Assignment 2

0. Use *Rmarkdown* to do the following tasks (2). Please note that the presentation of the document and the range of Rmarkdown features/functions used are matter.

1. Describe a real-world application that uses topic modelling and explain how the topic model works. (4)

2. Download the Twitter dataset (rdmTweets-201306.RData) from the course website and do the following. (8)

   a. Text cleaning: remove URLs, convert to lower case, and remove non-English letters or space.
   b. Count the frequency of words "data" and "mining".
   c. Plot the word cloud.
   d. Use a topic modelling algorithm to fit the Twitter data to 8 topics. Find the top 6 frequent terms (words) in each topic.

3. Provide a real-world example of a system or an application that utilises stream-data. In your example, explain the challenges faced by algorithms in analysing stream data and suggest some ideas to address those challenges (6)

4. Create a data stream of two dimensions data points. The data points will follow Gaussian distribution with 5% noise and belong to 4 clusters. Compare the performance of the following clustering methods in terms of precision, recall, and F1. (6)

   a. Use Reservoir sampling to sample 200 data points from 500 data points of the stream. Use K-means to cluster the points in the reservoir into 5 groups, and use 100 points from the stream to evaluate the performance of K-means.
   b. Use Windowing method to get 200 data points from 500 data points of the stream. Use K-means to cluster the points in the window into 5 groups, and use 100 points from the stream to evaluate the performance of K-means.
   c. Apply the D-Stream clustering method to 500 points from the stream with gridsize=0.1, and use 100 points from the stream to evaluate the performance of D-stream.

5. Explain a real-world application of geographical information system. (4)

6. Use spatial data analysis packages in R do the following tasks. (10)

   a. Draw a map of Australia where each city is represented as a dot. Highlight cities with population more than one million people. Map should have only the borders at country and state levels.
   b. Use the ***shapefile*** provided in the course website to draw a map of "South Australia". Keep all borders in the map. Use a colour palette to highlight the *statistical areas level 4 (SA4).*
   c. Create a ***spatial vector*** of "Greater Adelaide". Aggregate the polygons to draw a map that shows only the borders for *statistical areas level 3 (SA3).*
   d. For this point you need to check the data in "crimeCounts.csv" available in the course website.
      - Use the variable "SA3_NAME21" to obtain a ***spatial vector*** of "Salisbury".
      - Create a new attribute with the name *crimeCounts* containing the offence count (July 2022 – June 2023) for the suburbs in Salisbury ***spatial vector***.
      - Create a ***spatial raster*** to display the *crimeCounts* in Salisbury. Select a colour palette so that high *crimeCounts* are represented in red colour.
      - Show Salisbury suburb names and borders in the map.
   e. Create a html page with an interactive map containing the markers of your top 5 restaurants in Adelaide. Include in your report a screenshot of the interactive map. Upload the html as additional file in your submission.