

This question paper must be returned. Candidates are not permitted to remove any part of it from the examination room.

STUDENT'S SURNAME _____

OTHER NAMES _____

STUDENT NUMBER _____



Macquarie University
Division of Information and Communication Sciences
Department of Computing
Mid-Year Examinations 2004

Unit : COMP348 Document Processing and the Semantic Web

Date : 2004

Time Allowed : 3 hours plus ten minutes reading time

Total Questions : 10

Total Marks : 60

Instructions : Answer ALL questions.
This examination consists of two sections, A and B.
The answers to each of Sections A and B must be written in a separate answer book. Each of Sections A and B is worth 30 marks. Write your name and student number on the cover of each answer book. Also write the section (A or B) on the cover of each answer book.
Write your name and student number at the top of this page. You may do rough working on this question paper, but all answers MUST be submitted as described above. Hand in this question paper at the end of the examination.
No calculators are allowed. No dictionaries are allowed.

SECTION A

(5 QUESTIONS, 30 MARKS)

Use a separate book for Section A

1. (3 marks) Write a Python program that will read in data of the following form

```
Stephen:27:jazz
Mary:25:linux and Elvish
Simon:24:philosophy
```

and print out output of the following form

```
Stephen is 27 years old and likes jazz.
Mary is 25 years old and likes linux and Elvish.
Simon is 24 years old and likes philosophy.
```

Your program should expect the name of the input file as a command-line argument, open the file, process each line of the input, and complete by closing the file.

2. (3 marks) Describe three basic approaches to identifying key sentences for document summarisation.
3. The following probabilities have been estimated from a small corpus.

$P(N \emptyset) = 0.7$	$P(V \emptyset) = 0.3$	$P(N N) = 0.2$	$P(V N) = 0.6$
$P(P N) = 0.2$	$P(N V) = 0.7$	$P(P V) = 0.3$	$P(N P) = 0.8$
$P(\text{oranges} N) = 0.2$	$P(\text{fruit} N) = 0.2$	$P(\text{flies} N) = 0.2$	$P(\text{like} P) = 0.2$
$P(\text{on} P) = 0.2$	$P(\text{flies} V) = 0.3$	$P(\text{like} V) = 0.2$	

- (a) (1 marks) How do you think the probability $P(N|V)$ has been estimated?
- (b) (2 marks) Draw the Hidden Markov Model that corresponds to the above probabilities.
- (c) (2 marks) Compute the probabilities of the following sequences of part-of-speech tags:
 1. fruit/N flies/N like/V oranges/N
 2. fruit/N flies/V like/P oranges/N

4. You are given the following two reference translations R1 and R2, and the following two candidate translations C1 and C2.

R1: Sing, O goddess, the anger of Peliad Achilleus, that brought countless ills upon the Achaeans.

R2: Sing, goddess, the anger of Peleus' son Achilleus and its devastation, which put pains thousandfold upon the Achaeans.

C1: Sing, Déesse, of Peliad Achilleus the disastrous anger, which infinite evils overpowered Achaeans.

C2: Sing, Goddess, of the Peliad Achilleus the disastrous anger, that overwhelmed the Achaeans of infinite pains.

- (a) (4 marks) Calculate the unigram modified precisions, as defined in the BLEU algorithm, for each candidate translation C1 and C2. Ignore all punctuation in this calculation.

You may wish to draw up a table with the following columns:

n-gram	cand. freq.	R1 freq.	R2 freq.	max.	mod. prec.
...

- (b) (3 marks) Calculate the bigram modified precisions, as defined in the BLEU algorithm, for each candidate translation C1 and C2. Ignore all punctuation in this calculation.
- (c) (2 marks) In the BLEU algorithm, modified n-gram precisions are combined by a geometric average. What are the geometric averages of unigram and bigram modified precisions for C1 and for C2? Consequently (ignoring the brevity penalty), which of C1 and C2 is the better translation?
- (d) (1 marks) Why does the BLEU algorithm use a brevity penalty? Give an example.

5. (a) (3 marks) You are given the following two translations from Klingon to English.

(1) puq legh yaS.
child see officer

The officer sees the child.

(2) ja'chuqmeH rojHom neH jaghla'.
confer-PURPOSE truce want enemy-commander

The enemy commander wants a truce (in order) to confer.

The vocabulary for this second example corresponds as follows.

ja'chuq	confer
-meH	-PURPOSE, "in order to"
rojHom	truce
neH	want
jaghla'	enemy commander

What are two types of translation divergences occurring here? Describe a transfer-based machine translation approach to resolving these types of divergence. Use tree diagrams to illustrate your answer.

