0	0	0	1

3	**Ham**	1	0	0	0	0	1	0	1	1
4	**Ham**	0	0	1	0	0	1	0	1	1
5	**Spam**	1	0	0	0	0	1	1	1	0
6	**Spam**	0	0	1	0	1	0	1	0	1
Σ HAM		**1**	**2**	**3**	**1**	**1**	**3**	**0**	**2**	**3**
Σ SPAM		**1**	**0**	**1**	**0**	**1**	**1**	**2**	**1**	**1**
Σ ALL		**2**	**2**	**4**	**1**	**2**	**4**	**2**	**3**	**4**

c) Apply the NB classifier to decide if the following message is spam: "free reward hurry"

Answer:

Two classes: Ham, Spam

N = 6

P(Ham) = 4/6; P(Spam) = 2/6

|V| = 9

ID	Document	Class	Σ of words
1	Enjoy great food experience	Ham	4
2	Enjoy free vocation experience	Ham	4
3	Congratulation great vocation reward	Ham	4
4	Reward great vocation experience	Ham	4
5	Congratulation great reward hurry	Spam	4
6	Experience free vocation hurry	Spam	4

n-ham = 4+4+4+4 = 16

n-spam = 4+4 = 8

P(free|Ham) = $\frac{1+1}{16+9} = \frac{2}{25}$; P(free|Spam) = $\frac{1+1}{8+9} = \frac{2}{17}$

P(reward|Ham) = $\frac{2+1}{16+9} = \frac{3}{25}$; P(reward|Spam) = $\frac{1+1}{8+9} = \frac{2}{17}$

P(hurry|Ham) = $\frac{0+1}{16+9} = \frac{1}{25}$; P(hurry|Spam) = $\frac{2+1}{8+9} = \frac{3}{17}$

P(ham|"free reward hurry") = P(ham).P(free|ham).P(reward|ham).P(hurry|ham)

P(ham|"free reward hurry") = $\frac{4}{6} * \frac{2}{25} * \frac{3}{25} * \frac{1}{25} = \frac{4}{15625} = \mathbf{0.000256}$

P(spam|"free reward hurry") = P(spam).P(free|spam).P(reward|spam).P(hurry|spam)

P(spam|"free reward hurry") = $\frac{2}{6} * \frac{2}{17} * \frac{2}{17} * \frac{3}{17} = \frac{4}{4913} = \mathbf{0.000814}$

P(ham | "free reward hurry") < P(spam | "free reward hurry")

⇨ **"free reward hurry" is more likely a spam message.**

d) Decide if the following message is spam: "great food reward"

Answer:

P(great|Ham) = $\frac{3+1}{16+9} = \frac{4}{25}$; P(great|Spam) = $\frac{1+1}{8+9} = \frac{2}{17}$

P(food|Ham) = $\frac{1+1}{16+9} = \frac{2}{25}$; P(food|Spam) = $\frac{0+1}{8+9} = \frac{1}{17}$

P(reward|Ham) = $\frac{2+1}{16+9} = \frac{3}{25}$; P(reward|Spam) = $\frac{1+1}{8+9} = \frac{2}{17}$

P(ham|"great food reward") = $\frac{4}{6} * \frac{4}{25} * \frac{2}{25} * \frac{3}{25} = \frac{16}{15625} = \mathbf{0.001024}$

P(spam|"great food reward") = $\frac{2}{6} * \frac{2}{17} * \frac{1}{17} * \frac{2}{17} = \frac{14}{14739} = \mathbf{0.00027}$

P(ham | "great food reward") > P(spam | "great food reward")

⇨ **"great food reward" should not be a spam message.**

Problem 3) Cosine similarity (16 pts.)

Based on the same training data as in Problem-1, answer the following questions: (Please sort the features in alphabetical order from 'a' to 'z')

a) Build the vector representation for the class "Ham"

Answer:

Vectors representation for class "Ham":

	DOC-1	DOC-2	DOC-3	DOC-4
congratulation	**0**	**0**	**1**	**0**
enjoy	**1**	**1**	**0**	**0**
experience	**1**	**1**	**0**	**1**
food	**1**	**0**	**0**	**0**
free	**0**	**1**	**0**	**0**
great	**1**	**0**	**1**	**1**
hurry	**0**	**0**	**0**	**0**
reward	**0**	**0**	**1**	**1**
vocation	**0**	**1**	**1**	**1**

Summation of all vectors in class "Ham" produce a resultant vector representing class "Ham":

$$\vec{v}_{ham} = \left[\vec{D}_1, \vec{D}_2, \vec{D}_3, \vec{D}_4\right] \rightarrow \sum \vec{v}_{ham} = (1,2,3,1,1,3,0,2,3)$$

b) Build the vector representation for the class "Spam"

Answer:

Vectors representation for class "Spam":

