This examination paper consists of 5 printed pages each of which is identified by the Code Number (COMP377601)

This is an open book examination.
Any written or printed material is permitted.

School of Computing

May/June 2018

COMP3776

Data Mining and Text Analytics

Answer all 3 questions

Time allowed: 2 hours

TURN OVER FOR QUESTIONS

Question 1

I run a research lab with a network of 7 PCs, a shared file server, and a scanner. Some but not all of the PCs are connected to the file server; and some but not all of the PCs are connected to the scanner. Some of the PCs have previously had viruses. I asked my systems manager to examine each PC, to assess whether it is at high risk of catching a new virus. I want to add some new PCs, and I want the new PCs to have a low risk of virus infection. I hire you as a data-mining consultant, to examine my data and advise me.

The following csv file represents data about my PC lab; 1= yes, 0 = no:

```
RiskAssessment, InfectedPreviously, ScannerConnection, FileserverConnection
1,1,0,1
0,0,1,0
1,0,1,1
0,1,1,1
0,1,0,0
1,0,0,1
1,0,1,1
```

(a) Construct a J48-style decision tree from this training data, to predict risk of virus infection with 100% accuracy (tested on the training set).
Justify your choice of features for decision points.

[5 marks: 2 method, 2 justification, 1 full decision tree]

(b) Apply your decision tree to predict the risk of virus infection for:
(i) a new, unused, PC added to the Lab, connected to scanner and file-server

[1 mark]

(ii) a second-hand PC, previously virus-infected, added to the Lab stand-alone, NOT connected to scanner or file-server.

[1 mark]

(c) Prune your J48-style decision tree from (a) to give the smallest decision tree with at least 85% accuracy when tested on the training set. Draw or write down your pruned decision tree, and justify your choice of features for decision points.

[2 marks: pruned decision tree, justification]

(d) Apply your pruned decision tree to predict the risk of virus infection for:
(i) a new, unused, PC added to the Lab, connected to scanner and file-server

[1 mark]

(ii) a second-hand PC, previously virus-infected, added to the Lab stand-alone, NOT connected to scanner or file-server.

[1 mark]

(e) Apply the Apriori algorithm to find all association rules linking 2 or more features, with at least 90% accuracy and coverage of at least 4 instances.

[6 marks]

(f) I want new PCs in my lab to have a low risk of virus infection. Based on the above data-mining analysis, what recommendation can you make? State your evidence to support this recommendation.

[3 marks: 1 recommendation, 2 evidence]

[Question 1 total: 20 marks]

Question 2

The University of Leeds and South West Jiao Tong University (SWJTU) in Chengdu, China have set up a Joint School to teach a number of Bachelor degree programmes in engineering and computing to Chinese students at SWJTU, including a BSc Computer Science. The School of Computing at Leeds University is collating a set of lecture slides from our Leeds Computer Science teaching; these may be useful in teaching the new SWJTU students. Data Mining researchers can extract the text of the lecture slides and use this as a specialised data-set of English computer science text documents, for research as well as teaching.

In evaluating how well an advanced method or algorithm works for a given data mining or text analytics task, a standard approach is to first measure the accuracy of a “baseline” algorithm, and then measure the accuracy of the advanced algorithm, to compare the two. A baseline is a fast and simple algorithm; the advanced algorithm may be slower but should make less mistakes than the baseline, and so achieve a higher accuracy.

For each of the following tasks: (i) outline a baseline method, modelling each text document as a bag of words, and state the lexical resource required; (ii) give an example instance which would be incorrectly analysed by this baseline method; (iii) outline an alternative better method not restricted to individual-word lexical information, which should analyse the instance from (ii) correctly.

Write answers (i) (ii) and (iii) for the following two tasks:
(a) Named entity recognition to tag persons, organisations and locations in the text;
(b) Sentiment analysis to classify the overall sentiment of each lecture.

