FIT3152 Data analytics: Assignment 1

This assignment is worth 20% of your final marks in FIT3152.

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. Some starting points:

- Describe your data: How active are participants, and are there periods where this increases or decreases? Is there a trend over time?
- Looking at the linguistic variables, do these change over time? Is there a relationship between them?

b. Analyse the language used by groups. Some starting points:

- Threads indicate groups of participants communicating on the same topic. Describe the threads present in your data.
- By analysing the linguistic variables for all or some of the threads, is it possible to see a difference in the language used by these different groups?
- Does the language used within threads change over time?

c. Challenge: Social networks online. We can think of participants communicating on the same thread at the same time (for example during the same month) as forming a social network. When these participants also communicate on other threads, they extend their social network.

- Can you define, graph and describe the social network that exists at a particular point in time, for example over one month? How does this change in the following months? Note: you only need to analyse a short time period overall. We will cover social network analysis in Lecture 5.

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 are (see the language manual for more detail and examples):

Column	Brief Descriptor
ThreadID	Unique ID for each thread (a group of posts on a theme)
AuthorID	Unique ID for each author (-1 is anonymous)

Date	Date
Time	Time
WC	Word count of the text of the post
Analytic	LIWC Summary (analytical thinking)
Clout	LIWC Summary (power, force, impact)
Authentic	LIWC Summary (using an authentic tone of voice)
Tone	LIWC Summary (emotional tone)
ppron	LIWC (all personal pronouns)
i	LIWC ("I, me, mine" words) First person singular
we	LIWC ("We, us, our" words) First person plural
you	LIWC ("You" words) Second person
shehe	LIWC ("She, he, her, him" words) Third person singular
they	LIWC ("They" words) Third person plural
number	Quantities and ranks
affect	LIWC (expressing sentiment)
posemo	LIWC (Positive emotions)
negemo	LIWC (Negative emotions)
anx	Words indicating anxiety
anger	Words indicating anger
social	Words referring to social processes
family	Words referring to family
friend	Words referring to friends/friendship
leisure	Words referring to leisure
money	Words referring to money
relig	Words referring to religion
swear	Swear words
QMark	Question Mark (Punctuation)

Submission. Due 8th May 2020. Suggested length: 6–8 A4 pages + appendix.

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. 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.

Assessment criteria will include:

The quality of your analysis and description of your analytical process; Graphics and tables supporting your analysis; The quality of graphics used in the report. Justification of your findings and the degree of proof you can offer (for example statistical tests); Readability and quality of your written report; Insights gained from the data; Novelty of your approach.

Factors you should consider (starting points, not a complete list):

Techniques: summary/descriptive statistics, identification of important variables, networks, etc.
Major grouping variables: author, thread, date and/or time., or a combination of these.
Time window (days, weeks, months, years…); Subsets of the data to be analysed.
Graphics to communicate your analysis and insights (histograms, scatterplots, heatmaps, time series are some basic starting points, but see https://datavizproject.com/ for inspiration.