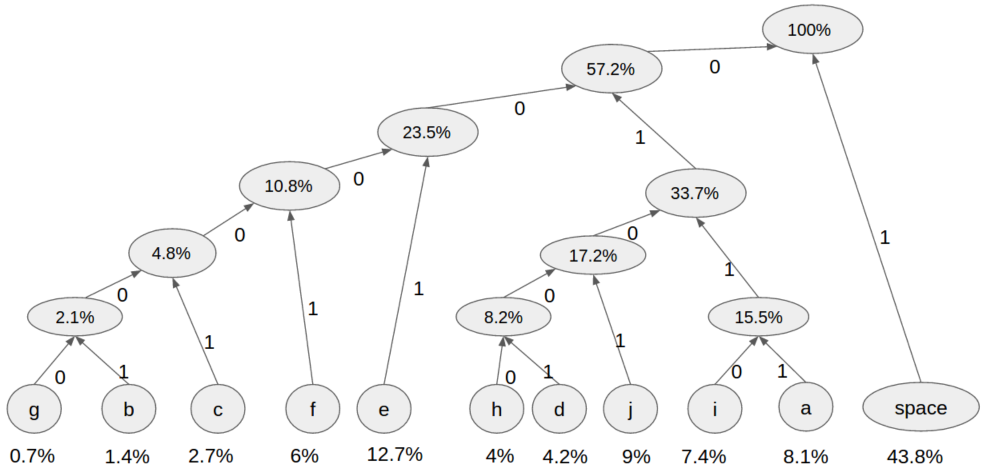
91.542 - Natural Language Processing
Homework Assignmen 3

Chang Liu
chang_liu@student.uml.edu

October 23, 2015

1 Problem 1. [20 pts] Huffman code

Answer:
(1) The building process of the huffman code is as follows:



The basic algorithm is to sort the nodes from low to high, and merge each two elements with the lowest frequency into one node, and then do it iteratively to connect all the nodes. After that, mark the left route as '0' and right path as '1'. At last, the huffman code for each symbol is represented by visiting the nodes and listing all the numbers in the path.

(2) The word "headi" is consist of five different symbols, which is 'h', 'e', 'a', 'd', 'i', we just need to get its code correspondingly and then concanate them, which is '01000', '001', '0111', '01001', '0110'. So the overall representation is

‘01000001011101001011 0’.

(3) First we can build a table that represents all the symbols, representations and its length, as follows:

Symbol	Representation	Length	Probability
a	0111	4	8.1%
b	000001	6	1.4%
c	00001	5	2.7%
d	01001	5	4.2%
e	001	3	12.7%
f	0001	4	6%
g	000000	6	0.7%
h	01000	5	4%
i	0110	4	7.4%
j	0101	4	9%
space	1	1	43.8%

So the average value is:

$$\begin{aligned}average &= 4 * 8.1\% + 6 * 1.4\% + 5 * 2.7\% + 5 * 4.2\% + 3 * 12.7\% + 4 * 6\% \\ &+ 6 * 0.7\% + 5 * 4\% + 4 * 7.4\% + 4 * 9\% + 1 * 43.8\% \\ &\approx 2.71\end{aligned}$$

(4) For entropy, use the following equation and the probability in the table from the question description, we can get the value:

$$\begin{aligned}H(p) &= H(X) \\ &= -\sum_{x \in X} p(x) \log_2 p(x) \\ &= -(8.1\% * \log_2 8.1\% + 1.4\% * \log_2 1.4\% + ... + 43.8\% * \log_2 43.8\%) \\ &\approx 2.68\end{aligned}$$

2 Problem 2. [30 pts] Evaluating a clustering solution

Answer:

(1) According to following equation, we can know the entropy of a clustering solution:

$$Entropy(C, S) = \sum_i \frac{|c_i|}{n} \sum_j \frac{|c_i \bigcap s_j|}{|c_i|} \log \frac{|c_i \bigcap s_j|}{|c_i|}$$

In the equation, c_i is a cluster from C, and s_j is a cluster from solution S, so we can get the entropy as follows:

$$\begin{aligned}
entropy &= \sum_{i\in C}\frac{|c_i|}{n}\sum_{j\in S}\frac{|c_i\bigcap s_j|}{|c_i|}\log\frac{|c_i\bigcap s_j|}{|c_i|}\\
&= \frac{1}{10}*[(\frac{2}{4}*\log_2\frac{2}{4}*2+0+0)*4+(0+\frac{1}{3}*\log_2\frac{1}{3}+\frac{2}{3}*\log_2\frac{2}{3}+0)*3\\
&+(0+\frac{1}{2}*\log_2\frac{1}{2}*2+0)*2+(\frac{1}{1}*\log_2\frac{1}{1}+0+0+0)*1]\\
&\approx -0.88
\end{aligned}$$

