Department of Computing & Information Systems

# COMP90049 Knowledge Technologies

# Final exam, Semester 1, 2017

**Date:** 9 June, 2017

**Time:** 08:30am

**Reading Time allowed:** 15 minutes

**Writing Time allowed:** 2 hours

**Number of pages:** 7 including this page, and the blank page overleaf

**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.
There are 85 marks in total, or 1 mark per 1.4 minutes. Note that questions are not of equal value. All questions should be interpretted 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* |
|---|---|---|---|---|---|---|---|---|---|
| 8 | 21 | 5 | 4 | 4 | 11 | 13 | 7 | 4 | 8 |
| | | | | | | | | | |

(This page has been intentionally left blank.)

*end of exam*

1. (a). crawling: finding and downloading documents from the web, gather information

parsing: transform the data into canonical form. after some steps like stemming, case folder. turning documents into a list of tokens.

indexing: create a inverted ~~list~~ index of ~~each~~ tokens for each document.

query: When we got the query = tokenise (in a similar manner to our documents), we can apply query algorithm like TF-IDF to ~~apply~~ model produce a document ranking.

Week 7

(b)

(c) We only focus on the query term. So we ignore the term which doesn't appear in query.

$W_{q,Spain} = \frac{2}{2} = 1$

$W_{q,in} = \frac{2}{2} = 1$

$W_{q,hat} = \frac{2}{1} = 2$

$W_{q,t} = \langle 0,1,0,1,2,0,0,0 \rangle$

2. (a).

the → 1(2) → 2(2)

Spain → 1(1) → 2(3)

Cat → 1(3)

in → 1(1) → 2(3)

hat → 1(1)

rain → 2(1)

mainly → 2(1)

plains → 2(1)

DocA: $\langle 2,1,3,1,1,0,0,0 \rangle$

DocB: $\langle 2,3,0,3,0,1,1,1 \rangle$

$\therefore \cos(DocA, W_{q,t}) = \frac{\langle 2,1,3,1,1,0,0,0 \rangle \cdot \langle 0,1,0,1,2,0,0,0 \rangle}{|\langle \quad \rangle| \, |\langle \quad \rangle|}$

$= \frac{1+1+2}{\sqrt{2^2+1^2+3^2+1^2+1^2} \cdot \sqrt{1^2+1^2+1^2}}$

$= \frac{4}{\sqrt{16} \cdot \sqrt{6}} = \frac{4}{4\sqrt{6}} = \frac{1}{\sqrt{6}}$

(b)

| | Spain | in | hat |
|---|---|---|---|
| DocA | 1 | 1 | 1 |
| DocB | 1 | 1 | 0 |

DocA has three words, return DocA

$\cos(DocB, W_{q,t}) = \frac{\langle \quad \rangle \cdot \langle \quad \rangle}{|\langle \quad \rangle| \, |\langle \quad \rangle|}$

$= \frac{3+3+0}{\sqrt{3+4+9+9} \sqrt{3}}$

$= \frac{6}{\sqrt{3} \cdot \sqrt{25}} = \frac{6}{5\sqrt{6}} > \cos A$

$\therefore$ DocB will be higher rank.

DocA: $\langle spain \rangle \cap \langle in \rangle \cap \langle hat \rangle$

$= \langle 1 \rangle \cap \langle 1 \rangle \cap \langle 1 \rangle = 1$

DocB: ..

$= \langle 1 \rangle \cap \langle 1 \rangle \cap \langle 0 \rangle = 0.$

## Part I : Information Retrieval [38 marks in total]

1. Consider building a "ranked Information Retrieval engine" for Web documents:

   (a) Such an engine typically contains 4 main components: name them, and briefly (in a sentence or two) explain what the purpose of each one is. [6 marks]

   (b) Instead of the constantly–changing Web, consider building a ranked Information Retrieval engine for a mostly static database of newspaper articles. Give an example of a component that would be different. [2 marks]

2. For this question, consider the (very small) collection of documents, labelled A) and B) below (the label is not part of the document text):