	DOC-5	DOC-6
congratulation	**1**	**0**
enjoy	**0**	**0**
experience	**0**	**1**
food	**0**	**0**
free	**0**	**1**
great	**1**	**0**
hurry	**1**	**1**
reward	**1**	**0**
vocation	**0**	**1**

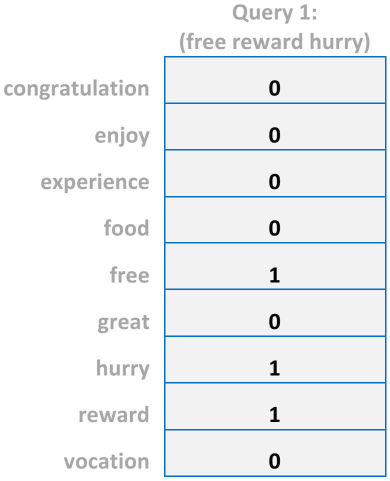
Summation of all vectors in class "Spam" produce a resultant vector representing class "Spam":

$\vec{v}_{spam} = [\vec{D}_5, \vec{D}_6] \rightarrow \sum \vec{v}_{spam} = (1,0,1,0,1,1,2,1,1)$

c) Build the vector representation for the test document: "free reward hurry", and calculate its Cosine similarities to both the ham and spam classes.

Answer:

Vector representation of Q1 = "free reward hurry":



	Query 1: (free reward hurry)
congratulation	0
enjoy	0
experience	0
food	0
free	1
great	0
hurry	1
reward	1
vocation	0

q1 = (0,0,0,0,1,0,1,1,0)

Using Cosine Similarity Formula

$$cosine(d_i, q) = \frac{d \cdot q}{\|d_j\| * \|q\|}$$

cosine(ham, q1) = 0.281

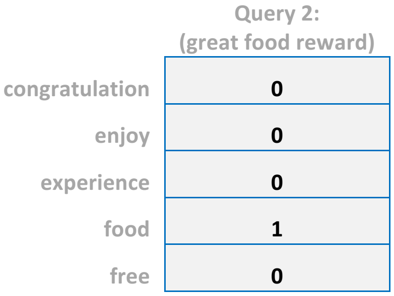
cosine(spam, q1) = 0.730

Since cosine(spam,q1) > cosine(ham,q1), therefore Q1="free reward hurry" is more likely a spam message.

d) Build the vector representation for the test document: "great food reward", and calculate its Cosine similarities to both the ham and spam classes.

Answer:

Vector representation of Q2 = "great food reward":



	Query 2: (great food reward)
congratulation	0
enjoy	0
experience	0
food	1
free	0

great	**1**
hurry	**0**
reward	**1**
vocation	**0**

q2 = (0,0,0,1,0,1,0,1,0)

Using Cosine Similarity Formula

$$cosine(d_i, q) = \frac{d \cdot q}{\|d_j\| * \|q\|}$$

cosine(ham, q2) = 0.562
cosine(spam, q2) = 0.365

Since cosine(ham,q2) > cosine(spam,q2), therefore Q2="great food reward" is more likely a ham message.

Problem 4) Tf-Idf transform (24 pts.)
Based on the same training set as in Problem-1, construct the Tf-Idf transform either manually or using Sklearn. Answer the following questions: (Please sort the features in alphabetical order from 'a' to 'z')

a) What is the bag of words representation of the 6 training documents after Tf-Idf transform?

Answer:
We will re-use the document term matrix in previous Problem 2:

DOC-ID	congratulation	enjoy	experience	food	free	great	hurry	reward	vocation
1	0	1	1	1	0	1	0	0	0
2	0	1	1	0	1	0	0	0	1
3	1	0	0	0	0	1	0	1	1
4	0	0	1	0	0	1	0	1	1
5	1	0	0	0	0	1	1	1	0
6	0	0	1	0	1	0	1	0	1

Calculate TF table:
Since the maximum frequency of each term in each document is 1 (i.e. no duplicated word in each document), we can re-use the above table as the term frequency table.

Calculate IDF table:

$$idf_i = \log_2 \frac{Total\_no\_of\_documents}{Number\_of\_documents\_containing\_term\_i}$$

The total number of document N = 6. Thefore the idf values for all terms are:

congratulation	= log2(6/2)	1.5850
enjoy	= log2(6/2)	1.5850
experience	= log2(6/4)	0.5850
food	= log2(6/1)	2.5850
free	= log2(6/2)	1.5850
great	= log2(6/4)	0.5850
hurry	= log2(6/2)	1.5850
reward	= log2(6/3)	1.0000
vocation	= log2(6/4)	0.5850

Multiply tf and the idf score of each term, we have the TF-IDF matrix below:

	congratulation	**enjoy**	**experience**	**food**	**free**	**great**	**hurry**	**reward**	**vocation**
DOC-1	0.000	1.585	0.585	2.585	0.000	0.585	0.000	0.000	0.000
DOC-2	0.000	1.585	0.585	0.000	1.585	0.000	0.000	0.000	0.585
DOC-3	1.585	0.000	0.000	0.000	0.000	0.585	0.000	1.000	0.585
DOC-4	0.000	0.000	0.585	0.000	0.000	0.585	0.000	1.000	0.585
DOC-5	1.585	0.000	0.000	0.000	0.000	0.585	1.585	1.000	0.000
DOC-6	0.000	0.000	0.585	0.000	1.585	0.000	1.585	0.000	0.585

b) Calculate Cosine similarities of the test document "free reward hurry" to the two classes "Ham" and "Spam" using Tf-Idf weights.

