The University Of Sheffield.

Ancillary Material: Computer Answer Sheets

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

DEPARTMENT OF COMPUTER SCIENCE **September/October Resit 2019**

TEXT PROCESSING **2 hours**

Answer ALL questions.

The exam has two sections: Section A and Section B.

Section A consists of a collection of multiple-choice questions, which carry equal weight, and which together account for 25% of the overall credit for the exam. These questions should be answered on the Computer Answer Sheet provided.

Section B contains three questions that each account for 25% of the overall credit for the exam. In these questions, figures in square brackets indicate the percentage of available marks (out of 100) allocated to each part of a question. These questions should be answered in the Answer Book provided.

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

THIS PAGE IS BLANK.

SECTION A

INSTRUCTIONS: Please answer each of the following *multiple choice questions* by filling in the column(s) corresponding to your selected answer on the Computer Answer Sheet provided. Incorrect answers do **not** attract a negative mark. The questions carry equal weight.

1. Consider the sentence:

   After taking the tube from Camden to South Kensington John Handey visited the Victoria and Albert Museum.

   Which of the following best represents the standard labelling that the BIO approach adopts to labelling named entities for training a named entity recogniser, assuming the three entity types person, organisation and location.

   a) [O After taking the tube from] [B$_{LOC}$ Camden] [O to] [B$_{LOC}$ South Kensington] [[B$_{PER}$ John Handey] [O visited the] [B$_{ORG}$ [I$_{PER}$ Victoria] and [I$_{PER}$ Albert Museum]].

   b) [O After taking the tube from] [B$_{LOC}$ Camden] [O to] [B$_{LOC}$ South Kensington] [[B$_{PER}$ John Handey] [O visited the] [B$_{ORG}$ Victoria and Albert Museum].

   c) After taking the tube from [B Camden] to [B South Kensington] [I John Handey] visited the [O Victoria and Albert Museum].

   d) After$_O$ taking$_O$ the$_O$ tube$_O$ from$_O$ Camden$_{B_{LOC}}$ to$_O$ South$_{B_{LOC}}$ Kensington$_{I_{LOC}}$ John$_{B_{PER}}$ Handey$_{I_{PER}}$ visited$_O$ the$_O$ Victoria$_{B_{ORG}}$ and$_{I_{ORG}}$ Albert$_{I_{ORG}}$ Museum$_{I_{ORG}}$.

   e) After$_O$ taking$_O$ the$_O$ tube$_O$ from$_O$ Camden$_{B_{LOC}}$ to$_O$ South$_{B_{LOC}}$ Kensington$_{I_{LOC}}$ John$_{B_{PER}}$ Handey$_{I_{PER}}$ visited$_O$ the$_O$ Victoria$_{B_{PER}B_{ORG}}$ and$_{I_{ORG}}$ Albert$_{B_{PER}I_{ORG}}$ Museum$_{I_{ORG}}$.

2. **Shannon's Source Coding Theorem** can be informally stated as "a string of length N made of symbols from alphabet X cannot be compressed into fewer than $N \times H(X)$ bits without possible loss of information". Here H is the entropy of the probability distribution of X. Suppose we have a string $S =$ *fabdad* and know that $P(a) = 1/2$, $P(b) = 1/4$ and $P(d) = P(f) = 1/8$. Then the shortest lossless encoding of S is:

   a) 4.5 bits

   b) 10.5 bits

   c) 6 bits

   d) 9.5 bits

   e) 18 bits

3. **hexdump** is a Unix/Linux utility that prints out file contents as rows of 16 hexadecimal numbers, where each two hexadecimal digits represent one byte. By adding the `-C` flag then in addition to printing out file contents as hexadecimal numbers `hexdump` interprets the numbers as ASCII codes and prints out the corresponding characters. Suppose we run `hexdump` on a file and get this output:

```
> hexdump -C foo1
  00000000  52 65 64 20 51 75 65 65  6e 20 61 6e 64 20 57 68  |Red Queen and Wh|
  00000010  69 74 65 20 52 61 62 62  69 74 0a                 |ite Rabbit.|
  0000001b
```

   (the numbers at the left of each row are hexadecimal byte offsets into the file).

