CSC 4170
Web Intelligence and Social Computing

Homework Assignment #4
Date: Friday, 20 November 2009
Due Date: Friday, 4 December 2009 at 6:30 pm

`http://wiki.cse.cuhk.edu.hk/irwin.king/teaching/csc4170/2009`

---

1. You may form a group of no more than two persons to do the homework assignments.
2. If you decided to form a group with two persons, the final score for the assignment will be given to both persons. Moreover, since it is done by two persons, you are expected to do more than just a single person group.
3. Once the group has been formed, you should stay together throughout the class, including the class project as well.
4. Lastly, the final examination will be assessed individually.
5. Make sure you submit the electronic copy of your homework assignment to the VeriGuide system through the web as Assignment #1.

Homework Assignment

1. As a famous application of Human Computation, reCAPTCHA has been introduced both in lecture and tutorial. Now we have two words: "hello" (this word can be recognized by Optical Character Recognition) and "world" (this word can not be recognized by OCR). Suppose we have 100 unknown internet users to help recognize the word. Use the idea of reCAPTCHA to design a procedure to effectively recognize the word "world". (20 marks)
2. Social game-based human computation with online players has emerged as an effective way to leverage human knowledge, and the game mechanism contains three forms: collaborative games, competitive games, and hybrid game. ESP game is the pioneer of GWAP. Answer following questions related to the ESP game:
   (a) ESP game should belong to which kind of game mechanism? The justification is also needed. (6 marks)
   (b) Play a round of the ESP game as a registered user or guest, and describe the usage of taboo words. (6 marks)
   (c) If you are required to design a two players' game like the ESP game, how would you design a mechanism to handle the case that one player leaves the game before it ends? (8 marks)

3. In crowdsourcing, one existing challenge is that not every internet user is trustworthy. Suppose two internet users are involved in a "relevant" and "irrelevant" assessment task, and the agreement result between these two users are shown in Table 1:

Table 1: Agreement result of two users in question 3.

	Yes	No	Total
Yes	400	40	440
No	20	80	100
Total	420	120	540

Use a way to evaluate whether the results of two internet users in the assessment task are reliable. (Hint: Kappa statistics). (20 marks)

4. Language model is a very important technique in the Information Retrieval and Natural Language research, two document models are described in Table 2:

Table 2: Two document models

Model M_1	Model M_2
cse 0.2	cse 0.15
cuhk 0.1	cuhk 0.12
csc4170 0.01	csc4170 0.0002
assignment 0.01	assignment 0.0001
web 0.06	web 0.03
intelligence 0.07	intelligence 0.005

Suppose we have a word sequence s: cuhk csc4170 cse, and we use the unigram model in this question.

(a) Calculate the probabilities of two document models generate the sequence s: $p(s|M_1)$ and $p(s|M_2)$. (10 marks)

(b) Which document model do you think is more likely to generate the word sequence s? (10 marks)

5. Suppose we have two candidate documents and one query as follows.

- d_1: cuhk cse assignment
- d_2: csc4170 cse csc4170 web
- query: cuhk csc4170

We use the unigram language model, employ the maximum likelihood estimation, and assume the probabilities $p(d_1) = p(d_2)$.

(a) Use the basic query likelihood model to rank the d_1 and d_2 for the query according to their relevance. (5 marks)

(b) The basic query likelihood model suffers from some problems in this question, and you are required to suggest an advanced version of the query likelihood model that can tackle the problem. (Hint: smoothing) (15 marks)