   ```
   A) the Spain cat cat in the cat hat
   B) the rain in Spain mainly in the Spain plains in Spain
   ```

   and a query `Q) Spain in hat`

   (a) For a standard inverted index consistent with the lecture or workshop notation; give a representation (in words or as a diagram) of the "inverted lists" for the 8 terms in this collection. (There is no need to explicitly indicate the "search structure" or "mapping table".) [6 marks]

   (b) If we wish to apply the method of "Boolean querying" — assuming that the query is implicitly a conjunction of terms — describe the procedure by which the query engine would arrive at the result set {A}. [4 marks]

   (c) Determine which of the documents above would be returned higher (nearer to the top of the ranking) for a "ranked query engine", based on the following "TF-IDF model", suitably interpretted in the context of this subject:

   $$w_{d,t} = f_{d,t}$$
   $$w_{q,t} = \frac{N}{f_t}$$

   (Remember to show your work; there should be no need to simplify irrational square roots to solve this problem.) [7 marks]

continued ...

$w_{d,t} = 1 + \log_2 f_{d,t}$

$w_{q,t} = \log_2 (\frac{N}{f_t})$

$w_{q,\text{Spain}} = \log_2 1 = 0.$

$w_{q,in} = 0$

$w_{q,hot} = \log_2 2 = 1$

$\therefore w_{q,t} = \langle 0,0,0,0,1,0,0,0 \rangle$

$DocA = \langle 2, 1, 1+\log_2 3, 1, 1, 0, 0, 0 \rangle$

$DocB = \langle 2, 1+\log_2 3, 0, 1+\log_2 3, 0, 1, 1, 1 \rangle$

$\cos(A) = \frac{\langle \rangle \langle \rangle}{|\langle \rangle| |\langle \rangle|}$

$= \frac{1}{\sqrt{1} \cdot \sqrt{7 + (1+\log_2 3)^2}}$

$\cos B = \frac{\langle \rangle \cdot \langle \rangle}{|\langle \rangle| |\langle \rangle|} = 0$

$\therefore$ choose A.

(d) If we had instead used the following TF-IDF model, how do you expect the ranking to change compared to the model from (c)? 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]

3. Consider the Information Retrieval evaluation metric "Average Precision" (AP):

(a) Why is AP preferred over other metrics like Precision and Recall, in the context of evaluating a typical Information Retrieval engine? [2 marks]

(b) Explain the procedure by which AP is calculated; you may give an example if you think it will help your explanation. [3 marks]

$\surd\surd\times\surd\surd \quad \times\surd\times\surd\times$

$AP = \frac{1}{10}\left(\frac{1}{1} + \frac{2}{2} + \frac{3}{3} + \frac{4}{4} + \frac{5}{5} + \frac{6}{6} + \frac{7}{7} + \frac{8}{8} + \frac{9}{9} + \frac{10}{10}\right)$

$\frac{1}{6}\left(\frac{1}{1} + \frac{2}{2} + \frac{4}{4} + \frac{5}{5} + \frac{6}{7} + \frac{6}{9}\right)$

4. One possible extension to a ranked Information Retrieval engine is "link analysis", of which "PageRank" is an example.

(a) Define "link analysis", in the context of this subject. [1 marks]

(b) What is the main idea behind incorporating this information into the ranking? [1 marks]

(c) What are the two main components of the PageRank algorithm, and what user behaviours do they correspond to? [2 marks]

*continued ...*

5. $d(1,7) = \sqrt{(1-0)^2+(2-0)^2+(0-0)^2} = \sqrt{5}$

$d(2,7) = \sqrt{0^2+2^2+0^2} = \sqrt{4}$ ✓

$d(3,7) = \sqrt{3} = \sqrt{3}$ ✓

$d(4,7) = \sqrt{3^2} = \sqrt{9}$

$d(5,7) = \sqrt{1^2+1^2} = \sqrt{2}$ ✓

$d(6,7) = \sqrt{3^2} = \sqrt{9}$.

choose the top 3 nearest instance. 2, 3 and 5.