   We now run `hexdump -C` on a file containing just the 11 character string "`Turing Test`". What are the 11 hexadecimal ASCII codes we can expect to see?

   a) 53 75 72 69 6e 67 20 53 65 72 73

   b) 54 75 72 69 6e 67 20 54 65 73 74

   c) 74 75 72 69 6e 67 20 74 65 73 74

   d) 54 75 72 69 6e 66 20 54 65 73 74

   e) 54 75 72 69 6e 67 20 74 65 73 74

   (Hint: ASCII codes number sequentially from A-Z and a-z for both upper and lower case characters, though the ranges used for upper case and lower case are not adjacent.)

4. Suppose we are given:

   1. a static zero order probabilistic character model that tells us $P(a) = 0.4$, $P(b) = 0.25$ and $P(d) = .2$ and $P(f) = 0.15$; and
   2. the bit sequence $B = 111$.

   We are told B has been coded from source string S using **arithmetic coding**, where the ranges assigned to each character are ordered with the most probable character assigned the lowest range, the second most probable character assigned the next to lowest range and so on till the least probable character is assigned the highest range (i.e. as done in the examples in the Text Processing lectures). Then the source string S is:

   a) fad

   b) faa

   c) fab

   d) faf

   e) fafa

5. There is a document in our document collection that contains only the following text:

> "Information retrieval is the art of the possible. Most people – no matter what society they live in – are involved in some kind of information seeking. For example the Bushmen of the Kalahari need information on waterholes in the desert. Although they live in a non-technical society, their thirst brings about an information need, in this case to find a waterhole."

Assume that the term manipulation $capitalisation$ is applied during indexing, but not $stemming$ or $stoplisting$. Which of the following boolean retrieval queries would successfully retrieve this document?

a) (seeking AND water) BUT flowing

b) information AND extraction AND thirst

c) (society AND kalahari) BUT desert

d) kalahari AND NOT (seeking OR water)

e) waterhole AND (extraction OR retrieval)

6. In the Unicode model of text encoding UTF-8, UTF-16 and UTF-32 are all instances of **character encoding forms**. A character encoding form is:

a) a mapping from a set of sets of glyphs (visually differing representations of the same abstract character) to a set of characters.

b) a mapping from an abstract character repertoire to a set of integers (the codespace).

c) a mapping from a set of integers to a set of sequences of code units of specified width.

d) a mapping from a set of sequences of code units to a serialized sequence of bytes.

e) a mapping from a set of sets of glyphs (visually differing representations of the same abstract character) to a set of integers (the codespace)

7. The **basic multilingual plane** in Unicode is a codespace region:

a) of size 127 and with the same character code assignments as ASCII, covering those languages with the same alphabet as English.

b) of size $FFFF_{16}$ and covering most of the world's languages, but excluding the scripts of Eastern Asia (e.g. Chinese, Japanese (Kanji) and Korean).

c) of size $FFFF_{16}$ and covering almost all modern languages and a large number of symbols.

d) of size 256 and with the same character code assignments as ISO/IEC 8859, covering most the languages of Europe.

e) of size $10FFFF_{16}$ and covering almost all modern languages and a large number of symbols. as well as most historic writing systems .

