The University of Melbourne

Department of Computing and Information Systems

COMP90049

Knowledge Technologies

Sample

Mid-semester Test

Length: This paper has 4 pages including this cover page.

Authorised materials: None

Calculators: Not permitted

Time: 50 minutes, with no reading time

Instructions to students: This exam is worth a total of 10 marks and counts for 10% of your final grade. Please answer all questions in the provided spaces on the test page. Please write your student ID in the space provided below. The test may not be removed from the test venue.

Student id:

Examiner's use only:

Q1	Q2	Q3	Q4	Q5	Q6

COMP90049 Knowledge Technologies
Mid-semester Test

Sample

Total marks: 10

Students must attempt all questions

1. Describe, with the aid of an example, the difference between "concrete tasks" and "knowledge tasks." (1 mark)

   - Concrete tasks: task is well-defined, "correct" solution, data is transformed in a mechanical way; no context for users, limited contribution to human knowledge; e.g. add two numbers, crack a code

   - Knowledge tasks: data is unreliable, outcome is not well-defined, no single "correct" solution (typically many "good" solutions); computers mediate between users and data, context (for the user) is critical, enhance human knowledge; e.g. relevant results for a web query, translation between languages

2. For the "regular expression":

   ```
   \S(he)a[t]{1,2}i?
   ```

   which of the following strings would the expression match (circle each)? (1 mark)

   (a) `eai`
   (b) `heatt`
   (c) `cheaters`
   (d) `space heating`

   - (c)

3. Use the "global edit distance," as shown in the lectures, to find the distance **from** (deleting) the string `led` **to** (inserting) the string `deed`, based on the following parameter vector:

$$[m, i, d, r] = [-3, 1, 4, 2]$$

Use as much of the matrix below as you need. (2 marks)

	ε	l	e	d	
ε	0 ← ↑↖	4 ← ↖	8 ← ↖	12	
d	1 ↑↖	2 ← ↑↖	6	5	
e	2 ↑↖	3 ↑↖	-1 ← ↑↖	3	
e	3 ↑↖	4 ↑	0 ↑↖	1	
d	4	5	1	-3	

- Edit distance: -3 (Operations: `rmim` or `rimm` or `irmm`)

4. For the "Soundex algorithm":

   (a) Apply it to the strings `carter` and `clinton`, using the following modified table: (1 mark)

`aeiouwy`	0
`bdgjlmnrvz`	1
`cfhkpqstx`	2

   - `carter`:
     - `c01201`
     - `c01201`
     - `c121`
     - `c121`
   - `clinton`
     - `c101201`
     - `c101201`
     - `c1121`
     - `c112`

   (b) Briefly describe how you might use your results in part (a) to perform "approximate matching" for the string `collins`. (1 mark)

   - Find the Soundex code for `collins`
   - Find the (global) edit distance (or 2-gram distance, or...) between the Soundex representations of `collins` and `carter`, as well as `collins` and `clinton`
   - Choose the string with the better distance out of `carter` and `clinton`

5. In the context of Information Retrieval, what does it mean for a document to be "relevant"? (1 mark)

   - The document meets the user's information needs; that is, the document contains information which allows the user to solve their problem

6. Given the following document collection:

   ```
   A: morning afternoon evening good
   B: good good good good vibrations
   ```

   Use the cosine similarity to determine the "document ranking" for the query `Q: good morning`, based on the following TF-IDF model: (3 marks)

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

   - For terms `afternoon`, `evening`, `good`, `morning`, `vibrations`, and the given model:
     - $A : \langle 1, 1, 1, 1, 0 \rangle$
     - $B : \langle 0, 0, 4, 0, 1 \rangle$
     - $Q : \langle 0, 0, 1, 2, 0 \rangle$
   - Similarities:

   $$
   \begin{aligned}
   S(A, Q) &= \frac{A \cdot Q}{|A||Q|} \\
   &= \frac{1 \times 0 + 1 \times 0 + 1 \times 1 + 1 \times 2 + 0 \times 0}{\sqrt{1^2 + 1^2 + 1^2 + 1^2}\sqrt{0^2 + 0^2 + 1^2 + 2^2 + 0^2}} \\
   &= \frac{3}{\sqrt{4}\sqrt{5}} \\
   S(B, Q) &= \frac{4}{\sqrt{17}\sqrt{5}}
   \end{aligned}
   $$

   - Similarity of A is larger, so it gets returned higher than B.
   - (No calculator? $3 > \sqrt{4}$ but $4 < \sqrt{17}$!)

— End of Test —

End of Test