Graphical user interface, text

Description automatically generated

# Coversheet for submission of MSc coursework

DO NOT WRITE YOUR NAME ON YOUR WORK. Instead, please write your student number on this coversheet.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Student No. | 2 | 5 | 4 | 3 | 5 | 5 | 1 |

|  |  |
| --- | --- |
| Date Submitted: |  |
| Word Count: |  |

PLAGIARSIM DECLARATION:

This assignment is entirely my own work, and it adheres to the University of Bristol’s policy on plagiarism and academic integrity.

Quotations from secondary literature are indicated by the use of inverted commas around ALL such quotations AND by citations in the text or notes to the author concerned. ALL primary and secondary literature used in this piece of work is indicated in the bibliography placed at the end, and dependence upon ANY source used is indicated at the appropriate point in the text.

* I confirm that no sources have been used other than those stated.
* I confirm that I have not used artificial intelligence or chatbot software to create any writing or content in this assessment
* I confirm that I have not written this work in another language and translated it into English using translation tools
* I confirm that I have not used grammar checkers that suggest rewrites

**I understand that plagiarism, collusion, and cheating constitute misconduct and may result in disciplinary action being taken.**

Applied Health Data Science Summative Assessment:

Health Data Science Mini Project Report

*Please delete guidance text in italics before submission. The maximum word count for this report is 1,200 words. Sections 1 to 4 are worth equal marks.*

1. Data processing

*A description of the steps you undertook to pre-process the data (e.g. re-formatting, data cleaning).*

The whole data processing is divided into three steps.

The first step is to download the data. First, download the required PMIDs, save them as “pmids.xml”, and then download the corresponding article data according to the PMIDs in this file.

The second step is to filter out the PMID, the year of publication on PubMed and the article title and clear the xml tags from the article title. In this step, first set the names and paths of the input ("raw/article-data-\*.xml") and output ("clean/articles.tsv") files. Next, I wrote a loop to iterate through each file and finally store the cleaned data into a file called ‘articles.tsv’. This loop will filter out ‘PMID’, ‘Year’ under ‘PubDate’ and ‘ArticleTitle’, and then clear the xml tags in the article title. Then determine whether the article title exists or not, if it does not exist, the result will not be written to the output file.

The third step is to clean up the article titles using tidytext. There are several steps to this: splitting words, removing stop words and numbers, and reducing words to stems. Next these processed titles are regrouped by PMID and year and output as ‘processed\_titles.tsv’.

2. Data visualisation with description and justification of the approach

*One static visualisation produced in R and embedded in text describing the question you intended your plot to explore, the methods you used to visualise the data, the theoretical reasons you chose this approach, and your interpretation of what the plot shows.*

3. Reflection on ethical and governance considerations

*Imagine that you were going to request a data set from the Avon Longitudinal Study of Parents and Children (ALSPAC) cohort study (*[*https://www.bristol.ac.uk/alspac/*](https://www.bristol.ac.uk/alspac/) *) to explore how participants’ health changed over the course of the pandemic. This data set will contain individual-level sensitive medical data from people living in Bristol. What ethical or governance considerations should you take into account?*

4. Reflection on data management

*Now imagine that ALSPAC has approved your project to work with the data set you requested. How do you plan to store and manage the data in an appropriate way?*

References