[10 marks, 5 each a,b: 2 baseline method, 1 example error, 2 better method]

(c) Kaggle.com is a website offering free access to data-sets, tutorials, software and data-mining competitions, for enthusiasts as well as researchers. For example, the Kaggle Titanic data-set includes information about passengers on the Titanic, a British luxury ocean liner advertised as “unsinkable”, which sank on her first ocean voyage in 1912. Kaggle account holders can download the data-set, and use it to train Data Mining classifiers. Below is a simplified sample of 6 instances from the Titanic training data file, saved as train.csv:

```
PassengerId,Survived,Pclass,Name,Sex,Age,Ticket,Fare,Cabin,Embarked
21,0,3,"Fynney, Mr. Joseph J",male,35,239865,26,,S
22,1,2,"Sloper, Mr. William Thompson",male,28,113788,35.5,A6,S
23,1,2,"Beesley, Mr. Lawrence",male,34,248698,13,D56,S
24,0,3,"Palsson, Miss. Torborg Danira",female,8,349909,21.075,,S
25,1,3,"Masselmani, Mrs. Fatima",female,23,2649,7.225,,C
26,1,3,"McGowan, Miss. Anna",female,15,330923,8.0292,,C
```

The first step in data-mining this data-set is data understanding, State three observations about properties of this sample data-set which are relevant in experiments with classifiers to predict the class Survived (whether the passenger survived after the ship sank)

[3 marks: 3 observations relevant to data-mining]

TURN OVER

(d) I want to try to compare results from a wide range of WEKA Classifier algorithms trained on this sample of instances. In WEKA Explorer, in the **Preprocess** tab, I **Open file** to load the data into WEKA; then go to the **Classify** tab, then click **Choose** to choose a classifier algorithm from a list of options. However, I find that many of the WEKA classifier options are greyed out and not available; for example, I cannot try NaiveBayesMultinomial classifier with this data-set. Why is this? State TWO ways to solve this problem to enable experiments to compare the results of a wide range of WEKA classifiers with this data-set.

[3 marks: 1 reason, 2 solutions to this problem]

(e) Assuming I have found a way to solve the problem in (d), I want to proceed with experiments to compare the results of applying different WEKA classifiers to this data-set. WEKA Explorer Classify tab offers four Test options, and I need to decide which to use in my experiments with this data-set:

Use training set
Supplied test set
Cross validation
Percentage split

Which of these Test options should I use for my experiments with evaluating WEKA classifiers on the above sample data-set of 6 instances? Justify your answer.

[4 marks: 1 best test option, 3 justification]

[question 2 total: 20 marks]

Question 3

Wales is one of the countries making up the United Kingdom of Great Britain and Northern Ireland. The official languages of Wales are Welsh and English. English is used in many official and scientific documents, but many Welsh people speak Welsh as their first or preferred language. The Welsh Assembly or government wants to promote use of the Welsh language terms for plants and animals found in Wales, by replacing English words for these plants and animals with Welsh terms in all English-language official government documents. For example, in Welsh National Park official documents, references to pine trees will be replaced with the Welsh word for "pine".

To achieve this, the government have acquired some text analytics data-set resources which could be useful: a list of plants and animals found in Wales, with both English and Welsh names; a large XML-format corpus of existing Welsh Assembly English-language documents; and an XML version of an English dictionary, the LDOCE Longman Dictionary of Contemporary English.

However, some English words are ambiguous, for example "pine" also has another sense "to be sad, for example over the loss of a loved one". It is important that only the plant or animal sense of each word in English official documents is replaced by the Welsh word. The Welsh Assembly needs a method to automatically classify ambiguous words in their English documents, to solve the problem of identifying plant or animal sense uses of such ambiguous words.

(a) Outline a supervised machine learning solution to the problem. Explain why it could be expensive to fund development of the necessary data-sets.

[4 marks: 3 marks for outline ML solution, 1 mark for cost explanation]

(b) Outline TWO unsupervised or semi-supervised approaches to the problem. Explain why these could be less expensive than your answer to (a).

[6 marks: each approach, 2 marks for method, 1 mark for cost explanation]

(c) Outline how to comparatively evaluate the two approaches from (b), in terms of precision and recall. What data-set does this require?

[6 marks: 4 marks for explanation of method, 2 marks for data-set]

(d) The Welsh Assembly has no text analytics resources or tools resources for the Welsh language. Suggest FOUR core resources or tools that should be developed or acquired, to enable the Welsh Assembly to apply general text analytics methods to analysis of Welsh text data in official documents; and give an example of what each resource or tool could be used for.

[4 marks: 1 mark for each Welsh text resource or tool]

[Question 3 total: 20 marks]

THE END.