MONASH University

Faculty of Information Technology

FIT3152 Data analytics – 2024: Assignment 3

Your task	• The objective of this assignment is to gain familiarity with Natural Language Processing and network analysis using R. • This is an individual assignment.
Value	• This assignment is worth **20%** of your total marks for the unit. • It has 36 marks in total.
Suggested Length	• 8 – 10 A4 pages (for your report) + extra pages as appendix for your R script. • Font size 11 or 12pt, single spacing.
Due Date	**11.55pm Thursday 6th June 2024**
Submission	You will submit 3 files: • Submit your report as a single PDF file. • Submit your video file as an mp4, m4v etc. • Submit your corpus as either a zipped folder or csv file on Moodle. *Please do not put your pdf and video in your zip file.* • Use the naming convention: *FirstnameSecondnameID.{pdf, zip, csv, mp4}* • Turnitin will be used for similarity checking all written submissions.
Generative AI Use	• In this assessment, you must not use generative artificial intelligence (AI) to generate any materials or content in relation to the assessment task.
Late Penalties	• 10% (3 mark) deduction per calendar day for up to one week. • Submissions more than 7 calendar days after the due date will receive a mark of zero (0) and no assessment feedback will be provided.

Instructions and data

In this assignment, you will create a corpus of documents and analyse the relationships between them, as well as the relationships between the important words used in these documents.

Background material for this assignment was covered in Weeks 10, 11, and 12. You are free to consult any other references, including those listed at the end of the document.

There are two options for compiling your written report:
(1) You can create your report using any word processor with your R code pasted in as machine-readable text as an appendix, and save as a pdf, or
(2) As an R Markdown document that contains the R code with the discussion/text interleaved. Render this as an HTML file and save as a pdf.

Your video report should be less than 100MB in size. You may need to reduce the resolution of your original recording to achieve this. Use a standard file format such as .mp4, or mov for submission.

Tasks

1. Collect a set of (machine-readable text) documents from an area of interest. For example, these could be a set of news stories, movie reviews, blogs, factual or creative writing. There is no restriction on the type of material you can choose although please avoid texts that might be offensive to people. As a guide, you should aim for the following:
   - Each document should be at least 100 words in length. Collect at least 15 documents.
   - Ideally, each document should cover one main topic and you should have at least 3 different topic areas in your collection of documents. Label each document as belonging to one of the topic areas in your corpus.
   - You can collect the documents as PDFs or as copied text from web-based articles or as text or other files.
   - Reference the source of your documents (URL or bibliographic citation (APA or Harvard style). **(2 Marks)**

2. Create your corpus by first converting each document into a text format. The type of original material you collect will determine the way you need to do this. For some formats you can simply copy and paste the text into an empty text file. For Word documents, HTML, and PDFs etc., you may find it simpler to create the text document using "export", or "save as" function in software.
   - Describe the process you follow for this step in your report.
   - Create your corpus using one of the methods covered in lecture videos and applied sessions. This could either be a folder of text files or a suitably formatted CSV file. Use suitable identifiers for your text file names or document IDs so that you can recognize the document in your clustering or network graphs. **(3 Marks)**

3. Follow the text processing steps covered in lecture videos and applied sessions to create your Document-Term Matrix (DTM).
   - As part of this process, you may need to make particular text transformations to either preserve key words, or to remove unwanted terms, for example, characters or artefacts from the original formatting. Describe any processing of this kind in your report or state why you did not need to do so.
   - Your DTM should contain approximately 20 tokens after you have removed sparse terms. You will need to do this by trial-and-error to get the right number of tokens.
   - Include your DTM as a table in the appendix of your report. **(3 Marks)**

4. Create a hierarchical clustering of your corpus and show this as a dendrogram.
   - Use the cosine distance between each document for clustering.
   - Identify which cluster each document belongs to.
   - Calculate the accuracy with which the clustering groups documents by topic.
   - Give a qualitative description of the quality of the clustering. **(4 Marks)**

5 Create a single-mode network showing the connections between the documents based on the number of shared terms.
   - To do this you will need to first calculate the connections between each document using the method shown in Week 12, or another method of your choice.

- What does this graph tell you about the relationship between the documents? Are there any groups in the data you can clearly identify? What are the most important (central) documents in the network?
- Improve your graph over the basic example given in Week 12 to highlight two or more interesting features of your data, such as the strength of connections, the relative importance of nodes, communities in the network. **(4 Marks)**

6 Repeat all the activities in Question 5, but now looking at the words (tokens). **(4 Marks)**

7 Create a bipartite (two-mode) network of your corpus, with document ID as one type of node and tokens as the other type of node.
- To do this you will need to transform your data into a suitable format.
- What does this graph tell you about the relationship between words and documents? Are there any groups in the data you can clearly identify?
- Improve your graph over the basic example given in Week 12 to highlight two or more interesting features of your data, such as the strength of connections, the relative importance of nodes, communities in the network. **(4 Marks)**

8 Write a brief report (suggested length 8 – 10 pages).
- Briefly summarise your results identifying important documents, tokens and groups within the corpus. Comment on the relative effectiveness of clustering versus social network analysis to identify important groups and relationships in the data.
- Can you suggest improvements to text processing used in this assignment to better discriminate between the documents studied? Describe briefly how and why these methods work. To do this you may want to refer to recent references on Natural Language Processing. You are not required to implement these improvements.
- Include your R script as an appendix. Use commenting in your R script, where appropriate, to help a reader understand your code. Alternatively combine working, comments and reporting in R Markdown. **(8 Marks)**

9 Record a short presentation using your smart phone, Zoom, or similar method. Your presentation should be approximately 5 minutes in length and summarise your main findings, as well as describing how you conducted your research and any assumptions made. Pay particular emphasis to your results for the investigative tasks. **(Submission Hurdle and 4 Marks)**

Software

It is expected that you will use R for your data analysis and graphics and tables. You are free to use any R packages you need but please list these in your report and include in your R code.

References

Statistical Analysis of Network Data with R, Kolaczyk, E. D., Csárdi, G. Springer 2020. Chapters 1 – 4
A User's Guide to Network Analysis in R, Luke, D. A. Springer 2015.
Network visualization with R, PolNet 2018 Workshop https://kateto.net/
Bipartite/Two-Mode Networks in igraph, Phil Murphy & Brendan Knapp. https://rpubs.com/
tm and iGraph package manuals.
Text Data Mining, Chengqing Zong, C., Xia, R., Zhang, J., Springer Nature, Singapore, 2021.