Regarding the entropy, we have to get the negative value of the above equation, which should be **0.88**

Similarly, we can get the second clustering solutions, as follows:

$$\begin{aligned}
entropy &= \sum_{i\in C}\frac{|c_i|}{n}\sum_{j\in S}\frac{|c_i\bigcap s_j|}{|c_i|}\log\frac{|c_i\bigcap s_j|}{|c_i|}\\
&= \frac{1}{10}*[4*(\frac{3}{4}*\log_2\frac{3}{4}+\frac{1}{4}*\log_2\frac{1}{4})+3*(\frac{2}{3}*\log_2\frac{2}{3}+\frac{1}{3}*\log_2\frac{1}{3})\\
&+3*(\frac{1}{3}*\log_2\frac{1}{3}*3)]\\
&\approx -1.08
\end{aligned}$$

So for the second clustering solution, the entropy is **1.08**

(2) First, we can get the equation as follows:

$$Bcubed \quad Precision = \frac{\sum_e\frac{|c_e\bigcap s_e|}{|c_e|}}{n}$$

$$Bcubed \quad Recall = \frac{\sum_e\frac{|c_e\bigcap s_e|}{|s_e|}}{n}$$

For the first solution, the BCubed precision and recall is as follows:

$$
\begin{aligned}
Bcubed - precision &= \frac{\frac{2}{4}+\frac{2}{4}+\frac{1}{1}+\frac{2}{4}+\frac{2}{4}+\frac{1}{3}+\frac{1}{2}+\frac{2}{3}+\frac{2}{3}+\frac{1}{2}}{10} \\
&= \frac{17}{30} \\
&\approx 56.67\% \\
Bcubed - recall &= \frac{\frac{2}{3}+\frac{2}{3}+\frac{1}{3}+\frac{2}{4}+\frac{2}{4}+\frac{1}{4}+\frac{1}{4}+\frac{2}{2}+\frac{2}{2}+\frac{1}{1}}{10} \\
&= \frac{37}{60} \\
&\approx 61.67\% \\
F1 - score &= 2 * \frac{P * R}{P + R} \\
&= 2 * \frac{\frac{17}{30} * \frac{37}{60}}{\frac{17}{30} + \frac{37}{60}} \\
&\approx 0.59
\end{aligned}
$$

For the second solution, similarly we can calculate the value, the BCubed precision and recall is as follows:

$$Bcubed - precision \approx 51.67\%,\ Bcubed - reall = 65\%,\ F1 - score \approx 0.58$$

(3) The first system is better, since it has less entropy, according to the entropy measure.

(4) The first system is better, since it has higher F1-score.

3 Problem 3. [10 pts] Leave-one-out (LOO) cross-validation

Answer:

There is some misunderstanding of this question, but I think we can regard it in two ways:

1) This five result is just an experiment with five test set, so we don't need to calculate the average, so for the five test set, there are 3 right predictions and 2 wrong predictions. In all these 5 results, the accuracy is

$$\frac{3}{5} = 60\%$$

2) This five performs as 5 test experiments, so we need to calculate them and find the average.For this five experiments, using Leave-one-out(LOO) cross-validation, we need to calculate the accuracy for each one, and then get the average of them. For experiment (1), (3) and (5), all the labels are predicted correctly, so the accuracy for it is 100%. But for experiment (2) and (4), the

labels are predicted wrong, we can set the smiley face as positive, and the star as negative, then for (2), we can get this table:

Items	Positive	Negative
actual Positive	2	0
actual Negative	1	2

So the accuracy for the (2) experiment is

$$\frac{2+2}{2+2+1} = 80\%$$

The same goes for experiment (4), so the average accuracy is:

$$\frac{100\% + 80\% + 100\% + 80\% + 100\%}{5} = 92\%$$

4 Problem 4. [50 pts] Computing similarity

Answer:

A. For this problem, I wrote some code to calculate the value since it's too complex to calculate by hand, the output is as appended[see appendix], so the result is:

$$Dice(D_1, D_2) = 0.2$$

$$Dice(D_1, D_2)(numberic \quad version) = 0.2745$$

$$Jaccard(D_1, D_2) = 0.159$$

$$cos(D_1, D_2) = 0.299$$

$$Euclidean \quad distance(D_1, D_2) = 7.81$$

B. There are 8 lines in total, and the total lines that appears "little" has 3 lines, the total lines occurring "baby" is 4. So the probability of occuring "little" is 3/8 = 37.5%

And the probability of occurring "baby" is also 4/8 = 50%

For pointwise mutual information, from the equation here:

$$I(X;Y) = \sum_{x,y} p(x,y) * \log \frac{p(x,y)}{p(x)p(y)}$$

Here, p(x) = 37.5%, p(y)=50%, p(x,y)=3/8=37.5%, so the PMI = 0.375.

