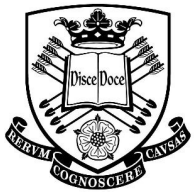

The University Of Sheffield.

Data Provided: None

DEPARTMENT OF COMPUTER SCIENCE **Autumn Semester 2007-08**

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) 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%]

   b) Explain what is meant by the *bag of words* model in text processing. Discuss the relevance of this model to work on *information retrieval*. [20%]

   c) Write down the Bayes Rule, and briefly explain what it means. State the key assumption that is made in a Naive Bayesian approach, and show how this allows the rule to be reformulated. [20%]

   d) Write a short Perl program that reads from a file (on `STDIN`), and which prints (to `STDOUT`) any line of the input file that contains the words "secret mission" (appearing together in this order) *irrespective* of their capitalisation (i.e. so that occurrences of e.g. "Secret Mission" and "SECret MiSsiON" should count as matches). Other lines from the input file should be ignored. [20%]

   e) Write a short Perl program that reads a file (on `STDIN`) containing an unsorted list of words (one word per line), and prints those words out (to `STDOUT`) in *alphabetically sorted order*. [20%]

2. a) Explain the differences between direct, transfer and interlingua approaches to Machine Translation. [30%]

   b) Describe the main features of the approach to statistical machine translation developed by the IBM research group. [40%]

   c) Do you believe that the approach cited in (b) is suitable for use with any pair of languages for which a large enough corpus of text is available? Justify your answer. [10%]

   d) Describe the approach used by the BLEU system for evaluation of Machine Translation systems. [20%]

3. a) Explain what is meant by the use of (i) a *stoplist* and (ii) *stemming* in the context of information retrieval. Discuss the potential benefits of using these methods for an information retrieval system. [20%]

   b) Explain what is meant by an *inverted index*, and why such indices are important in the context of information retrieval. Show how Documents 1 and 2 below would be stored in an inverted index, using stemming and a stoplist of your choice. [25%]

   Document 1: She sells sea shells on the sea shore.

   Document 2: Many shells and sea creatures are sold at the market.

   c) Calculate the similarity of Documents 1 and 2 using a measure of choice. Explain your measure and justify its choice. [25%]

Assume that we have a collection of documents which are stored in a *single* text file, in the manner shown in the box below. Note that the file includes XML-style tags, which mark the beginning and end of each document and also its identifier number.

```
<document docid=1>
Mary had a little lamb.
Its fleece was white as snow.
</document>
<document docid=2>
Little Miss Muffet, sat on her tuffet.
</document>
      ⋮
<document docid=N>
Mary, Mary, quite contrary.
How does your garden grow?
</document>
```

   d) Write a perl program which will read a file which is in this format and which will compute an inverted index. (NOTE: The requirement here is only that you should compute the information required for an inverted index and store it within an appropriate data structure. You do not need to print the index out to a file.) You may decide for yourself whether to record term frequency information or not. Let the definition of *term* be any (maximal) alphabetic sequence, i.e. do not try to accommodate terms that contain any non-alphabetic characters. Do not use a stoplist or stemming. [30%]

4. a) Explain what is meant by *summarisation*? List and explain a number of the ways in which different kinds of summary have conventionally been subdivided. [20%]

   b) Explain what is meant by 'deep' and 'shallow' approaches to automatic text summarisation, and discuss their differences. [10%]

   c) Discuss the criteria that have been used by *shallow* approaches to automated summarisation for identifying text segments containing significant information in a document. [20%]

   d) Describe the document summarisation approach of Kupiec et al. (1995). Your answer should address: (i) the nature of their training corpus, (ii) the features used as indicators of useful sentences, and (iii) how their system determines whether or not to include a sentence in the summary. [30%]

   e) Discuss alternative approaches to evaluating the performance of automatic summarisation systems. [20%]

END OF QUESTION PAPER