1.

(1) Define per-word entropy. (10 points) 6-58

信息量的大小隨著訊息長度的增加而增加

(2) Give an application that uses “mutual information” as the evaluator. (10 points) 6-63

Clustering words

word sense disambiguation

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)

**MTH1** affects directly on the **HSP82**

3. 7-13

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)

AN : Arab nations,early Monday,United Arab,diplomatic staff

NN : Saudi Arabia,Monday morning

(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)

ANN :United Arab Emirates,their diplomatic staff

ANN :early Monday morning

NAN :Four Arab nations,Qatar early Monday

NNN :Bahrain Egypt Saudi, Egypt Saudi Arabia

NPN :staff from Qutar

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

NNNNNNNNN 是好結果

4. 8-19

(1) What is the sparse data problem? (10 points)

Data is considered sparse when certain expected values in a dataset are missing

(2) Propose a combining method to resolve this problem. (10 points)

Use a discounting or smoothing technique.  
smoothing : An alternative solution is to smooth the probability estimate so that no word sequences s give a probability of 0.

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)

在這種annotated data不足的情況下，我們可以透過cross-validation方法來幫助我們解決annotated data 不足的問題，我們暫時將現有的data切成n等份，並且輪流用期中的1/n暫時當作testing data，剩下的當成training data，老師的講義有更詳盡的說明。

6. 10-26

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)

Lesk’s algorithm: a word’s dictionary definitions are likely to be good indicators for the sense they define。

Ex:

I can’t bear with the bears.