

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 2005

Unit : COMP348 Document Processing and the Semantic Web

Date : 2005

Time Allowed : 3 hours plus ten minutes reading time

Total Questions : 9

Total Marks : 180

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 90 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, 90 MARKS)

Use a separate book for Section A

1. (15 marks) You are given input consisting of strings of the form w/p , where w is a word and p is a part of speech; for example,

```
The/DT cost/NN of/IN a/DT local/JJ call/NN has/VBP fallen/VBN
Call/VBP me/PRP Ishmael/NNP
```

Write a Python program with the following characteristics:

- it expects as a command-line argument the name of a file with input as above;
- it builds a dictionary that contains the count of each **distinct word and part-of-speech pair** (that is, `call/NN` is distinct from `call/VBP`);
- it builds a dictionary containing the count of each part of speech (that is, `NN`, `VBP`, ...); and
- it prints out the contents of both of these dictionaries.

(As a guideline, if your program is longer than about 30 lines, you've probably done too much.)

2. (9 marks) Discuss three problems that occur in segmenting and tokenising human-language text (English or otherwise).
3. Consider the word *ring*: it can have the sense of *a piece of jewellery worn on the finger* (s_1) or *to call (someone)* (s_2). For example:

(1) I had pondered what I might do, should I have on my finger the One Ring made by the craft of Sauron.

(2) Don't just let the telephone ring on and on.

In (1), *ring* has the sense of *jewellery* (s_1), and in (2) the sense of *call* (s_2).

It has been determined that there are 5 keywords which can be used to distinguish these senses: *finger*, *made*, *one*, *telephone*, *mother*. Each instance i of *ring* can thus be represented by a vector v_i of keyword occurrences within a **7 word window**, where element $v_i[0]$ represents whether *finger* occurs within this 7 word window, $v_i[1]$ represents whether *made* occurs, and so on.

(a) (3 marks) What are the vectors corresponding to sentences (1) and (2)?

(b) (9 marks) You have, **in addition**, the following context vectors:

s_1	s_2
(1, 0, 1, 0, 0)	(0, 0, 0, 0, 1)
(0, 1, 1, 0, 0)	(0, 0, 1, 1, 0)
(1, 0, 0, 0, 1)	(1, 0, 0, 0, 1)
	(0, 0, 0, 0, 1)

Including the two vectors from part (a), calculate the following values:

$\Pr(s_1)$	$\Pr(s_2)$
$\Pr(v[0] = 1 \mid s_1)$	$\Pr(v[0] = 1 \mid s_2)$
$\Pr(v[1] = 1 \mid s_1)$	$\Pr(v[1] = 1 \mid s_2)$
$\Pr(v[2] = 1 \mid s_1)$	$\Pr(v[2] = 1 \mid s_2)$
$\Pr(v[3] = 1 \mid s_1)$	$\Pr(v[3] = 1 \mid s_2)$
$\Pr(v[4] = 1 \mid s_1)$	$\Pr(v[4] = 1 \mid s_2)$

(c) (6 marks) Consider the sentence

If it stops after one ring it's my mother.

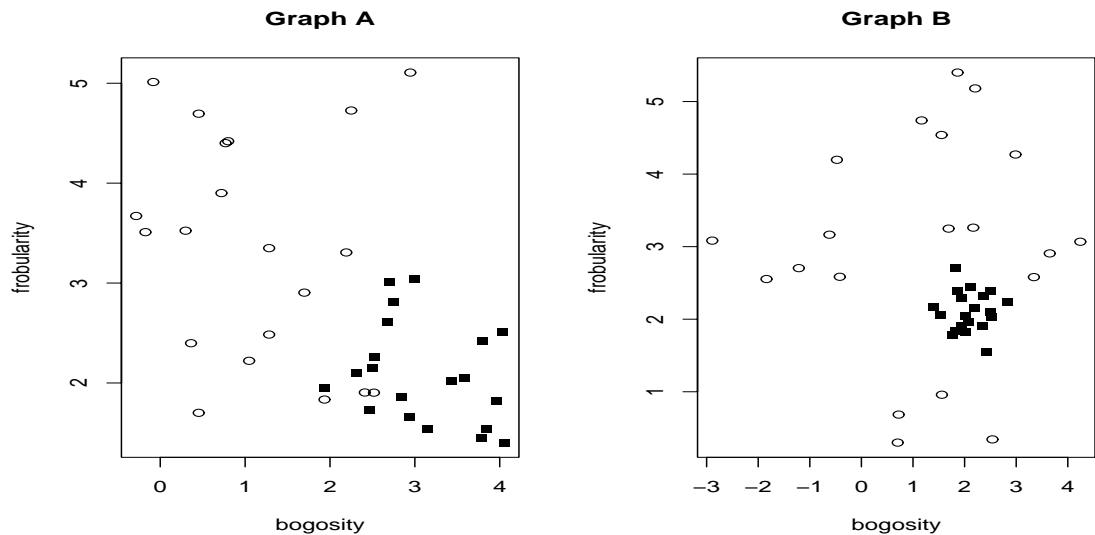
(You should assume *it's* is a single token.)