8. You receive a message containing the following Huffman canonical code consisting of a sequence of symbols followed by a sequence of integers:

   ```
   (a,d,n,i,~,b,s) (0,3,1,1,2)
   ```

   You then receive the following message encoded using this code:

   ```
   1000111111101110111110011110110
   ```

   You decode the message to get:

   a) dina sabi

   b) dina basi

   c) ani basin

   d) nasi sabi

   e) bas dinab

9. Bing Liu's model for an opinion is expressed as a quintuple of the form:

$$(o_j, f_{jk}, X, h_i, Y)$$

   Here X and Y are placeholders standing for missing elements. The missing elements are:

   a) X is t_m is the time at which opinion holder h_i holds sentiment or feeling f_{jk} towards object o_j; Y is e_{jk} is the evidence opinion holder h_i puts forward for his feeling f_{jk} towards object o_j.

   b) X is so_{ijkl} the sentiment value of the opinion of opinion holder h_i towards feature f_{jk} of object o_j at time t_l; Y is the time t_l at which the opinion is held.

   c) X is a_{jkl} the k-th aspect of object o_j towards which opinion holder h_i holds sentiment or feeling f_{jk} at time t_l; Y is the time t_l at which the opinion is held.

   d) X is so_{ijkl} the sentiment value of the opinion of opinion holder h_i towards feature f_{jk} of object o_j with evidence e_{jk}; Y is the evidence opinion holder h_i puts forward for his sentiment so_{ijkl} towards object o_j.

   e) X is a_{jkl} the k-th aspect of object o_j towards which opinion holder h_i holds sentiment or feeling f_{jk} with evidence e_{jk}; Y is the evidence opinion holder h_i puts forward for his feeling f_{jk} towards object o_j.

10. LZ77 is a popular compression method, used in common compression utilities such as *gzip*. The following shows some possible LZ77 encoder output (assuming the encoding representation presented in the lectures of the Text Processing module):

$$\langle 0,0,b\rangle\langle 0,0,a\rangle\langle 0,0,d\rangle\langle 3,3,b\rangle\langle 1,2,a\rangle\langle 1,2,d\rangle\langle 1,2,a\rangle\langle 9,2,a\rangle$$

What output would be produced by decoding the above representation?

a) `badbbbbaaddaaaa`

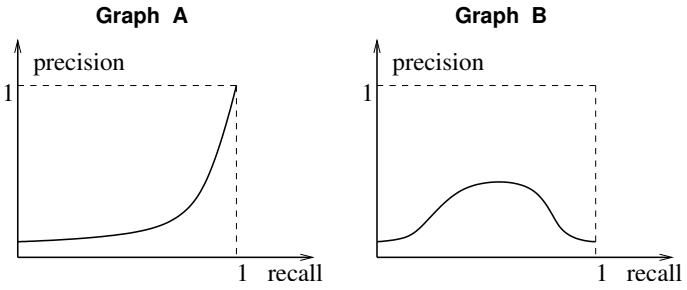
b) `badbadbbaaddabba`

c) `badbadbbbaaadddabba`

d) `badbadbaaddaba`

e) `badbadbdadddaddddddddda`

11. In regard to the precision and recall measures, as used in Information Retrieval, consider whether each of the following graphs is a *possible* precision/recall graph. Choose the option that correctly describes the situation.



a) Graph A *IS* a possible precision/recall graph.
Graph B *IS* a possible precision/recall graph.

b) Graph A *IS* a possible precision/recall graph.
Graph B *IS NOT* a possible precision/recall graph.

c) Graph A *IS NOT* a possible precision/recall graph.
Graph B *IS NOT* a possible precision/recall graph.

d) Graph A *IS NOT* a possible precision/recall graph.
Graph B *IS* a possible precision/recall graph.

12. Which of the following is **NOT** one of the comparative relation types identified by Bing Liu as occurring in comparative, as opposed to direct, opinions:

    a) superlative.

    b) contrastive.

    c) gradable.

    d) equative.

    e) non-gradable comparitive.

13. Consider the following opinion lexicon and the sentence (S1).

    Lexicon:

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

    (S1) *While not action-packed, the plot is very compelling and cinematography beautiful!*

    Using the graded lexicon-based approach to sentiment analysis, as presented in the lectures of the Text Processing module, what are the CNeg and CPos scores for (S1) using the given lexicon and any other relevant contextual rules?

    a) CNeg = -3 and CPos = 6

    b) CNeg = -4 and CPos = 7

    c) CNeg = -3 and CPos = 7

    d) CNeg = -4 and CPos = 5

    e) CNeg = -3 and CPos = 6

14. Which of the following are **subjective** statements?

S1 Barack Obama has said nachos are his favourite snack food.