Answer:

Vectors representation of class "Ham" and "Spam" based on the summation of {D1,D2,D3,D4} and {D5,D6} on TF-IDF matrix:

	HAM	SPAM
CONGRATULATION	1.585	1.585
ENJOY	3.170	0.000
EXPERIENCE	1.755	0.585
FOOD	2.585	0.000
FREE	1.585	1.585
GREAT	1.755	0.585
HURRY	0.000	3.170
REWARD	2.000	1.000
VOCATION	1.755	0.585

c-ham$_{\text{TF-IDF-ed}}$ = (1.585, 3.170, 1.755, 2.585, 1.585, 1.755, 0, 2, 1.755)
c-spam$_{\text{TF-IDF-ed}}$ = (1.585, 0, 0.585, 0, 1.585, 0.585, 3.170, 1, 0.585)

Q1 = "free reward hurry", q1 = (0,0,0,0,1,0,1,1,0)

cosine(ham, q1) = 0.350
cosine(spam, q1) = 0.804

Since cosine(spam,q1) > cosine(ham,q1), therefore Q1="free reward hurry" is more likely a spam message.

c) Calculate Cosine similarities of the test document "great food reward" to the two classes "Ham" and "Spam" using Tf-Idf weights

Answer:

c-ham$_{\text{TF-IDF-ed}}$ = (1.585, 3.170, 1.755, 2.585, 1.585, 1.755, 0, 2, 1.755)
c-spam$_{\text{TF-IDF-ed}}$ = (1.585, 0, 0.585, 0, 1.585, 0.585, 3.170, 1, 0.585)

q2 = (0,0,0,1,0,1,0,1,0)

cosine(ham, q2) = 0.619
cosine(spam, q2) = 0.221

Since cosine(ham,q2) > cosine(spam,q2), therefore Q2="great food reward" is more likely a ham message.

Problem 5) Rocchio Text Classifier (30 pts.)

Using the training data in Problem-1, find the centroids for the "spam" and "ham" classes of the Rocchio Text Classifier as discussed in the class. Recall that the Rocchio classifier computes the centroid C_i for each class i from relevant and irrelevant documents as follows:

$$\mathbf{c}_i = \frac{\alpha}{|D_i|}\sum_{\mathbf{d}\in D_i}\frac{\mathbf{d}}{\|\mathbf{d}\|} - \frac{\beta}{|D-D_i|}\sum_{\mathbf{d}\in D-D_i}\frac{\mathbf{d}}{\|\mathbf{d}\|}$$

For the current problem, make the following assumptions:

(a) Tf-IDF weight is NOT used.
(b) The weights $\alpha = 1$ and $\beta = 0.5$.

Answer:

Term frequency matrix

1	2	3	4	5	6
Ham	**Ham**	**Ham**	**Ham**	**Spam**	**Spam**
0	0	1	0	1	0
1	1	0	0	0	0
1	1	0	1	0	1
1	0	0	0	0	0
0	1	0	0	0	1

1	0	1	1	1	0
0	0	0	0	1	1
0	0	1	1	1	0
0	1	1	1	0	1

$$\vec{v}_{ham} = \left[\vec{D}_1, \vec{D}_2, \vec{D}_3, \vec{D}_4\right] \rightarrow \sum \vec{v}_{ham} = (1,2,3,1,1,3,0,2,3)$$

$$\vec{v}_{spam} = \left[\vec{D}_5, \vec{D}_6\right] \rightarrow \sum \vec{v}_{spam} = (1,0,1,0,1,1,2,1,1)$$

$$\vec{c}_{ham} = \frac{1}{4} * \frac{(1,2,3,1,1,3,0,2,3)}{2} - \frac{0.5}{2} * \frac{(1,0,1,0,1,1,2,1,1)}{2}$$
$$= (0., 0.25, 0.25, 0.125, 0., 0.25, -0.25, 0.125, 0.25)$$

$$\vec{c}_{spam} = \frac{1}{2} * \frac{(1,0,1,0,1,1,2,1,1)}{2} - \frac{0.5}{4} * \frac{(1,2,3,1,1,3,0,2,3)}{2}$$
$$= (0.1875, -0.125, 0.0625, -0.0625, 0.1875, 0.0625, 0.5, 0.125, 0.0625)$$