There 2 tomorrow. 1 yesterday

⇒ choose TOMORROW

6. 1). SVM will find a hyperplane to linearly seperate two classes, If it isn't linearly seperable, then it would apply a kernal function to make it linearly ~~seperatable~~ seperable

2) Week 12

Partition the training data based on the best line (hyperplane) that divide the positive instance of the class the we're looking for from negative instances.

**Part II: Data Mining/Machine Learning** **[47 marks in total]**

For Questions 5–8 in this section, we have a training dataset comprised of the following 6 instances, 3 attributes, and two classes YESTERDAY and TOMORROW, and a single test instance labelled with ?:

| | tfw | ftw | wtf | CLASS |
|---|---|---|---|---|
| 1 | 1 | 2 | 0 | YESTERDAY |
| 2 | 0 | 2 | 0 | YESTERDAY |
| 3 | 1 | 1 | 1 | TOMORROW |
| 4 | 0 | 0 | 3 | TOMORROW |
| 5 | 1 | 0 | 1 | TOMORROW |
| 6 | 0 | 3 | 0 | TOMORROW |
| 7 | 0 | 0 | 0 | ? |

5. Classify the test instance according to the (simple) "3-Nearest Neighbour" method, given the following "distance" metric:

$$D(a,b) = \sqrt{\sum_i (a_i - b_i)^2}$$

(Show your work; it should not be necessary to simplify irrational square roots to solve this problem.) [4 marks]

6. Consider applying the method of Support Vector Machines to classify this test instance:

   (a) Briefly explain (in a couple of sentences) the logic behind training a Support Vector Machine. You may use some logical simplifications (like "line" for "hyperplane"); you should aim to avoid the mathematical formulation unless you *really* understand it. [4 marks]

   (b) For the given instances above, we will find that there is no solution. Explain what this would look like, if we were to graph these instances. (You may include a diagram if you think it will help your explanation; it can be approximate — please don't attempt to graph more than 2 dimensions simultaneously.) [3 marks]

   data is not ~~linearly separated~~ seperable

   (c) There are two main alternatives for building an SVM when we discover that there is no solution: briefly explain them (one sentence each). [4 marks]

(1) Soft margin. relax the notion of linear separability by allowing some number of points on the wrong side of the line.

(2) Kernel function: use this function to map the current points into higher dimensions, and find a hyperplane.

*continued . . .*

7.(c) a) sore the attribute on values

(2) linearly scan these values, each time updating the count matrix and computing Gini index at points where class label changes

(3) choose the split position that has the Gini index.

| Few | 0 | 0 | 1 | 2 | 2 | 3 |
|---|---|---|---|---|---|---|
| class: | tomorrow | tomorrow | tomorrow | Yesterday | Yesterday | tomorrow. |
| | | | ↑ pointA | | ↑ pointB | |

for few ≤ 1, 3 tomorrow

$$GINI = 1 - \left(\frac{3}{3}\right)^2 - 0^2 = 0$$

few > 1,

$$1 - \left(\frac{2}{3}\right)^2 - \left(\frac{1}{3}\right)^2 = \frac{4}{9}$$

$$\therefore GINI_{split} = \frac{1}{2} \times 0 + \frac{1}{2} \times \frac{4}{9} = \frac{2}{9}$$

$$GINI = 1 - \left(\frac{3}{5}\right)^2 - \left(\frac{2}{5}\right)^2 = 1 - \frac{9}{25} - \frac{4}{25} = \frac{12}{25}$$

$$\therefore GINI_{split} = \frac{12}{25} \times \frac{5}{6} + 0 \times \frac{1}{6} = \frac{2}{5} > \frac{2}{9}$$

$\therefore$ split at A ☆

a) measure the impurity of a node. Maximum value of GINI when records are equally distributed among all classes. most impure. Minimum when records belong to 1 class. most pure. Choose the lowest GINI.