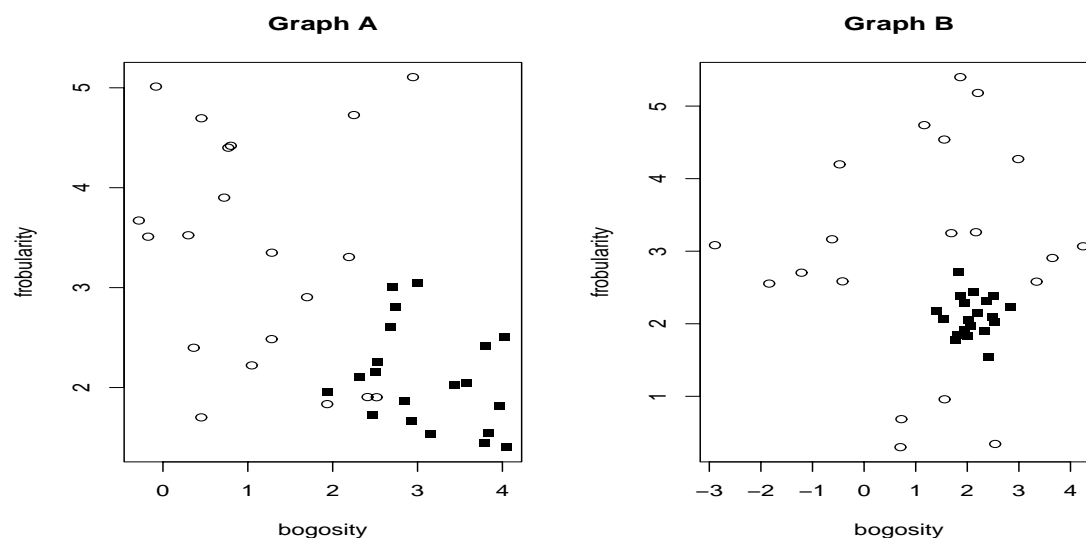
- (b) (5 marks) Describe how you might use Python to carry out a transfer-based translation from English to Klingon. In particular:
1. show the data structure you would use to represent the first English sentence above (using NLTK module definitions or otherwise);
 2. give a Python dictionary that will give the Klingon translation of each English word in the first sentence; and
 3. write a Python function that will translate from the data structure you have chosen for English in (i) above for the first sentence, into a corresponding data structure for Klingon.
- (c) (1 marks) Explain the two important conceptual components of a statistical machine translation system, in terms of the properties of the translation that they aim to maximise. (No more than 8 lines.)

SECTION B

(5 QUESTIONS, 30 MARKS)

Use a separate book for Section B

1. The graphs below show two sets of data points which represent document sets in a two dimensional feature space. The features (*bogosity* and *frobularity*) are continuous values.



- (a) (2 marks) Comment on the usefulness of these features in separating the sets of documents shown in the graphs. Is one feature any better than the other in either case?
 - (b) (6 marks) Discuss **two** text classification methods with specific reference to why they might be applicable to the data shown in each of the graphs.
2. (4 marks) What are the principal features used in *word sense disambiguation*? It is said that using WSD methods in *information retrieval* does not provide significant increased recall. With reference to the algorithms typically used for both tasks, suggest why this might be the case.
 3. (a) (4 marks) Describe the features that make Information Retrieval on the Web different to that in, say, a user's personal files. How do these differences impact the way that Google is implemented?
 - (b) (3 marks) Describe Google's PageRank algorithm. How is the PageRank integrated into the other IR measures to come up with an overall ranking for a set of search results?

4. (a) (3 marks) Give an example of a commonly used RDF *ontology* and describe how its use might enable some of the promised features of the Semantic Web to be realised.

- (b) (2 marks) Draw the RDF graph corresponding to the following fragment of N3:

```
@prefix foaf: <http://xmlns.com/foaf/0.1/> .
@prefix ical: <http://www.w3.org/2002/12/cal/ical#> .
```

```
:jim a foaf:Person;
    foaf:name "Jim Smith";
    foaf:mbox <mailto:jim@mail.com>;
    foaf:knows :fred .
```

```
:fred a foaf:Person;
    foaf:name "Fred Dag";
    foaf:mbox <mailto:dag@mail.com>;
    foaf:knows :jim .
```

```
:event1 a ical:Vevent;
    ical:summary "Project Supervision Meeting";
    ical:rrule [
        ical:freq "WEEKLY";
        ical:byday "WE";
        ical:interval "1";
    ];
    ical:attendee :jim, :fred .
```

5. (a) (4 marks) Describe in general terms the Burrows-Wheeler *block sorting* algorithm and explain why it improves the compression of text using Huffman or Arithmetic coding.

- (b) (2 marks) What conditions must be fulfilled to allow direct indexing of compressed file collections?