Introduction to Data Science and Artificial Intelligence 2020/2021 Resit Questions

— Do not turn this page before the official start of the exam! —

First Name, Surname: _____

Student ID: _____

Program: Bachelor Data Science and Knowledge Engineering

Course code: KEN1110

Examiners: Dr. Rachel Cavill, Prof. Anna Wilbik, & Dr. Pietro Bonizzi

Date/time: Monday 25th January, 2021, 14.00-16.00h

Format: Closed book exam

Allowed aides: Pens, simple (non-programmable) calculator from the DKE-list of allowed calculators.

Instructions to students:

- The exam consists of 11 questions on 15 pages (excluding the 1 cover page(s)).
- Fill in your name and student ID number on each page, including the cover page.
- This exam consists of two parts. Both worth 50 points.
- Answer every question at the reserved space below the questions. If you run out of space, continue on the back side, and if needed, use the extra blank page.
- Ensure that you properly motivate your answers.
- Before answering the questions, please first read all the exam questions, and then make a plan to spend the two hours.
- Do not use red pens, and write in a readable way. Answers that cannot be read easily cannot be graded and may therefore lower your grade.
- You are not allowed to have a communication device within your reach, nor to wear or use a watch.
- You have to return all pages of the exam. You are not allowed to take any sheets, even blank, home.
- If you think a question is ambiguous, or even erroneous, and you cannot ask during the exam to clarify this, explain this in detail in the space reserved for the answer to the question.
- If you have not registered for the exam, your answers will not be graded, and thus handled as invalid.
- **Success!**

The following table will be filled by the examiner:

Question:	1	2	3	4	5	6	7	8	9	10	11	Total
Points:	10	10	10	10	10	7	10	8	10	5	10	100
Score:												

Part I - Artificial Intelligence

1. (10 points) Two students Alice and Bob are building agents to play the popular computer game 'Fortnite', a multi-player third person shooter.
   Alice has arranged for her agent to have access to the data in the main server, so her agent has information about the whole map and all enemies wherever they are positioned. Bob on the other hand is programming his agent to receive the same information a human player would receive in the game.

   a. Which feature of the environment will be different for Alice's and Bob's agents? Explain your answer, stating which value this feature will have for each student's agent.

   b. Alice decides that a simple reflex agent will be sufficient for her agent.
   Explain what a simple reflex agent is, and give a simple example of how this might work in this scenario.

c. Bob says that a simple reflex agent won't be appropriate for his agent.
Why would a simple reflex agent be ok for Alice, but not work for Bob?

d. When designing their agents Alice and Bob remember the acronym PEAS from their IDA course.
What does the A stand for in this acronym? Give an example appropriate to the task being performed by Alice and Bob's agents.

2. (10 points) Consider the following graph as a representation of travel times between 5 cities: Aura, Bumbletown, Copper Falls, Donken, Fulton

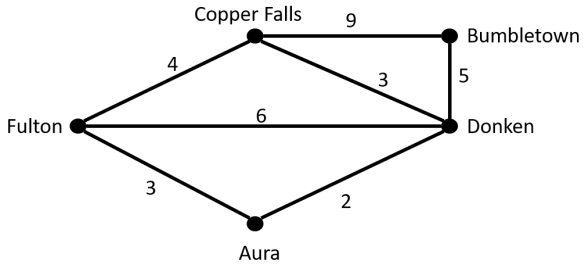


Figure 1: Travel time between 5 cities

a. Draw a search tree associated with this graph using breath first search for a route from Fulton to Bumbletown. Use the alphabetic order as direction preference. You may use the first letter of the city name as its representation.

b. What is the fastest route between Fulton to Bumbletown?

c. Which search technique will give result first? Include the search orders.