S2 The new Panasonic combination microwave is brilliant for cooking quiche.

S3 I believe the England will win the next World Cup.

S4 My pizza arrived covered in jalapenos, which I had not ordered.

S5 The Blade Runner is a fabulous movie, stunning, well crafted with good action and good sci-fi story telling.

a) S1, S2

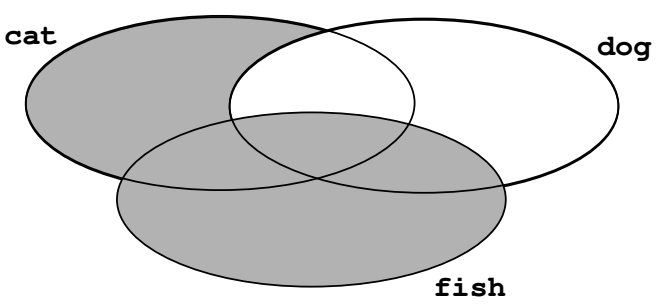
b) S2, S3, S5

c) S2, S4, S5

d) S3, S4, S5

e) S2, S5

15. The three ovals in the following diagram represent, for each word specified, the set of documents in our document collection that contain that word. Which of the boolean queries specified below would correctly pick out the set of documents corresponding to the shaded area of the diagram?



a) cat AND fish BUT dog

b) (dog OR NOT fish) and cat

c) (cat AND NOT dog) OR fish

d) cat AND NOT (dog OR fish)

e) (cat BUT dog) AND fish

SECTION B

16. 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 16(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 16(b)(ii), and suggest an explanation for any difference observed between the two values. [20%]

17. 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

a) Show the form that these documents would take after the following term manipulations were applied: *stemming*, *capitalisation* and *stop-word removal*. Use the stop-list {a, and, any, for, is, the, they, to}, and assume that punctuation is removed. [10%]

b) Show how Documents 1 and 2 (after the above term manipulations) would be represented using an *inverted index* that includes term frequency information. [10%]

c) Compute the similarity between the query and each document using the cosine metric, using term frequency values for the term weights in the document vectors. Which document is most similar to the query? [30%]

d) Two ranked retrieval systems – System 1 and System 2 – each return 10 documents for a given query. The following table shows, for each system, whether the document returned at each rank is relevant (✓) or not (×). It is known that there are 8 relevant documents for this query, within an overall collection of 100 documents.

Rank	System 1	System 2
1	×	×
2	✓	✓
3	×	×
4	×	✓
5	✓	×
6	✓	×
7	×	×
8	×	✓
9	×	×
10	✓	✓

(i) Explain the Precision, Recall and F-measure metrics, as used to evaluate information retrieval systems. Compute the performance of these two systems in terms of these three metrics. [15%]

(ii) Given that the two systems return the same number of relevant documents, suggest an alternative metric for evaluating their performance, and apply it to determine which system does better. [15%]

e) Describe *manual* and *automatic* approaches to indexing for Information Retrieval. What are the advantages and disadvantages of the different approaches? [20%]

18. a) One approach to Sentiment Analysis is the corpus-based supervised learning approach.

(i) Give the mathematical formulation of the Naive Bayes classifier and explain how such a classifier can be used to predict the polarity class (positive or negative) of a subjective text. [10%]

(ii) Explain how Naive Bayes classifier can be trained to carry out sentiment classification, given a corpus of sentences labelled as to whether they express positive or negative sentiment. [10%]

(iii) 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.

[30%]

QUESTION CONTINUED ON THE NEXT PAGE

b) 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. [25%]

(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. [25%]

END OF QUESTION PAPER