Stellenbosch University

Industrial Engineering
DA774 & 874 (Data Analytics)
Additional Assignment 2023

Total: [???]
Deadline: 15 January 2024 at 23:59

Instructions

1. Students registered for Data Analytics 874 should answer all of the questions. Students registered for Data Analytics 774 should answer all questions excluding questions labeled as DA874.
2. You are allowed to consult any literature, provided that you include proper citations to such literature.
3. You are not allowed to make use of any Artificial Intelligence tools.
4. You are not allowed to discuss the questions with anyone.
5. You have to submit your own work.
6. Submit your **typed** answers on or before 15 January 2024, 23:59, via email to `engel@sun.ac.za`. Use as subject heading of your email `Data Analytics x74 Additional Assignment`, where you replace x with a 7 if your are a PGDip student and with an 8 if your are and MEng student. All submissions have to be in pdf format. Please do not submit in any other format, and make sure that your pdf document does not contain any errors.
7. Note that this is an optional assignment, and therefore no deadline extensions will be given, and no late submissions will be accepted.
8. Give answers in your own words. Do not copy from any other text. Marks will be given for insight and interpretation, not just facts. **Give motivations for all your answers**.
9. Name your file `????????DAx74.pdf`, where the question marks are replaced with your SU number and the `x` with a 7 or 8 depending on whether you are a PGDip or MEng student. Also provide at least your student number in the header of the first page of your pdf document.
10. Please note that your lecturers will be on leave during this period.
11. Note that if you submit the additional assignment, the mark obtained will replace the 60% mark of your post-block assingments. After inclusion of the 20% contribution from your pre-block assignment and the 20% contribution from the quizzes, the maximum mark that will be awarded is 50%.

12. Only students who have submitted all three post-block assignments, and who have not received a final mark yet for the module, qualify for the additional assignment.

Introduction

ArXiv is a free distribution service and open-access archive for nearly 2.5 million scholarly articles in the fields of (i) *physics*, (ii) *mathematics*, (iii) *quantitative biology* (iv), *computer science*, (v) *quantitative finance*, (vi) *statistics*, (vii) *electrical engineering and systems science*, and (viii) *economics*.

When publishing an article on ArXiv, an author must select the most applicable field and subject area. For example, an author can choose the field of *computer science* and the subject area *artificial intelligence*. Authors can also select multiple areas by cross-listing an article and choosing additional subject areas.

Much like data science, subject areas on ArXiv have expanded over the years, and definitions have evolved. Consequently, users now encounter challenges in finding articles of interest. ArXiv has enlisted your help in addressing the growing difficulty users face in finding relevant articles. Your task is to develop a new categorization scheme that can be retroactively applied to existing articles.

To develop the system, ArXiv has provided you with a database of articles, each described by a title and an abstract.

Ultimately you will have to convince ArXiv that the categories you suggested are meaningful and correct. It is therefore important that for all the questions you critically evaluate whether the categories found are meaningful and interpretable. For instance, if you propose a category *artificial intelligence* you must be able to explain what this category entails and how articles will be assigned to this category. For instance, *group 1 contains articles that can be classified as artificial intelligence*, articles in this category includes *expert systems, theorem proving, knowledge representation, planning and uncertainty but not machine learning.*

Section A: Clustering a small dataset

Total: [50]

Consider a small subsample of the dataset which consists of the 15 articles shown in Table 1. These 15 articles can be roughly categorised as (i) *technical aspects of clustering*, (ii) *clustering in medicine* and (iii) *clustering in agriculture*. In Section A, you need to cluster the documents provided in Table 1 into three groups by answering the questions below.

1. Using the tiles provided in Table 1, develop a vocabulary. The tokens in the vocabulary are defined as (i) words separated by whitespaces, (ii) ignoring the case of the word, (iii) treating hyphenated words as one word and (iv) excluding the tokens: "and", "be", "do", "in", "for", "from", "not", "of", "on", "the", "to", "with", "use", "using".
   (a) How many tokens are in the vocabulary? (1)
   (b) Provide a list of the vocabulary sorted alphabetically. (3)
2. Develop a count matrix, where each row represents an article and each column represents a token in the vocabulary. Remove all the tokens that only occur in a single article title.
   (a) What are the dimensions of the count matrix **before** the tokens that only appear in a single article title were removed? (1)
   (b) What are the dimensions of the count matrix **after** the tokens were removed? (1)
   (c) Provide the count matrix with clear labels and columns (2)
3. Calculate the Manhattan distance between each article based on the count matrix after the tokens that only appeared in a single article were removed.
   (a) Provide the proximity matrix (2)
4. Conduct hierarchical clustering with single linkage utilising the proximity matrix computed in the preceding question.
   (a) Represent the clustering results using a dendrogram, where the x-axis should contain the article number as provided in Table 1, and the y-axis should indicate the distance. (3)

