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?
b) If the query is "vocation OR food AND experience", which documents will be retrieved?

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

Answers:

a) 1, 2, 3, 4
b) 1, 2, 6

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.
b) Build "bag of words" representation for the training set. Please sort the words in alphabetical order from 'a' to 'z'.
c) Apply the NB classifier to decide if the following message is spam: "free reward hurry"
d) Decide if the following message is spam: "great food reward"

Answers:

a) Vocabulary

[Enjoy great food experience free vocation Congratulation reward hurry]

b) Bag of words

	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

P(ham) = 2/3, P(spam) = 1/3,

P(free | ham) = (1+1)/(16+9), P(free | spam) = (1+1)/(8+9)

P(reward | ham) = (2+1)/(16+9), P(reward | spam) = (1+1)/(8+9)

P(hurry | ham) = (0+1)/(16+9), P(hurry | spam) = (2+1)/(8+9)

P(great | ham) = (3+1)/(16+9), P(great | spam) = (1+1)/(8+9)

P(food | ham) = (1+1)/(16+9), P(food | spam) = (0+1)/(8+9)

c) P(ham | free reward hurry) = 2/3 * 2/25 * 3/25 * 1/25 = 4/(25*25*25)
P(spam | free reward hurry) = 1/3 * 2/17 * 2/17 * 3/17 = 4/(17*17*17)
"free reward hurry" is a spam

d) P(ham | great food reward) = 2/3 * 4/25 * 2/25 * 3/25 = 16/(25*25*25) = 0.001
P(spam | great food reward) = 1/3 * 2/17 * 1/17 * 2/17 = 4/(3*17*17*17) = 0.00027
"free reward hurry" is not a spam

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

Answers:

a) Vector representation for ham class: [1 2 3 1 1 3 0 2 3]

b) Vector representation for spam class: [1 0 1 0 1 1 2 1 1]
c) Vector representation for "free reward hurry": [0 0 0 0 1 0 1 1 0]
    a. Cosine similarity with ham = 0.281
    b. Cosine similarity with spam = 0.73
d) Vector representation for "free reward hurry": [0 0 0 0 1 0 1 1 0]
    a. Cosine similarity with ham = 0.562
    b. Cosine similarity with spam = 0.365

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

Answers:

TF-IDF for ham+spam:

[[0.1527 0.4293 0.4582 0.2146 0.1527 0.4582 0. 0.3054 0.4582]

[0.2682 0. 0.2682 0. 0.2682 0.2682 0.7539 0.2682 0.2682]]

a) TF-IDF for corpus:

[[0. 0.5762 0.4099 0.5762 0. 0.4099 0. 0. 0.]

[0. 0.6301 0.4483 0. 0.4483 0. 0. 0. 0.4483]

[0.5 0. 0. 0. 0. 0.5 0. 0.5 0.5]

[0. 0. 0.5 0. 0. 0.5 0. 0.5 0.5]

[0.4483 0. 0. 0. 0. 0.4483 0.6301 0.4483 0.]

[0. 0. 0.4483 0. 0.4483 0. 0.6301 0. 0.4483]]

b) Cosine similarity for 'free reward hurry'
    a. Ham = 0.2298
    b. Spam = 0.8005
c) Cosine similarity for 'experience great reward'
    a. Ham = 0. 7054
    b. Spam = 0. 4646

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 $\beta = 0.5$.

Answers:

hamMean= [0.125 0.25 0.375 0.125 0.125 0.375 0. 0.25 0.375]

spamMean= [0.25 0. 0.25 0. 0.25 0.25 0.5 0.25 0.25]

hamRocchio= [0. 0.25 0.25 0.125 0. 0.25 -0.25 0.125 0.25]

spamRocchio= [0.1875 -0.125 0.0625 -0.0625 0.1875 0.0625 0.5 0.125 0.0625]