The University of Sheffield

DEPARTMENT OF COMPUTER SCIENCE

Autumn Semester 2002-2003 **2 hours**

TEXT PROCESSING

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.

1. Answer **each** of the following short answer questions.

a) Explain each of the following terms and make clear the relations between them: **language**, **script**, **character**, **glyph**, **font**. Give examples of each. [20%]

b) Briefly describe the function of, and give an example of, each of the <!ELEMENT>, <!ATTLIST>+ and <!ENTITY> as found in SGML document type definitions. [20%]

c) Briefly describe the Unicode coding model, making clear the levels in the model, explaining the differences between, and the motivations for, UTF-8 and UTF-16. [20%]

d) Define each of the following terms as they are used in discussing Perl regular expressions and give an example of each: **metacharacters**, **metasymbols**, **anchors**, **quantifiers**, **back references**. [20%]

e) What would be an appropriate Perl data structure to hold an email folder, where an email folder is viewed as a numbered sequence of messages each of which consists of a collection of fields, such as **To**, **From**, **Subject**, etc.?
Write a Perl subroutine which, given a sender's name and a reference to the email folder data structure, will print out the subject field of all messages from the sender, one per line. [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 occurence of each of these characters in the corpus and, using these frequencies as estimates of the probability of occurence 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 does length does your code tree encoding save over a minimal fixed length binary character encoding for a 7 character alphabet? [10%]

   b. 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 be decoded,

```
<0,0,b><0,0,a><0,0,d><3,3,b><1,3,a><1,3,d><1,3,a><11,2,a>
```

assuming the encoding representation presented in the lectures? [10%]

3. a) In order for documents to be retrieved efficiently from an information retrieval system, information is stored in an inverted index. Explain what this means with an illustration of the data structure used to store the information in documents. Use the text below as two documents in the collection. [30%]

Document 1:
If your data is corrupted and your data can't be hashed,,

Document 2:
Then your system is corrupt, and your system is gonna crash!

b) Now illustrate the data structure with Documents 1 and 2 above, if we are developing a system where stop words are eliminated and stemming is used. You may assume the stemming algorithm behaves perfectly. Assume the following words are in the stop list: { if is and the can't be your} [30%]

c) Give an example of how document 1 might be represented in a vector space model if you assume that stemming is used and stop words listed in 1 (b) have been removed. Assume the only words in the universe are those in Documents 1 and 2 above. Indicate some ways in which the document representation might vary in a vector space model. [20%]

d) What information in a document would be represented in a Boolean system? Illustrate your answer with the text of Document 1. Assume stemming and stop-word removal. [5%]

e) Discuss the difference between how queries are compared to documents in a Boolean system versus a ranked retrieval system. [15%]

4. a) Discuss why stemming might be used in an IR system. Give an example of a rule that might be used in a stemmer. Discuss why it is difficult to implement a good stemmer. [15%]

b) Discuss the similarities and differences between the google search engine and one that you built in class. [15%]

c) Describe what is meant by the following phrases:

(i) text classification;

(ii) email filtering.

by describing in each case the task that the user might want to do. [20%]

d) Explain how the IR system that you built in class might be modified to perform each of the two tasks mentioned in part (c).
[20%]

e) Discuss what we mean by the precision and recall of a system for a given query.
[10%]

f) What does it mean to find the precision at 50% recall. [10%]

g) Explain what is meant by inverse document frequency? Why might document frequency be used in an information retrieval system? [10%]

END OF QUESTION PAPER