Student Number: □

The University of Melbourne, School of Computing & Information Systems

COMP90049 Knowledge Technologies

Final exam, Semester 2, 2018

Reading Time allowed: 15 minutes

Writing Time allowed: 2 hours

Number of pages: 8 including this page

Instructions to candidates:

This paper counts for 50% of your final grade.

Answer all questions on the ruled pages in the script book(s) provided, unless otherwise indicated.

There are 85 marks in total, or 1 mark per 1.5 minutes. Note that questions are not of equal value. All questions should be interpreted as referring to concepts given in this subject, whether or not it is explicitly stated.

No external materials may be used for this exam, but calculators are permitted (although not necessary). You may leave square roots and logarithms without integer solutions (like $\sqrt{2}$) unsimplified.

Unless otherwise indicated, you must show your working for each problem. Please indicate your final answers clearly for problems where you show intermediate steps.

Instructions to invigilators:

The students require script books.

Calculators are permitted; other materials are not authorised.

The examination paper should not leave the examination hall; this exam is to be held on record in the Baillieu Library.

Examiner's use only:

Q1	*Q2*	*Q3*	*Q4*	*Q5*	*Q6*	*Q7*	*Q8*	*Q9*	*Q10*	*Q11*	*Q12*
2	8	5	13	11	5	5	7	6	3	4	16

Part I : Text Processing **[28 marks in total]**

1. Describe two (or more) steps that we would typically perform in the "tokenisation" process for an Information Retrieval collection, according to the description from this subject. [4 marks]

2. It has been claimed that there are thre primary types of "information need" in a web search context: "informational", "navigational", and "transactional". Briefly describe each of these, optionally with the aid of an example. [4 marks]

3. In the context of Information Retrieval:

   (a) Explain how "data retrieval" is different to "information retrieval". [2 marks]

   (b) Give an example of a method or source of information that we might incorporate into our engine, that is specific to Web–scale information retrieval. [1 marks]

4. ...and other questions to add up to the marks as stated above. `:-)`

continued . . .

Part II: Data Mining/Machine Learning [57 marks in total]

For these questions, we have a training dataset comprised of the following 6 instances, 3 attributes, and two classes F and T, with a single test instance labelled with "?":

`ele`	`fed`	`aus`	CLASS
1	1	1	F
1	0	0	F
1	1	0	T
1	1	0	T
1	1	1	T
1	1	1	T
0	0	0	?

5. Classify the test instance according to the method of "Naive Bayes", as described in this subject. [4 marks]

6. Explain why "1-Nearest Neighbour" will give a different prediction to "3-Nearest Neighbour" on the given test instance. (It is not necessary to show your work for this question; an explanation which refers to the data should suffice.) [2 marks]

7. Consider the method of "Random Forests":

   (a) Briefly explain how a Random Forest would be constructed on the training data above. [4 marks]

   (b) Is there any evidence that a Random Forest would label the given test instance differently to a regular "Decision Tree"? [3 marks]

8. Exclude the CLASS labels from the dataset, and cluster all 7 instances using the method of "k-means". Apply the Manhattan Distance as a similarity measure; use the second (1,0,0) and third (1,1,0) instances as seeds. [4 marks]

9. ...and other questions to add up to the marks as stated above. `:-)`

end of exam