MONASH University

Faculty of Information Technology

FIT3152 Data analytics – 2022: Assignment 1

Your task	• Analyse the activity, language use and social interactions of an on-line community using metadata and linguistic summary from a real on-line forum and submit a report of your findings. • This is an individual assignment.
Value	• This assignment is worth **20%** of your total marks for the unit. • It has 30 marks in total.
Suggested Length	• 6 – 8 A4 pages (for your report) + extra pages as appendix (for your code) • Font size 11 or 12pt, single spacing
Due Date	**11.55pm Friday 22nd April 2022**
Submission	• PDF file only. Naming convention: *FirstnameSecondnameID.pdf* • Via Moodle Assignment Submission. • Turnitin will be used for similarity checking of all submissions.
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

Submit the results of your analysis, answering the research questions and report anything else you discover of relevance. If you choose to analyse only a subset of your data, you should explain why.

You are expected to include at least one multivariate graphic summarising key results. You may also include simpler graphs and tables. Report any assumptions you've made in modelling, and include your R code as an appendix.

There are two options for compiling your report:
(1) You can submit a single pdf with R code pasted in as machine-readable text as an appendix, or
(2) As an R Markup document that contains the R code with the discussion/text interleaved. Render this as an HTML file and print off as a pdf and submit.

Regardless of which method you choose, you will submit a single pdf, and your R code will be machine readable text. We need to conform to this format as the university now requires all student submission to be processed by plagiarism detection software.

Submit your report as a single PDF with the file name ***FirstnameSecondnameID.pdf*** on Moodle.

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 document these in your report and include in your R code.

Questions

Activity, language use and social interactions in an on-line community. Analyse the metadata and linguistic summary from a real on-line forum and submit a report of your findings. Do the following:

(a) Analyse activity and language on the forum over time:

1. How active are participants over the longer term (that is, over months and/or years)? Are there periods where activity increases or decreases? Is there a trend over time? **(3 Marks)**
2. Looking at the linguistic variables, do the levels of these change over the duration of the forum? Is there a relationship between linguistic variables over the longer term? **(3 Marks)**

(b) Analyse the language used by threads:

We can think of threads as groups of participants posting on the same topic.

1. Using the relevant linguistic variables, is it possible to see whether or not particular threads are happier or more optimistic than other threads, or the forum in general, at different periods in time. **(3 Marks)**

(c) Analyse social networks online:

We can think of authors posting to the same thread at similar times (for example during the same month) as having a connection to each other, forming a social network. This is called a two-mode network. When an author posts to more than one network during the same time period their social network extends to include authors from both networks, and so on. We will cover social network analysis in Lecture 5.

1. Create a non-trivial social network of all authors who are posting over a particular time period. For example, over one month. To create this, your social network should include at least 30 authors, some of whom will have posted to multiple (2 or more) threads during this period. Your social network should be connected, although some authors may be disconnected from the main group. Present your result as a network graph. **(3 Marks)**
2. Identify the most important author in the social network you created. Looking at the language they use, can you observe any difference between them and other members of their social network? **(3 Marks)**

(d) Overall considerations:

- The quality and clarity of your reasoning and assumptions. **(3 Marks)**
- The strength of support for your findings. **(3 Marks)**
- The quality of your writing in general and communication of results. **(3 Marks)**
- The quality of your graphics throughout, including at least one high-quality multivariate graphic. **(3 Marks)**
- The quality of your R coding. **(3 Marks)**

Data

The data is contained in the file webforum.csv and consists of the metadata and linguistic analysis of posts over the years 2002 to 2011. You will each work with 20,000 posts, randomly selected from the original file. The linguistic analysis was conducted using Linguistic Inquiry and Word Count (LIWC), which assesses the prevalence of certain thoughts, feelings and motivations by calculating the proportion of key words used in communication. See http://liwc.wpengine.com/ for more information, including the language manual http://liwc.wpengine.com/wp-content/uploads/2015/11/LIWC2015_LanguageManual.pdf

Create your individual data as follows:

```
rm(list = ls())
set.seed(XXXXXXXX) # XXXXXXXX = your student ID
webforum <- read.csv("webforum.csv")
webforum <- webforum [sample(nrow(webforum), 20000), ] # 20000 rows
```

Data fields given. (see the language manual for more detail and examples):

Column	**Brief Descriptor**	**Column**	**Brief Descriptor**
ThreadID	Unique ID for each thread	we	"We, us, our" words
AuthorID	Unique ID for each author	you	"You" words
Date	Date	shehe	"She, her "him words
Time	Time	they	"They" words
WC	Word count of the text of the post	posemo	Expressing positive emotions
Analytic	Summary: Analytical thinking	negemo	Expressing negative emotions
Clout	Summary: Power, force, impact	anx	Indicating anxiety
Authentic	Summary: Authentic tone of voice	anger	Indicating anger
Tone	Summary: Emotional tone	sad	Indicating sadness
ppron	"I, we, you" words	focuspast	Expressing a focus on the past
i	"I, me, mine" words	focuspresent	Expressing a focus on the present
focusfuture	Expressing a focus on the future	focusfuture	Expressing a focus on the future

End.