Department of Computing & Information Systems

COMP30018/90049 Knowledge Technologies

Final exam, Semester 1, 2016

Date: 21 June, 2016

Time: 08:30am

Reading Time allowed: 15 minutes

Writing Time allowed: 2 hours

Number of pages: 5 including this page

Instructions to candidates:

For COMP30018, this paper counts for 60% of your final grade.
For COMP90049, this paper counts for 50% of your final grade.
Answer all questions on the ruled pages in the script book provided.
Note that questions are not of equal value.
There are 90 marks in total, or 3 marks per 4 minutes.

No external materials or calculators may be used for this exam. You should be able to compare fractions without a calculator; square roots and logarithms with integer solutions should be simplified, but non-integral forms (like $\sqrt{2}$) should be left 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.
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*
36	29	25

1. **Information Retrieval** **[36 marks in total]**

For this question, assume that we have collected the following 6 documents, numbered 1) through 6) below (the number is not part of the document text):

```
1) Australian Election.  Australian Federal Election.  Elections
Australia.  Federal Elections Australia.
2) election
3) .  @Election Federal Elections!  Federal Elections!!
#Federal #Election
4) vote federal elections, VOTING FEDERAL
5) Ausrlian FeDeRaL ElEcTiOn #Australia
6) .  @Vote @Australia vote federal federals elections
election vote vote vote!
```

(a) Describe two (or more) steps that we would typically perform in the Tokenisation process for an Information Retrieval collection, according to the discussion in this subject. [4 marks]

(b) Give a substitution-style Regular Expression which performs one of the tokenisation steps you have described in part (a), using syntax consistent with the lecture materials for this subject. [2 marks]

(c) According to a typical tokenisation process over the document collection above, give a representation (in either words or as a diagram) of an Inverted Index suitable for Boolean Querying, consistent with the methods described in this subject.
(Your index should preferably contain five terms.) [4 marks]

(d) What other data would need to be stored, so that we can perform Ranked Querying without reading the original documents, according to the methods described in this subject? [3 marks]

(e) Using the following TF-IDF model, along with the Cosine Similarity, calculate the ranking of the documents above for a Ranked Query Engine consistent with the description in this subject, when given the query `australia election`:

$$\begin{aligned} w_{d,t} &= f_{d,t} \\ w_{q,t} &= \frac{N}{f_t} \end{aligned}$$

(Note that you do not need to simplify irrational square roots to solve this problem, although you may need to simplify and compare fractions.) [9 marks]

continued ...

(f) If we had ascertained that, among the documents above, that document 5 was truly relevant for the query `australia election`:

i. What does "truly relevant" mean in this context, and how might we have discovered that this document was truly relevant? [2 marks]

ii. What is the $P@2$ of the ranked query engine from (e) on the given query? [1 marks]

(g) If we had instead used the following TF-IDF model and the query `australia election`, how do you expect the ranking to change compared to the model from (e)? Why is this the case?

$$w_{d,t} = \begin{cases} 1 + \log_2 f_{d,t} & \text{if } f_{d,t} > 0 \\ 0 & \text{otherwise} \end{cases}$$

$$w_{q,t} = \begin{cases} \log_2(\frac{N}{f_t}) & \text{if } f_{q,t} > 0 \\ 0 & \text{otherwise} \end{cases}$$

(Note that you do not need to calculate the steps of the model for this question.) [4 marks]

(h) Of the following improvements to a basic IR model, as described in this subject, choose **one** to discuss according to the following points: (i) Where in the standard four-part IR pipeline such an improvement would be integrated, (ii) A short description of changes to the system, which are required to incorporate the improvement, (iii) How such an improvement would impact a rank query engine on the collection given in this question (for example, for the query `australia election`).

- the Thresholding strategy for Accumulators
- Link Analysis
- Phrase Querying
- Spelling Correction
- Zoning

(Note that if you describe more than one improvement, we will only consider the final one, and ignore all earlier attempts, according to the page ordering in the script book(s).) [9 marks]

continued ...

2. Data Mining/Machine Learning **[29 marks in total]**

For this question, we have a training dataset comprised of the following 6 instances, 5 attributes, and two classes FACEBOOK and TWITTER, and a single test instance labelled with ?:

`ele`	`fed`	`aust`	`vot`	`ausr`	CLASS
Y	Y	Y	N	N	FACEBOOK
Y	N	N	N	N	FACEBOOK
Y	Y	N	N	N	TWITTER
Y	Y	N	Y	N	TWITTER
Y	Y	Y	N	Y	TWITTER
Y	Y	Y	Y	N	TWITTER
N	N	N	N	Y	?

(a) Classify the given test instance using the method of Naive Bayes, as described in this subject. [7 marks]

(b) According to the GINI coeffecient, which of the five attributes would be chosen as the root of a Decision Tree, as described in this subject, and why? How would such a Decision Tree classify the given test instance?
(Note that some steps in the model-building algorithm are unnecessary when answering the above questions; in particular, there is no need to build the entire tree.) [7 marks]

(c) The correct label of the given test instance is TWITTER.

i. What can you conclude about the above models? [2 marks]

ii. If we altered the dataset to remove the `ausr` attribute, and the test instance instead has `aust`=Y , would the classification of the above models on the test instance change? Why or why not? [4 marks]

(d) State **one** assumption made by the Naive Bayes method, as described in this subject. Is there any evidence that this subject is valid/invalid on the above dataset? [3 marks]

(e) Is there any evidence that using the method of Bagging, as described in this subject, would lead to a better Decision Tree algorithm on the above dataset? Why or why not? [3 marks]

(f) Give an example of a non-trivial Association Rule that can be constructed with respect to the above dataset. Calculate its Support and Confidence. [3 marks]

3. **Application** **[25 marks in total]**

The instances given in the dataset for Question 2 are derived from the documents given in the collection for Question 1. Based on this observation:

(a) The dataset given in the Question 2 is not suitable for building a k-Nearest Neighbour classifier, as we have discussed it in this subject. Why? What change can be made to the data to build such a classifier? [3 marks]

(b) Based on the change you made in part (a), the data becomes linearly separable. How might you observe this? What implications does this have if you are considering building a Support Vector Machine? [3 marks]

(c) What would be the significance of performing Classification on the dataset given in Question 2? What is a knowledge problem that you could solve with such a system? [4 marks]

(d) What would be the significance of performing association rule mining on the dataset from Question 2? What does it tell you about the collection from Question 1? [3 marks]

(e) Suggest another attribute from the collection in Question 1, which you could incorporate into one of the Data Mining systems in Question 2. Explain why it might improve the system. [3 marks]

(f) Observing whether the Classification models from Question 2 correctly label the given test instance is inadequate for formally evaluating models of this kind. Describe a method that would allow us to compare these models more formally; in particular, you should discuss (i) Where the dataset would come from, (ii) A suitable evaluation metric, (iii) A mechanism for using the dataset and the evaluation metric to compare the models. [9 marks]

end of exam