Table 1: List of Articles

Article Number	Title
0	Research on K-value selection method of K-means clustering algorithm
1	Performance of K-means clustering algorithm with different distance metrics
2	Introduction to the K-means clustering algorithm based on the elbow method
3	Effect of distance metrics in determining k-value in K-means clustering using elbow and silhouette method
4	K-means clustering with incomplete data
5	Use of latent class analysis and K-means clustering to identify complex patient profiles
6	Covid-19 cases and deaths in southeast Asia clustering using K-means algorithm
7	K-means clustering of Covid-19 cases in Indonesia's provinces
8	Brain tumor segmentation using K-means clustering and deep learning with synthetic data augmentation for classification
9	Skin cancer detection from dermoscopic images using deep learning and fuzzy K-means clustering
10	Diagnosis of grape leaf diseases using automatic K-means clustering and machine learning
11	K-means algorithm for clustering system of plant seeds specialization areas in east Aceh
12	Plant disease detection and recognition using K-means clustering
13	Plant leaf recognition using texture features and semi-supervised spherical K-means clustering
14	Segmentation of leaf spots disease in apple plants using particle swarm optimization and K-means algorithm

(b) Explain how you can assign instances to three clusters using the dendrogram provided above. Redraw the dendrogram, clearly indicating which instances will be assigned to which clusters. (2)

(c) Do the three clusters obtained from the dendrogram successfully cluster the documents into the three categories: (i) *technical aspects of clustering*, (ii) *clustering in medicine* and (iii) *clustering in agriculture*? Motivate your answer (1)

5. Consider how you can improve the clustering results obtained above.

(a) Which of the following approaches do you believe will improve the clustering results the **most**? (1)

i. Additional or different preprocessing steps
ii. Using a different distance metrics
iii. Using a different clustering algorithm

(b) Motivate your answer given in the previous question by:

i. explaining why the previous approach provided poor results, (4)
ii. and explaining how the approach that you have selected will address the problem. (2)

6. One of your team members suggested that using a self-organising map (SOM) can improve the clustering results. Consider the **example** self-organising map provided in Table 2. This self-organising map consists of a square structure of 25 neurons where each neuron has three weights. Cluster the weights of the self-organising map using single linkage hierarchical clustering and Euclidean distance.

Table 2: SOM Weights

	0	1	2	3	4
0	[0.37 0.95 0.73]	[0.6 0.16 0.16]	[0.06 0.87 0.6]	[0.71 0.02 0.97]	[0.83 0.21 0.18]
1	[0.18 0.3 0.52]	[0.43 0.29 0.61]	[0.14 0.29 0.37]	[0.46 0.79 0.2]	[0.51 0.59 0.05]
2	[0.61 0.17 0.07]	[0.95 0.97 0.81]	[0.3 0.1 0.68]	[0.44 0.12 0.5]	[0.03 0.91 0.26]
3	[0.66 0.31 0.52]	[0.55 0.18 0.97]	[0.78 0.94 0.89]	[0.6 0.92 0.09]	[0.2 0.05 0.33]
4	[0.39 0.27 0.83]	[0.36 0.28 0.54]	[0.14 0.8 0.07]	[0.99 0.77 0.2]	[0.01 0.82 0.71]

(a) Represent the clustering results using a dendrogram, where the x-axis should contain the neuron position in the grid e.g. (0,0), and the y-axis should indicate the distance. (4)

(b) Using the dendrogram, cluster the neurons into three groups. Modify Table 2, by colouring each grid space according to the cluster assignment. (4)

(c) Based on the results obtained in the previous question, do you think the SOM was trained using the correct hyperparameters? Motivate your answer. (2)

7. Figure 1 shows the cluster assignment of the Self-Organizing Map (SOM) trained on the count matrix **after** the tokens were removed, where each color represents a different cluster. Figure 2 shows the SOM grid based on the values for each feature, where yellow indicates a high value, and purple indicates a low value.