5 Appendix - Code output for P4

```
    %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
D1.count = 24, D1 = [’dance’, ’little’, ’baby’, ’dance’, ’up’, ’high’, ’never’,
’mind’, ’baby’, ’mother’, ’is’, ’nigh’, ’crow’, ’and’, ’caper’, ’caper’, ’and’, ’crow’,
’there’, ’little’, ’baby’, ’there’, ’you’, ’go’]

    D2.count = 27, D2 = [’up’, ’to’, ’the’, ’ceiling’, ’down’, ’to’, ’the’, ’ground’,
’backwards’, ’and’, ’forwards’, ’round’, ’and’, ’round’, ’dance’, ’little’, ’baby’,
’mother’, ’will’, ’sing’, ’with’, ’the’, ’merry’, ’coral’, ’ding’, ’ding’, ’ding’]

    set1 set([’and’, ’caper’, ’little’, ’there’, ’nigh’, ’dance’, ’crow’, ’is’, ’never’,
’mind’, ’up’, ’high’, ’go’, ’mother’, ’baby’, ’you’])

    set2 set([’and’, ’will’, ’ceiling’, ’backwards’, ’forwards’, ’dance’, ’coral’, ’up’,
’down’, ’to’, ’merry’, ’mother’, ’baby’, ’sing’, ’the’, ’little’, ’with’, ’round’, ’ding’,
’ground’])

    intersection set([’and’, ’little’, ’dance’, ’up’, ’mother’, ’baby’])

    union set([’and’, ’ding’, ’ceiling’, ’nigh’, ’dance’, ’crow’, ’is’, ’never’, ’mind’,
’high’, ’baby’, ’merry’, ’go’, ’sing’, ’with’, ’the’, ’ground’, ’caper’, ’down’, ’little’,
’backwards’, ’forwards’, ’coral’, ’there’, ’up’, ’will’, ’to’, ’mother’, ’you’, ’round’])

    union size 30
    dict1 ’and’: 2, ’caper’: 2, ’little’: 2, ’never’: 1, ’nigh’: 1, ’dance’: 2, ’crow’:
2, ’is’: 1, ’there’: 2, ’mind’: 1, ’up’: 1, ’high’: 1, ’baby’: 3, ’mother’: 1, ’go’: 1,
’you’: 1

    dict2 ’and’: 2, ’down’: 1, ’ceiling’: 1, ’backwards’: 1, ’forwards’: 1, ’dance’:
1, ’coral’: 1, ’up’: 1, ’will’: 1, ’to’: 2, ’merry’: 1, ’mother’: 1, ’baby’: 1, ’sing’:
1, ’the’: 3, ’little’: 1, ’with’: 1, ’round’: 2, ’ding’: 3, ’ground’: 1

    total number of common words 7

    Dice(D1, D2)(numberic version) 0.274509803922 Dice 0.2
    Jaccard(D1, D2) 0.159090909091

    vec1: ’and’: 2, ’ding’: 0, ’forwards’: 0, ’dance’: 2, ’crow’: 2, ’is’: 1, ’mind’:
1, ’high’: 1, ’go’: 1, ’ground’: 0, ’caper’: 2, ’little’: 2, ’nigh’: 1, ’there’: 2, ’to’:
0, ’you’: 1, ’ceiling’: 0, ’never’: 1, ’merry’: 0, ’baby’: 3, ’sing’: 0, ’with’: 0,
’down’: 0, ’backwards’: 0, ’coral’: 0, ’up’: 1, ’will’: 0, ’mother’: 1, ’the’: 0,
’round’: 0

    vec2: ’and’: 2, ’ding’: 3, ’forwards’: 1, ’dance’: 1, ’crow’: 0, ’is’: 0, ’mind’:
0, ’high’: 0, ’go’: 0, ’ground’: 1, ’caper’: 0, ’little’: 1, ’nigh’: 0, ’there’: 0, ’to’:
```

2, 'you': 0, 'ceiling': 1, 'never': 0, 'merry': 1, 'baby': 1, 'sing': 1, 'with': 1, 'down': 1, 'backwards': 1, 'coral': 1, 'up': 1, 'will': 1, 'mother': 1, 'the': 3, 'round': 2

total=13, distance1=24, distance2=27

cos 0.299028409048

Euclidean distance 7.81024967591

Jaccard(numeric version) 0.159090909091