The University Of Sheffield.

Data Provided: None

DEPARTMENT OF COMPUTER SCIENCE **Autumn Semester 2009-2010**

TEXT PROCESSING **2 hours**

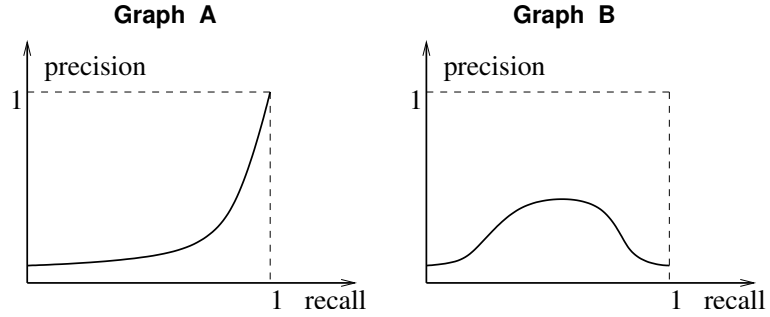
Answer THREE questions.

All questions carry equal weight. Figures in square brackets indicate the percentage of available marks allocated to each part of a question.

Registration number from U-Card (9 digits) — to be completed by student

1. a) Explain the difference between *stemming* and *morphological analysis*. Suggest a text processing context where stemming might be used. Suggest a text processing context where morphological analysis might be useful, but where simple stemming would *not* be useful. [20%]

   b) Define the precision and recall measures in IR. Is Graph A a possible precision/recall graph? Is Graph B a possible precision/recall graph? Explain your answers. [20%]



   c) A common approach to classification in text processing is to assign a category label $v \in V$ to an instance based on a number of feature values $f_1 \ldots f_n$ that serve to describe the instance or its context, where the label v that is chosen is the most probable or MAP (*maximum a posteriori*) hypothesis, as follows:

$$v_{MAP} = \underset{v \in V}{\operatorname{argmax}} P(v|f_1 \ldots f_n)$$

   Show how this approach can be reformulated using Bayes Theorem, and an assumption of *conditional independence*, to give a *Naive Bayes classification* method. Explain the benefits of this reformulation in relation to the problem of *data sparseness*. [20%]

   d) Indicate what will be printed by each of the following pieces of Python code, explaining your answer:

   (i) [10%]

```
import re
s = 'abracadabra'
for x in re.findall('[ar].',s):
    print x
```

   (ii) [10%]

```
import re
pattern = re.compile('^[a-z]+$')
s = "My baby don't love nobody but me."
for x in s.split():
    if pattern.search(x):
        print x
```

   e) What is *part-of-speech-tagging*? Explain why this task is difficult, i.e. why it requires something more than just simple dictionary look-up. [20%]

2. a) Text compression techniques are important because growth in volume of text continually threatens to outstrip increases in storage, bandwidth and processing capacity. Briefly explain the differences between:

   (i) **symbolwise** (or statistical) and **dictionary** text compression methods; [10%]

   (ii) **modelling** versus **coding** steps; [10%]

   (iii) **static**, **semi-static** and **adaptive** techniques for text compression. [10%]

   b) The script for the fictitious language Gavagese contains only the 7 characters a, e, u, k, r, f, d. You assemble a large electronic corpus of Gavagese and now want to compress it. You analyse the frequency of occurrence of each of these characters in the corpus and, using these frequencies as estimates of the probability of occurrence of the characters in the language as a whole, produce the following table:

Symbol	Probability
a	0.25
e	0.20
u	0.30
k	0.05
r	0.07
f	0.08
d	0.05

   (i) Show how to construct a Huffman code tree for Gavagese, given the above probabilities. [30%]

   (ii) Use your code tree to encode the string $dukerafua$ and show the resulting binary encoding. For this string, how much length does your codetree encoding save over a minimal fixed length binary character encoding for a 7 character alphabet? [10%]

   c) One popular compression technique is the LZ77 method, used in common compression utilities such as $gzip$.

   (i) Explain how LZ77 works. [20%]

   (ii) How would the following LZ77 encoder output

$$\langle 0,0,b\rangle\langle 0,0,a\rangle\langle 0,0,d\rangle\langle 3,3,b\rangle\langle 1,3,a\rangle\langle 1,3,d\rangle\langle 1,3,a\rangle\langle 11,2,a\rangle$$

   be decoded, assuming the encoding representation presented in the lectures? Show how your answer is derived. [10%]

3. a) What is a *stoplist* (or list of *stop words*) in the context of an information retrieval system? Discuss the significance of using a stoplist in relation to both the effectiveness of retrieval and computational efficiency. [10%]

   b) Discuss the advantages and disadvantages of *boolean* versus *ranked* approaches to information retrieval. [15%]

   c) Describe and explain the TF.IDF term weighting scheme. Include the formula (or formulae) for computing TF.IDF values as part of your answer. [20%]

   d) Explain what is meant by an *inverted index*, and why such indices are important in the context of information retrieval. Suggest a suitable data strucure (e.g. in Python) for storing an inverted index. [15%]

   Assume that you are provided with a Python class `Collection` which provides access to the documents of a particular document collection in the following way: an instance of this class is an *iterator* that will successively return the documents of the collection one at a time. The documents are returned as instances of a `Document` class, that have two attributes: an attribute `docid` whose value is a document identifier string (e.g. `"doc0032"`), and an attribute `lines` whose value is a list of strings for the lines of text in the document. For example, the following code would print the identifier and *first* line of each document in the collection:

```
import Collection
collection = Collection.Collection()
for doc in collection:
    print "ID: ", doc.docid
    print doc.lines[0]
```

   e) Write Python code (preferably as a function) which computes an *inverted index* for the above document collection (i.e. accessed using the `Collection` class). Your index should store the *counts* of terms appearing in the different documents. [25%]

   f) Write Python code (preferably as a function) which, when given a single term and an inverted index, uses the index to identify one document in the collection containing that term and prints it out (or prints a warning if no documents contain the term). [15%]

4. a) Explain the differences between *direct*, *transfer* and *interlingual* approaches to Machine Translation. [20%]

   b) Describe the main features of the approach to statistical machine translation developed by the IBM research group in the 1990s. [30%]

   c) What is *summarisation*? Explain what is meant by *sentence selection* as an approach to summarisation. Discuss some of the common problems for summaries produced in this way. [20%]

   d) Discuss the criteria that have been used by *shallow* approaches to automated summarisation for identifying the most significant sentences of a document. How might the indications of several such criteria be combined together for sentence selection? [30%]

END OF QUESTION PAPER