

# COMP90049 Knowledge Technologies

[illegible]

**Part I : Text Processing****[25 marks in total]**

1. **Regular Expressions:** Given the following “regular expression”:

```
\s*f?re{1,3}\w
```

For each of the following strings, indicate (with “yes” or “no” in your script book) whether the regular expression will match. [2 marks]

- (a) freedom
- (b) a really good book
- (c) sfrew
- (d) \tf\tfr\tfre\tfree

2. **Approximate String Search:** By referring to the definitions from this subject:

- (a) Explain the problem that we wish to solve in “approximate string search”, including one typical strategy and any necessary data structure(s). [2 marks]
- (b) Explain why “approximate string search” is a “Knowledge Technology” (or “knowledge task”). [2 marks]

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

A) it happens over and over and over again  
 B) over and under in and out  
 C) over over over over

and a query Q) over and out

- (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”.)  
[4 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 process the inverted index, to arrive at the result set {B}. (The **and** in the query is a term, not the Boolean operator AND.)  
[3 marks]
- (c) Determine the document ranking for a “ranked query engine”, based on the following “TF-IDF model”, suitably interpreted in the context of this subject:

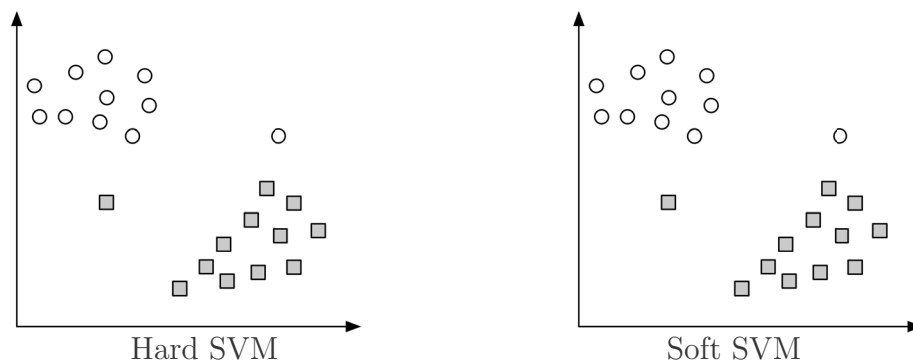
$$\begin{aligned} w_{d,t} &= f_{d,t} \\ w_{q,t} &= \frac{N}{f_t} \end{aligned}$$

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

4. The method of “accumulators” is often used when determining a document ranking in the context of Information Retrieval on the World Wide Web:
- (a) What information is stored in an accumulator? How is this used to build a document ranking?  
[2 marks]
  - (b) The full accumulator method is typically not employed: explain why.  
[1 marks]
  - (c) In this subject, two heuristic strategies for simplifying the accumulator method were discussed. Choose **one**, and briefly explain it. [3 marks]

**Part II: Data Mining/Machine Learning****[51 marks in total]****5. Support Vector Machines (SVMs)**

- (a) **This question should be answered by drawing directly on the diagrams below.** Draw the **maximum margin hyperplane** that would be generated by a Support Vector Machine (SVM), based on the dataset visualised in the diagrams below. On the left, using a **hard margin** SVM; on the right, using a **soft margin** SVM. [2 marks]



- (b) Do we get different support vectors from the hard margin and soft margin SVMs on the dataset in the graphs above? Why? [2 marks]
- (c) State two characteristics of the dataset above, which would lead us to believe that it is practical to learn an SVM here. [2 marks]
- (d) What are the main parameters in the “primal” formulation of an SVM? What are the main parameters in the “dual” formulation of an SVM, and how do they relate to “support vectors”? [3 marks]

**6. Association Rule Mining:****[4 marks]**

TIME	DRINK	BAKED GOOD
morning	coffee	pastry
morning	tea	pastry
evening	coffee	pastry
evening	tea	roll

Given the instances above, apply the “*a priori* principle” to find all of the Association Rules with at least  $\frac{1}{3}$  Support and at least  $\frac{3}{4}$  Confidence.

