

EXAMINATIONS – 2019

TRIMESTER 1

COMP 423

INTELLIGENT AGENTS

Time Allowed: TWO HOURS

CLOSED BOOK

Permitted materials: Only silent non-programmable calculators or silent programmable calculators with their memories cleared are permitted in this examination.

Non-electronic foreign/English language dictionaries are permitted.

Instructions: Answer all questions.

The exam will be marked out of 100.

Questions

1. Text Representation [25]
2. Text Classification [20]
3. Text Clustering [25]
4. Web Intelligence [30]

1. Text Representation

(25 marks)

- (a) (10 marks) Very briefly describe two commonly used models for document representation. Which one is better in the context of text classification? Justify your answer.

- (b) (5 marks) Briefly describe how Word2Vec works and name its two different models.

- (c) (10 marks) In your opinion, what are the main limitations of the current text representation models? Suggest some solutions that may address these limitations.

2. Text Classification

(20 marks)

- (a) (10 marks) Suppose you are using Python to implement a very simple text classification system. You have a collection of documents as plain text files and each document starts with a manually assigned label. Write pseudocode to show the main steps that are needed to process the data so the data can be feed into a simple classifier such as K Nearest Neighbour.

- (b) (10 marks) Convolutional Neural Network(CNN) has been successfully used in many areas such as image classification. Very briefly explain what a convolutional layer does and what a pooling layer does in the context of image classification. Explain the main differences between CNN for image classification and CNN for text classification.

3. Text clustering

(25 marks)

- (a) (5 marks) What are the main differences between text classification and text clustering? Name one evaluation measure for classification and one for clustering, and briefly explain how they are calculated.
- (b) (10 marks) Suppose you want to design a system that automatically groups all the unique words in a book into different clusters, so that the words in the same cluster are semantically related. Describe how you are going to represent the data, what clustering algorithm you might use and why, and how you are going to evaluate the performance of your system.
- (c) (10 marks) Name at least five clustering algorithms. Choose one algorithm and very briefly outline how it works. Describe one application or task where the chosen algorithm would perform poorly, explaining the reasons why it would not be suitable.

4. Web Intelligence

(30 marks)

- (a) (10 marks) Suppose you want to design a house recommender system to help people to find the right house to buy. What are the challenges? Which approach is most suitable for house recommendation? Justify your answer.
- (b) (5 marks) What are the main approaches in query expansion?
- (c) (5 marks) State the main challenges in natural language understanding.
- (d) (10 marks) Describe a basic algorithm that Google uses to rank the search results. Explain its advantages and disadvantages.