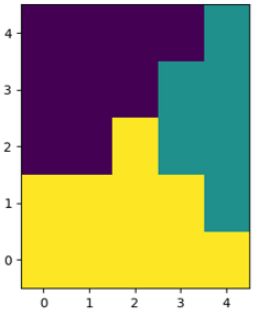
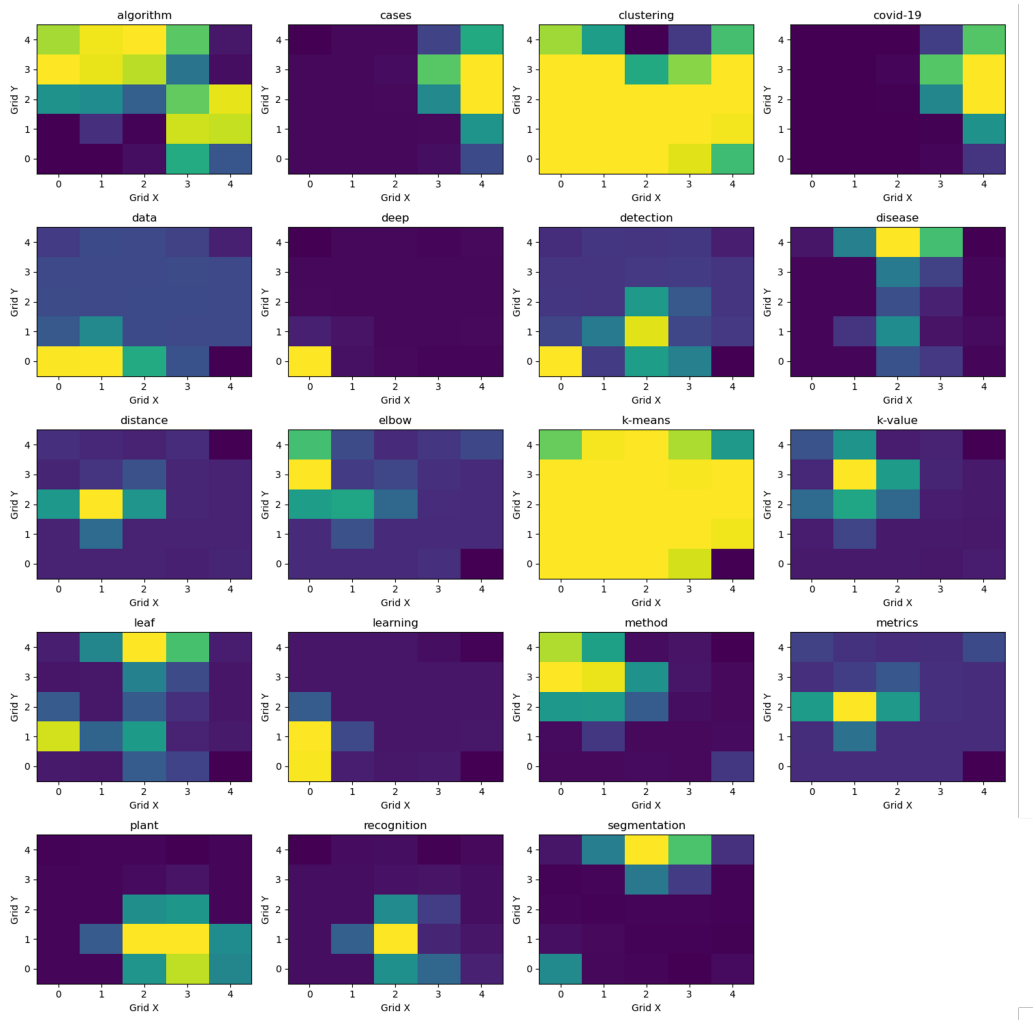


Figure 1: Cluster assignment of the SOM trained on the count matrix after token removal.



words.png

Figure 2: SOM grid based on feature values, where yellow indicates high values and purple indicates low values.

(a) Explain what steps you need to follow to generate the images in Figure 2? Assume that you have the SOM weights available. (4)

(b) Intepret the graph labelled k-means in Figure 2. What insights can you derive from the graph? (2)

(c) Intepret the graph labelled covid-19 and cases in Figure 2. Why do these graphs look similar? Motivate your answer (2)

(d) Intepret the graph labelled covid-19 and cases in Figure 2. Why do these graphs look similar? Motivate your answer (2)

(e) Provide the index of the best matching neuron for article 0. Explain your logic and state any assumptions that you gave made (2) (3)

(f) Are the results obtained using a SOM better or worse then the results obtained using hierarchical clustering? Motivate your answer (4)

Section B: Clustering a large data set

Total: [30]

Develop a system to cluster the "arxiv2017.csv" dataset. Your system should consist of the following components:

- Component A: converts text into a suitable numerical representation
- Component B: reduce the dimension of the numerical representation
- Component C: cluster the reduced numerical representations

While you should describe and motivate each component in your report, your system will be evaluated purely on the quality of the clusters obtained. Discuss at least:

1. the number of clusters obtained,
2. the quality of the clusters obtained measured using (i) *topic coherence* and (ii) *topic diversity*,
3. and the top five words associated with each cluster and the label that you have decided to provide to the cluster.

   Note that DA874 students should use a clustering technique that can assign an article to more than one cluster

Section C: Self-organising map

Total: [40]

Use a SOM to explore the "arxiv2017.csv" dataset with the prupouse of identifying suitable clusters.

While you should describe and motivate how you constructed your SOM, Section C will be evaluated purely on the quality of the SOM obtained as measured based on the visual representations provided. You should carefully decide how you can visualise your SOM to effectively communicate to the reader the (i) number of clusters (ii) what each cluster represents and (iii) the feature that distinguishes each cluster.

Note that DA874 students should develop custom visualisations to support their analysis. In other words, you are not allowed to use the visualisations provided by SOM packages. You need to design your own SOM visualisation that works well for text