Remarks for NLP Final Examination

自然語言處理期末考注意事項

1. Please use A4 size as your answer sheets. 答案卷請使用 A4 紙大小。
2. The first page is the cover page which only shows your Student ID and name. 第一頁為封面頁，只需寫上你的學號與姓名。
3. The answers are written starting from page 2. Please answer in sequence and you don't need to copy the problem. Staple the answer sheets on the top left-hand corner. 請自第二頁開始，依題號順序作答，不需抄題，繳交時答案卷需以訂書針在左上角裝訂好。
4. Return your answers to assistant Yang in the class during 14:20-15:00, June 13, 2019. Remember sign your name on the signature form. 於 6/13 (四)下午 14:20-15:00 間在上課教室將答案卷交給助教，並在簽名表上簽名。
5. If you don't follow the rules mentioned above, i.e., rules about answer sheets and returning deadline, you will get a point deduction. 答案卷格式不符或遲交者，將會扣分。

Natural Language Processing Final Examination
(6/13/2019)

1. (1) Define per-word entropy. (10 points)

   (2) Give an application that uses “mutual information” as the evaluator. (10 points)

2. In biological literature, researchers often use some verbs to express the relationships between two biological entities like proteins. The following shows an example, where **MTH1** and **HSP82** are two proteins, and *interact* denotes a relationship.

   MTH1 *interacts* directly with the **HSP82**.

   Given a corpus of biological literature, you are asked to mine the verbs that are used to represent the relationships. Please show your method. (20 points)

3. Given the following two sentences:

   > Four Arab nations cut diplomatic ties to Qatar early Monday morning. Bahrain, Egypt, Saudi Arabia and the United Arab Emirates all announced they would withdraw their diplomatic staff from Qatar.

   (1) List the bigrams of the above sentences with tag patterns “A N” and “N N” where “A” represents adjective and “N” means noun. (5 points)

   (2) List the trigrams of the above sentences with tag patterns “A A N,” “A N N,” “N A N,” “N N N” and “N P N” where “P” represents preposition. (5 points)

   (3) In your opinion, list good collocations for these bigrams and trigrams. (5 points)

4. (1) What is the sparse data problem? (10 points)
   (2) Propose a combining method to resolve this problem. (10 points)

5. If large amount of annotated data is not available, how do you evaluate your model? In other words, you have to use the data both in training and in testing, and avoid the overtraining problem. (10 points)

6. Lesk proposed an algorithm for word sense disambiguation in 1986. Please briefly describe the idea of Lesk’s algorithm, and give an example to show the process of word sense disambiguation. (15 points)