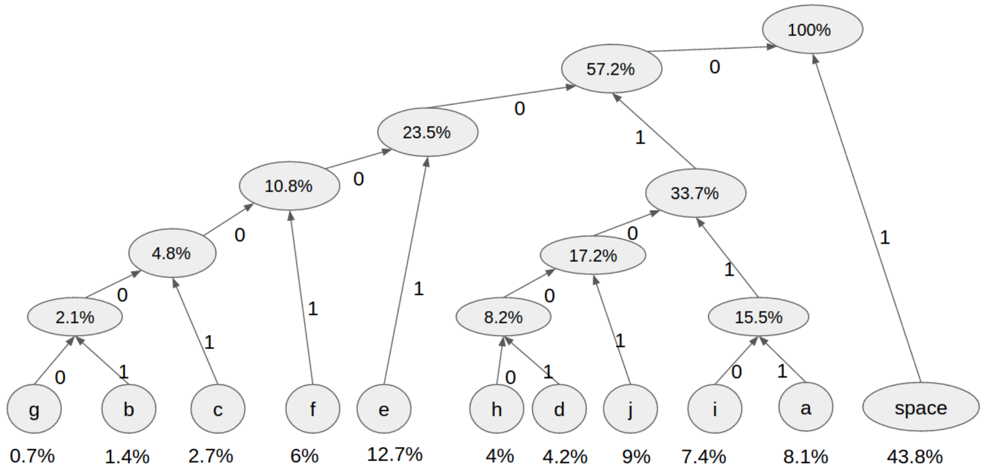
91.542 - Natural Language Processing
Homework Assignmen 3

Chang Liu
chang_liu@student.uml.edu

October 19, 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

‘010000010111010010110’.

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

Symbol	Representation	Length
a	0111	4
b	000001	6
c	00001	5
d	01001	5
e	001	3
f	0001	4
g	000000	6
h	01000	5
i	0110	4
j	0101	4
space	1	1

So the average value is:

$$\frac{4+6+5+5+3+4+6+5+4+4+1}{11} = \frac{47}{11} \approx 4.27$$

(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:

$$\begin{aligned}
entropy &= \sum_{i \in C} \sum_{j \in S} \frac{|c_i \bigcap s_j|}{|c_i|} \log \frac{|c_i \bigcap s_j|}{|c_i|} \\
&= (\frac{2}{4} * \log_2 \frac{2}{4} * 2 + 0 + 0) + (0 + \frac{1}{4} * \log_2 \frac{1}{4} + \frac{2}{3} * \log_2 \frac{2}{3} + 0) \\
&+ (0 + \frac{1}{2} * \log_2 \frac{1}{2} * 2 + 0) + (\frac{1}{1} * \log_2 \frac{1}{1} + 0 + 0 + 0) \\
&\approx -2.89
\end{aligned}$$

Regarding the entropy, we have to get the negative value of the above equation, which should be **2.89**

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

$$\begin{aligned}
entropy &= \sum_{i \in C} \sum_{j \in S} \frac{|c_i \bigcap s_j|}{|c_i|} \log \frac{|c_i \bigcap s_j|}{|c_i|} \\
&= (\frac{3}{4} * \log_2 \frac{3}{4} + \frac{1}{4} * \log_2 \frac{1}{4}) + (\frac{2}{3} * \log_2 \frac{2}{3} + \frac{1}{3} * \log_2 \frac{1}{3}) \\
&+ (\frac{1}{3} * \log_2 \frac{1}{3} * 3) \\
&\approx -3.31
\end{aligned}$$

So for the second clustering solution, the entropy is **3.31**

(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

4 Problem 4. [50 pts] Computing similarity