Use a Naïve Bayes approach to determine the probability of senses s_1 and s_2 for this sentence. You may leave the final answer as an expression consisting of probabilities multiplied together. Which sense is more likely?

(d) (2 marks) What other information might help you disambiguate these two senses of *ring*?

4. (a) (3 marks) In text classification, the text to be classified is typically represented by a vector, just as for Word Sense Disambiguation in Question 3. Explain briefly what a *feature* is in the context of text classification, and how it relates to this vector.

(b) (7 marks) The following diagram represents a set of data points measured by features (*bogosity* and *frobularity*), and their class by colour (black or white).



- State whether the data is linearly separable in Graph A only, in Graph B only, in both graphs or in neither graph.
- What is required for the data to be linearly separable?
- Name one text classification method other than Naïve Bayes, and describe broadly how it separates the data into classes.

(c) (6 marks) Consider the following code discussed in lectures, which uses the Natural Language Toolkit to classify words by their part of speech. How would you

add a feature LEN representing the length of the word to be classified? Write lines of Python code to do this, and indicate using the numbered comment lines where in the program these lines of code should be placed. (Not all need be used.)

```

from nltk.classifier import *
from nltk.feature import *
from nltk.tokenizer import *
import re

def my_feature_demo():

    # Load the training data, and split it into test & train.
    import nltk.corpus
    text = nltk.corpus.brown.read('cr01', add_contexts=True)
    toks = text['WORDS']

    # -- OPTIONAL CODE ADDITION [1] --
    split = len(toks) * 3/4
    train, test = toks[:split], toks[split:]

    # -- OPTIONAL CODE ADDITION [2] --
    # Create the feature detector.
    from nltk.feature.word import BagOfContextWordsFeatureDetector
    detector = MergedFeatureDetector(
        PropertyFeatureDetector('TEXT'),
        PropertyFeatureDetector('TAG'),
        # -- OPTIONAL CODE ADDITION [3] --
        BagOfContextWordsFeatureDetector(window=2))

    # Run feature detection on the training data.
    for tok in train:
        detector.detect_features(tok)

    # Build a feature encoder, based on the training data.
    encoder = learn_encoder(train, unseen_cutoff=0)

    # Run the feature encoder on the test data.
    for tok in test[:10]:
        print 'Input token: ', tok.exclude('CONTEXT')
        detector.detect_features(tok)
        # -- OPTIONAL CODE ADDITION [4] --
        items = tok['FEATURES'].items()
        print ('Feature dict: { ' +
              (',\n'+16*' ').join(['%r: %r' % i for i in items]) +
              ' }')
        encoder.encode_features(tok)
        print 'Feature vector:',
        assignments = tok['FEATURE_VECTOR'].assignments()
        assignments.sort()
        print ', '.join(['v[%d]=%d' % (index, val)
                        for index, val in assignments])
        print
        # -- OPTIONAL CODE ADDITION [5] --

if __name__ == '__main__':
    my_feature_demo()

```

- (d) (8 marks) Also starting from the above code, how would you add a feature `WITH_W` representing whether the word to be classified contains the letter w ?
5. (a) (4 marks) Explain the ideas behind transfer-based machine translation and interlingua-based machine translation, and briefly compare the two approaches.
- (b) (4 marks) Explain briefly the notions of faithfulness and fluency in statistical machine translation.
- (c) (4 marks) We want to find the most probable translation of the following Dutch sentence:

Ik houd van een kop sterke koffie.

The first part of our sentence is already generated:

I love a cup of

The task in this question is to find the most likely ending for this sentence in English. To do this we need to translate the last two words in Dutch and try the different combinations.

To calculate the *faithfulness* part the following statistics are given:

Dutch	English	$P(e d)$
koffie	coffee	$\frac{9}{10}$
koffie	caffeine	$\frac{1}{10}$
sterke	powerful	$\frac{6}{10}$
sterke	strong	$\frac{4}{10}$

Table 1:

English	Dutch	$P(d e)$
coffee	koffie	$\frac{8}{10}$
caffeine	koffie	$\frac{2}{10}$
powerful	sterke	$\frac{8}{10}$
strong	sterke	$\frac{4}{10}$

Table 2:

Initially we want to maximize the probability of an English sentence, given the initial Dutch sentence, and we estimate this probability by using Bayes Rule. Which of the above two tables do we therefore need, and which is not relevant? Why?

- (d) (10 marks) Following on from part (c), to calculate the *fluency* we use the following statistics:

e_1, e_2	e_3	$P(e_3 e_1, e_2)$
cup,of	strong	$\frac{8}{10}$
cup,of	powerful	$\frac{2}{10}$
of,strong	coffee	$\frac{8}{10}$
of,strong	caffeine	$\frac{2}{10}$
of,powerful	coffee	$\frac{9}{10}$
of,powerful	caffeine	$\frac{1}{10}$

Copy the following table, complete it and indicate the best translation.