*continued ...*

7. **Hierarchical Clustering:** Given the proximity matrix below (in terms of **distance**) for five instances  $p_1 \dots p_5$ , build a hierarchical clustering by applying the method of agglomerative clustering, using the **complete link** updating heuristic. Be sure to show the proximity matrix at each step of the clustering, and the resulting **dendrogram**. [3 marks]

	$p_1$	$p_2$	$p_3$	$p_4$	$p_5$
$p_1$	0	1	5	9	10
$p_2$	1	0	3.5	8	7
$p_3$	5	3.5	0	3	4
$p_4$	9	8	3	0	0.5
$p_5$	10	7	4	0.5	0

8. **Naïve Bayes:** What is the primary assumption of the Naive Bayes Classifier? Explain why it is necessary, and why it is often untrue. [3 marks]
9. **Decision Tree:** Given the dataset below, with 8 instances, 3 binary (T or F) attributes  $a_1 \dots a_3$ , and the class LABEL:

$a_1$	$a_2$	$a_3$	LABEL
T	T	T	1
T	T	T	1
F	T	T	1
F	T	F	1
T	T	F	0
T	T	F	0
F	T	F	0
F	T	T	0

- (a) Use the Information Gain to determine which attribute should be placed at the root of the Decision Tree which would be constructed based on this dataset. Be sure to show your work; the following formula might help: [3 marks]

$$\Delta = I(\text{parent}) - \sum_j \frac{N_j}{N} I(j)$$

- (b) Does a decision tree exist, which can perfectly classify the given instances? If yes, draw that decision tree, otherwise, explain why not, by referring to the data. [2 marks]

continued ...

### 10. Neural Networks

What are steps involved in the “back propagation” algorithm for a neural network? You should explain the significance of the “learning rate parameter” in your answer. [5 marks]

### 11. Application Question (Long Response) [22 marks in total]

Imagine that, after completing Knowledge Technologies and graduating from the University of Melbourne, you are hired as a data scientist. For your first project, you are given a dataset from the university’s library, and your job is to build a *classification model*, as well as a *recommendation system*. As shown in Table 1, the dataset includes the list of books available in the library (columns) and the students who borrowed them (rows), and the ranking for each item (ranking value is between 0–5, 0 if the book was not borrowed and 1–5 indicates the student’s interest). The metadata for the books (e.g. titles) are not readily available to us, we just have the book IDs (e.g. Book #i). The dataset also includes the students’ field of study (in total there are 10 fields), which can be used for the classification task. Answer the following questions, considering that there are 500,000 students and 100,000 books in this dataset.

Table 1: Library data set.

Student ID	Book #1	...	Book #100,000	Label (Field of Study)
Student # 1	3	...	2	Computer science
Student # 2	5	...	0	Biology
⋮	⋮	⋮	⋮	⋮
Student # 500,000	1	...	4	Mathematic

[Question continues on the following page.]

**Task1: Classification Model** — The aim of this task is to use the label information (the most right column in the table) and build a classification model that classify students based on the list of books that they borrowed.

- (a) Consider the following supervised machine learning methods, and for each one, explain why it would be appropriate or inappropriate to use for this problem:
  - i. Support Vector Machines [2 marks]
  - ii.  $k$ -Nearest Neighbour [2 marks]
  - iii. Neural Networks [2 marks]
  - iv. Naive Bayes [2 marks]
- (b) Would “feature selection” be useful here? Explain why, by referring to a single machine learning method. [3 marks]
- (c) Explain how you would evaluate the effectiveness of your system: you should briefly describe an evaluation strategy and an evaluation metric that are suitable for this data. What might be an example of a baseline? [4 marks]

**Task2: Recommendation System** — Now assume that we do not have the label information and we would like to build a recommendation system for the library.

- (e) In this subject we studied content-based and collaborative filtering methods. Which recommendation method and similarity measure would you choose to address this problem? Justify your answer. [3 marks]
- (f) What are the challenges of building a recommendation system for datasets such as the library dataset? [2 marks]
- (g) As a recommendation system designer what goals do you need to consider? [2 marks]



THE UNIVERSITY OF  

---

MELBOURNE

## Library Course Work Collections

**Author/s:**

Computing and Information Systems

**Title:**

Knowledge Technologies, 2017, Semester 2, COMP90049

**Date:**

2017

**Persistent Link:**

<http://hdl.handle.net/11343/212959>

**File Description:**

COMP90049