Calculators may be used in this examination provided they are not capable of being used to store alphabetical information other than hexadecimal numbers

UNIVERSITY OF BIRMINGHAM

School of Computer Science

Intelligent Data Analysis

Main Summer Examinations 2024

Time allowed: 2 hours

[Answer all questions]

Note

Answer ALL questions. Each question will be marked out of 20. The paper will be marked out of 60, which will be rescaled to a mark out of 100.

Question 1

(a) Wind directions in angles $\alpha_i \in [0, 2\pi)$ (clockwise from the North direction) have been measured daily at noon in 50 designated locations across greater London, indexed by $i = 1, 2, ..., 50$. Observations, conducted over the last 10 years, are collected in a data collection $\mathcal{D}$. Meteorologists would like to understand the data structure in $\mathcal{D}$ through Principal Component Analysis (PCA).

(i) Suggest a suitable representation of the data $\mathcal{D}$ so that the PCA methodology can be readily applied. **[7 marks]**

(ii) Some meteorologists are skeptical. They do not think that the PCA method is useful in this case. They argue that it is not reasonable to expect that a significant dimensionality reduction on $\mathcal{D}$ will be possible such that a large portion of the data variability will still be preserved in a low-dimensional subspace. Do you agree with them? Justify your answer. **[6 marks]**

(b) Formulate the *term-frequency-inverse-document-frequency* (TFIDF) representation of documents in the bag-of-words framework. Informally explain its meaning.

[7 marks]

Question 2

Algorithm 1 encodes the PageRank algorithm in pseudocode.

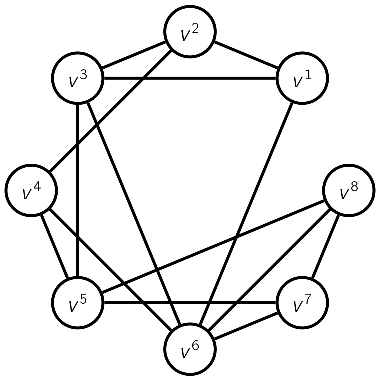
```
Algorithm 1: PageRank
   Data: $G$: A directed graph
   Data: $d \in [0,1] \subset \mathbb{R}$: A dumping parameter
   Data: $\mathrm{e} : e^i \in \mathbb{R}$: A rank sources vector
   Data: $\mathrm{p0} : p0^i \in \mathbb{R}$: PageRank scores initialization vector (initial prestiges)
   Data: $maxIter \in \mathbb{N}$: Number of maximum iterations
   Data: $tol \in \mathbb{R}$: Tolerance error
   Data: $Dangling$: Outedges to be assigned to any dangling nodes
   Result: $\mathrm{p} : p^i \in \mathbb{R}$: PageRank scores (final prestiges)
 1 /* Initialization:                                                              */
 2 /* - Set up the various parameters and build the normalized adjacency matrix    */
 3 $A \leftarrow adjacency(G)$ /* Retrieve adjacency matrix                        */
 4 if $A = \emptyset$ then
 5     $\mathrm{p} \leftarrow \emptyset$
 6 end if
 7 $A^{norm} \leftarrow A$
 8 $a^{norm}_{ij} \leftarrow \frac{a^{norm}_{ij}}{\sum_j a^{norm}_{ij}}$ /* Normalize */
 9 if $\sum_i p0^i \neq 1$ then
10     $p0^i \leftarrow \frac{p0^i}{\sum_i p0^i}$
11 end if
12 if $\sum_i e^i \neq 1$ then
13     $e^i \leftarrow \frac{e^i}{\sum_i e^i}$
14 end if
15 $Dangling_{ij} \leftarrow \frac{Dangling_{ij}}{\sum_j Dangling_{ij}}$ /* Normalize dangling */
16 $DanglingNodes \leftarrow v^i : \sum_j a_{ij} = 0$
17 for $v^i \in DanglingNodes$ do
18     $a^{norm}_{ij} \leftarrow Dangling_{ij}$
19 end for
20 /* Main loop:                                                                   */
21 $\mathrm{p} \leftarrow \mathrm{p0}$
22 for $iter = 1 : maxIter$ do
23     $\mathrm{p}^{last} \leftarrow \mathrm{p}$
24     $\mathrm{p} \leftarrow d(A^{norm^T} * \mathrm{p}^{last}) + (1-d)\mathrm{e}$
25     /* Check convergence, using L1 norm                                         */
26     if $\left\|\mathrm{p} - \mathrm{p}^{last}\right\|_1 < tol$ then
27         break
28     end if
29 end for
```

- Explain *what* normalization achieves and *why* it is necessary in PageRank. **[5 marks]**
- Replace the current lines 7 to 19 in the pseudocode so that dangling pages are dealt with using a dummy page, v^r such that:
  - Page v^r has a link to itself, and
  - Every dangling page is made to point to v^r

  Algorithm 1 and the new pseudocode may differ in the number of lines of code. **[15 marks]**

Question 3

Let $G = (V, E)$ be:



In each question below, do not simply answer with the final outcome. Explain and justify the steps taken as well as the intermediate calculations if any.

(a) What is the most central node according to degree centrality? **[10 marks]**

(b) What is the most central node according to closeness centrality? **[10 marks]**

This page intentionally left blank.

Do not complete the attendance slip, fill in the front of the answer book or turn over the question paper until you are told to do so

Important Reminders

- Coats/outwear should be placed in the designated area.
- Unauthorised materials (e.g. notes or Tippex) must be placed in the designated area.
- Check that you do not have any unauthorised materials with you (e.g. in your pockets, pencil case).
- Mobile phones and smart watches **must** be switched off and placed in the designated area or under your desk. They must not be left on your person or in your pockets.
- You are not permitted to use a mobile phone as a clock. If you have difficulty seeing a clock, please alert an Invigilator.
- You are not permitted to have writing on your hand, arm or other body part.
- Check that you do not have writing on your hand, arm or other body part – if you do, you must inform an Invigilator immediately
- Alert an Invigilator immediately if you find any unauthorised item upon you during the examination.

Any students found with non-permitted items upon their person during the examination, or who fail to comply with Examination rules may be subject to Student Conduct procedures.