3. (10 points) a. Does $A \vee B \models A \wedge B$? Show truth tables.

b. Give an example of a statement which would correctly complete the following, showing a truth table.
$(A \Rightarrow B) \equiv$

c. For each term give an example of a statement using two variables A and B where that term applies;
i. Valid
ii. Unsatisfiable

4. (10 points) A small dataset containing 4 samples has measured 3 variables x, y and z, and we want to use them to predict the class c.

x	y	z	c
1	0	0	II
1	1	1	I
1	1	0	I
0	0	1	II

a. State the entropy of this dataset, explain your answer.

b. Calculate the information gain for each of the three variables.
Using $E(S) = -p^+ log_2 p^+ - p^- log_2 p^-$
$Gain(S, A) = E(S) - \sum_{v \in Values(A)} \frac{|S_V|}{|S|} E(S_V)$ and $S_V = \{s \in S | A(s) = V\}$

c. Draw a decision tree which correctly classifies all the samples in this dataset.

5. (10 points) a List the main steps performed to generate the list of postings from a collection C containing 200 documents D_i.

b. Five web pages have the following link structure and previous pagerank.

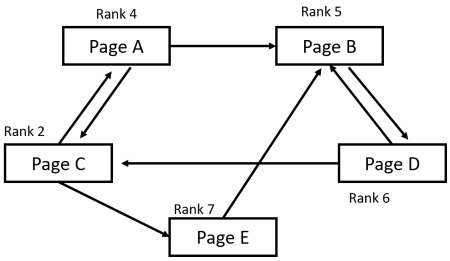


Figure 2: Link structure and previous pagerank of 5 web pages

What is the new rank of pages A and B after 1 iteration?

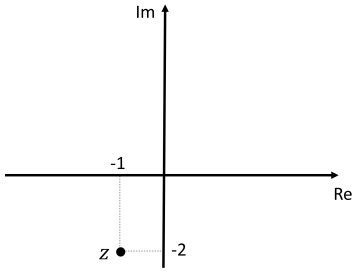
Part II - Data Science

6. (7 points) Given the following number: 01012021 (data)

   a. Think of any meaning which could be associated with this number (information), and how you could use it to perform a specific action (knowledge).

   b. Express the knowledge generated at the previous point into a rule which could be implemented into a knowledge based system.

7. (10 points) Given the complex number z represented below in the complex plane:



a. Express z both in algebraic ($z = a + ib$) and trigonometric form.

b. Given $w = i$, compute $\frac{z}{w}$

c. Find $x \in \mathbb{C}$ such that: $x + i = \bar{x}$

8. (8 points) Given the three sequences: $x = \{1, 2, -1, 0\}$, $y = \{1, 3, -1, 1\}$ and $z = \{1, -2, -1, 0\}$

   a. Compute the Euclidean distances $d(x, y)$, $d(x, z)$, and $d(y, z)$. What are the most similar time series?

   b. Given the following formula to measure the distance between two time series w and v

$$F(w, v) = \sum_{i=1}^{2} |w_i - v_i| + \max_{j=3,4} (|w_j - v_j|)$$

   Does this measure fulfill the properties of a distance measure for points A={1,2,3,1} and B={-1,0,2,1}? For the triangular inequality property consider point C={0,1,3,0}.

9. (10 points) In a physical experiment four different measurements y_k are obtained over four different x_k

x_k	0	1	1	2
y_k	1	1	2	3

The physicist who collected those is interested in fitting these measurements to a model of the form: $y_k = ax_k + b$

Compute the optimal values for a and b according to the chosen model and the data collected during the experiment.

10. (5 points)

a. Consider the experiment of picking a ball at random from a bag, noting the color of the ball, and putting it back. The bag has 25 balls, of which 3 are Red, 10 are Green, 7 are Orange, and 5 are Yellow. Complete the following table summarizing the probability distribution of this experiment:

Outcome value (color)	Probability	Cumulative probability
Red		
Green		
Orange		
Yellow		

b. Write a rule that would allow you to simulate this experiment with a computer.

11. (10 points) Given the following game:

		Player 2	
Player 1	Strategy	1	2
	1	1,-1	2,-2
	2	0,0	1,1

a. Does the game have a Nash equilibrium?

b. Is this a zero-sum game?