Possible Translations	$P()$	$P(e)$	$P(e) * P()$
I love a cup of strong caffeine	$\frac{1}{100}$	$\frac{1}{100}$	$\frac{1}{10000}$
I love a cup of powerful caffeine	$\frac{1}{100}$	$\frac{1}{100}$	$\frac{1}{10000}$
I love a cup of strong coffee	$\frac{1}{100}$	$\frac{1}{100}$	$\frac{1}{10000}$
I love a cup of powerful coffee	$\frac{1}{100}$	$\frac{1}{100}$	$\frac{1}{10000}$

End of Section A.
Use a new booklet for Section B.

SECTION B

(5 QUESTIONS, 90 MARKS)

Use a separate book for Section B

6. Given the following newspaper extract:

“These are the people who are out there raising our families, paying their mortgages, working hard, building this nation and they get \$6 a week. It’s not good enough” said Labor leader Kim Beazley.

Mr Costello taunted Mr Beazley over his opposition to the cuts, saying Labor’s position was futile because the government gained control of the Senate on July 1 and would pass them.

The Treasurer also revealed he had held talks with Tax Commissioner Michael Carmody over whether an interim position could be put in place to provide tax relief if the laws are blocked.

- (a) (10 marks) Identify the *Named Entities* in the passage along with their type. Comment on any difficult cases.
- (b) (10 marks) Describe the problem of co-reference resolution and illustrate it with examples from the passage. What strategies exist to address this problem?
7. (a) (15 marks) List five different factors that can be used to rank matches in a web information retrieval application. Describe how each factor is extracted or calculated and why it might be useful in ranking.
- (b) (5 marks) What advantages does the vector space matching algorithm have over a simple boolean query processor in information retrieval?
8. (a) (15 marks) What is the status of the Dublin Core metadata set within the Semantic Web community? What are the advantages of using Dublin Core in RDF descriptions? What applications might this enable?
- (b) (15 marks) Sketch an *ontology* for Australian Television that might be used to represent information about programming and other activities of the major broadcasters. You may wish to illustrate your answer with example RDF triple descriptions of some entities from the domain.
- (c) (10 marks) What is the main difference between the different levels of the OWL standard (OWL Lite, OWL DL and OWL Full)? OWL DL is based on Description Logic; what useful properties does DL have and what kinds of data can be represented easily in DL?

9. Figure 1 displays a simple RDF graph.

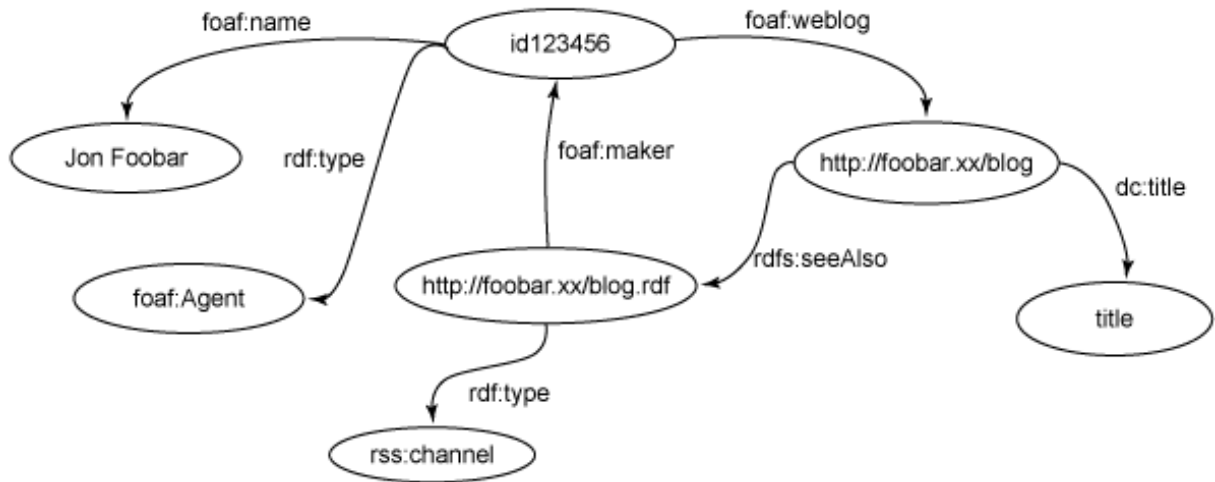


Figure 1: A simple RDF graph

- (7 marks) Write down a serialisation of this graph in N3 format; you may make up suitable URLs for the namespace prefixes.
- (3 marks) Write a SPARQL query to find the URL of the weblog of *John Foobar*.