The University Of Sheffield.

Data Provided: None

PLEASE LEAVE THIS EXAM PAPER ON YOUR DESK.
DO NOT REMOVE IT FROM THE HALL.

DEPARTMENT OF COMPUTER SCIENCE **September 2018**

TEXT PROCESSING **2 hours**

Answer THREE questions.

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

1. a) Describe *manual* and *automatic* approaches to indexing for Information Retrieval. What are the advantages and disadvantages of the different approaches? [15%]

   b) Consider the following documents and query:

   Document 1: They sailed to the port for a good dinner and sailed home after.

   Document 2: Ruby port is very good after a meal, any meal.

   Query: good ports after dinner

   (i) Show how Documents 1 and 2 would be stored in an inverted index, using stemming and the following list of stopwords: {`a, and, any, for, is, the, they, to` } [20%]

   (ii) Compute the similarity between the query and each document using the cosine metric and using term frequency values for the term weights in the document vectors. [30%]

   (iii) Explain why the vector space model for Information Retrieval would not identify Document 2 as the top ranked document for the query. [15%]

   c) Define the precision and recall measures used in Information Retrieval. Explain why it is easier to compute precision than recall for web-based Information Retrieval systems such as Google or Bing. [20%]

2. a) Outline the algorithm for Huffman coding, i.e. for generating variable-length codes for a set of symbols, such as the letters of an alphabet. What does it mean to say that the codes produced are *prefix-free*, and why do they have this property? [30%]

   b) We want to compress a large corpus of text of the (fictitious) language *Fontele*. The writing script of Fontele employs only the six letters found in the language name (`f,o,n,t,e,l`) and the symbol ␣, used as a 'space' between words. Corpus analysis shows that the probabilities of these seven characters are as follows:

Symbol	Probability
`e`	0.3
`f`	0.04
`l`	0.26
`n`	0.2
`t`	0.04
`o`	0.1
␣	0.06

   (i) Show how to construct a Huffman code tree for Fontele, given the above probabilities for its characters. Use your code tree to assign a binary code for each character. [30%]

   (ii) Given the code you have generated in 2(b)(i), what is the average bits-per-character rate that you could expect to achieve, if the code was used to compress a large corpus of Fontele text? How does this compare to a minimal fixed length binary encoding of this character set? [20%]

   (iii) Use your code to encode the message "`telefone␣noel`" and show the resulting binary representation. Compare the average bits-per-character rate achieved for this message to the expected rate that you computed in 2(b)(ii), and suggest an explanation for any difference observed between the two values. [20%]

3. a) Explain the graded lexicon-based approach for Sentiment Analysis. Given the following sentences and opinion lexicon, apply this approach to classify *each* sentence in S1-S3 as **positive**, **negative** or **objective**. Show the final emotion score for each sentence and also how this score was generated. Give any general rules that you used to calculate this score as part of your answer. Explain these rules when they are applied. [40%]

Lexicon:

action-packed	4
boring	-3
beautiful	3
compelling	2
dull	-2
great	3
horrid	-4
mostly	1
tedious	-3
very	1

(S1) Tedious movie, mostly boring characters, DULL, DULL, DULL.

(S2) While not action-packed, the plot is very compelling and cinematography beautiful!

(S3) Not great but not horrid either.

b) A second approach to Sentiment Analysis is the corpus-based supervised learning approach.

(i) Explain how a Naive Bayes classifier can be trained and then used to predict the polarity class (positive or negative) of a subjective text. Give the mathematical formulation of the Naive Bayes classifier as part of your answer. [20%]

(ii) Suppose you are given the following set of labelled examples as training data:

Doc	Words	Class
1	Amazing movie, the perfect way to make a sequel.	Positive
2	Hypnotic, surrealist, and most of all, maybe the most beautiful movie of the year.	Positive
3	Beautiful film. Well-paced; never felt it was overly long.	Positive
4	Visually stunning and amazing. A bit long, perhaps, but never boring.	Positive
5	Great plot but bad acting. Too long, boring in the middle.	Negative
6	Very boring, not entertaining, too artsy, plot holes galore, too long.	Negative
7	Visually beautiful but way too long and the soundtrack was annoying	Negative

Using as features just the adjectives (underlined words in the examples), how would a Naive Bayes sentiment analyser trained on these examples classify the sentiment of the new, unseen text show below?

Doc	Words	Class
9	Beautiful sets but too long and sooo boring.	???

Show how you derived your answer. You may assume standard pre-processing is carried out, i.e. tokenisation, lowercasing and punctuation removal. You do not need to smooth feature counts.

[40%]

4. *Relation extraction* is one of the main tasks in the sub-area of text processing known as *information extraction*.

   a) Briefly explain what the task of relation extraction is and illustrate your answer with an example. [10%]

   b) Relation extraction is sometimes split into two sub-tasks, *relation detection* and *relation classification*. Explain what relation detection and relation classification are, making clear how they differ, and illustrate your answer with an example (you may use the same example as in the preceeding part, but are not required to do so). [10%]

   c) Various linguistic features of natural language make relation extraction hard. Identify **three** such features and give an example of each. [20%]

   d) Supervised learning approaches to relation extraction have been quite successful but have the drawback of requiring substantial amounts of manually annotated training data. Two approaches that have been devised to address this problem are the *distant supervision approach* to relation extraction and the *bootstrapping approach* to relation extraction.

      (i) Briefly explain how the *distant supervision approach* to relation extraction works, give an example of how it works and briefly identify the strengths and weaknesses of this approach. [30%]

      (ii) Briefly explain how the *bootstrapping approach* to relation extraction works, give an example of how it works and briefly identify the strengths and weaknesses of this approach. [30%]

END OF QUESTION PAPER