7. If we wished to build a "Decision Tree" to classify the given test instance:

   (a) Explain the logic behind the GINI coefficient ($G = 1 - \sum P(j)^2$), and how we can use it to build the tree. [2 marks]

   use this to find how the tree pure or not, if ↓, the more pure the tree is. the lowest number of GINI coefficient

   (b) If we treated these attributes as "categorical", which attribute would be placed at the root of the tree? Consequently, what would be the predicted label for the test instance? (You do not have to show your work; an explanation which refers to the data is sufficient.) [4 marks]

   (b) 0, 1, 2, 3. ~~ftw and wtf is the root~~; tomorrow a different. ftw, tomorrow

   (c) If we treated these attributes as "continuous", it is more difficult. Briefly demonstrate how we could determine the root attribute, working through the process for either (not both) of `ftw` or `wtf`? [4 marks]

   (d) The method of "Bagging" usually wouldn't help with the procedure from (c) — but in this case, it could. Explain what might happen to vastly simplify the process, referring to the dataset where necessary. [3 marks]

~~does~~ doesn't select (d) ⇒ ftw ≥2, <2, reduce the nodes, only works for the example because the training set is small.

8. The method of "Naive Bayes" that we discussed in this subject could only be applied to "categorical" attributes. Treat each number as a distinct value when answering this question.

   (a) To "train" a Naive Bayes model, we must estimate two types of probabilities. Name the two types. Give one example calculation for each, based on the training data above. [3 marks]

   (b) There is a single crucial factor which causes this Naive Bayes model to label the given test instance as TOMORROW. Explain what this is, referring to the data as necessary. (You do not need to work through the entire procedure, but if you are unsure, we will accept the working as partial credit.) [4 marks]

a) prior probabilities: $P(\text{Yesterday}) = \frac{2}{6} = \frac{1}{3}$

conditional probabilities $P(\text{ftw}=1 \mid \text{Yesterday}) = \frac{1}{2}$

Crucial factor.

b) For tomorrow: $P(T) \cdot P(\text{ftw}=0 \mid T) \cdot P(\text{ftw}=0 \mid T) \cdot P(\text{wtf}=0 \mid T)$

$= \frac{2}{3} \times \frac{1}{2} \times \frac{1}{2} \times \frac{1}{4} = \frac{1}{24}$ ⇒ tomorrow

*continued ...*

For Yesterday: $P(Y) \cdot P(\text{ftw}=0 \mid Y) \cdot P(\text{ftw}=0 \mid Y) \cdot P(\text{wtf}=0 \mid Y)$

$= \frac{1}{2} \times \frac{1}{2} \times \frac{e}{6} \times 1 = \frac{e}{6}$

| predict \ Actual | Y | N |
|---|---|---|
| Y | 40 | 10 |
| N | 20 | 30 |

9. Consider evaluating a Machine Learning classifier, such that:
We have 100 instances in our development data, of which 60 are actually labelled as Y, and the rest as N. Our classifier predicted that 40 of the actual Y instances were indeed Y and 10 of the actual N instances were Y.

   (a) Show the "Confusion Matrix" that summarises all of the classifications. [2 marks]

   (b) Find the "Accuracy" of this classifier. [1 marks]

   (c) Find the "Precision" (of class Y) for this classifier. [1 marks]

(b) Accuracy $= \frac{40+30}{100} = 0.7 = 70\%$

(c) Precision (Y) $= \frac{40}{40+10} = \frac{4}{5} = 80\%$

10. Recall the definition of "knowledge tasks" from this subject:

   (a) The process of "Clustering" seems to clearly fit the most of criteria of things that are knowledge tasks. Give an example of this. [3 marks]

   (b) "Classification", on the other hand, mostly doesn't fit these criteria. Explain why, and explain how we might end up with "knowledge" nonetheless. [3 marks]

   (c) An example of an "Association Rule" is {tfw}→{ftw}. Explain how such a rule encodes "knowledge". [2 marks]

a) knowledge task: outcome are not well-defined, clustering ~~class~~ group. Solution based on similarity, based on users.

b) concrete task.

*end of exam*