CSC 4170

Web Intelligence and Social Computing

Homework Assignment #4

Sample Answer

1, (20 marks)

Based on the idea of reCAPTCHA, we can form a CAPTCHA picture by combining the word “hello” and “world”, and requires the users to type in two words. If a user types two words and the first word “hello” is typed right, this case is referred to as a valid case. If the number of words the user types does not equal to 2 or the first word “hello” is not typed in correctly, this case is referred to as an invalid case. After a certain period, we can count the words’ frequencies of the second word in the valid cases, and set a threshold to determine whether the most typed word is reliable. For example, if there are 100 valid cases and 80 of them contains “world” as the second word. Suppose we set the threshold to 0.7, then 80/100 > 0.7, and we can consider the “world” recognized by the reCAPTCHA procedure is reliable.

2, (1) (6 marks)

ESP game belongs to the collaborative game. Reason: It requires 2 players to work together so that they can match a common word for each image.

(2) (6 marks)

“Taboo” words are words that cannot be entered as possible labels. Because taboo words have been entered too many times by the users, it is not very valuable to the system. Thus, the system prefers to collect some not so frequent words that can also describe the images.

(3) (8 marks)

If one user leaves a game, EPS game could place a robot against the remaining user by annotating labels used by previous users.

3. (20 marks)

We can use the Kappa Statistics to evaluate the assessment task to see whether it is reliable.

P(A) = (400+80)/540 = 0.8889

P(nonrelevant)=(100+120)/(540+540) = 0.2037

P(relevant)=(440+420)/(540+540) = 0.7963

$P(E)=P(\text{nonrelevant})^2 + P(\text{relevant})^2 = 0.2037^2 + 0.7963^2 = 0.6756$

k = (P(A)-P(E))/(1-P(E)) = (0.8889-0.6756)/(1-0.6756)=0.6575

It can be seen as substantial agreement. (Refer to tutorial Evaluation.ppt)

4.(1) (10 marks)

P(s|M1) = 0.1 * 0.01 * 0.2 = 0.0002

P(s|M2) = 0.12 * 0.0002 * 0.15 = 0.0000036

(2) (10 marks)

The document model 1 is more likely to generate the word sequence s.

5 (1) (5 marks)

$$P(d|q) = P(q|d)P(d)/P(q)$$

P(d1|q) = P(q|d)p(d)/P(q) = p(q|d) = 1/3 * 0/3 = 0

P(d2|q) = 0/4 * 2/4 = 0

(2) (15 marks)

We can use the Jelinek-Mercer Smoothing technique to smooth. Suppose we set the λ =0.5:

P(d1|q) = ((1/3+1/7)*0.5) * ((0/3+2/7)*0.5)= 0.0340

P(d2|q) = ((0/4+1/7)*0.5) * ((2/4+2/7)*0.5)= 0.0281

So the d1 is more relevant compared with d2.