FIT3152 Data analytics: Assignment 1

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

The social and linguistic dynamics of an on-line community.

There is a theory in social science, that people adopt similar patterns of language use when they interact. (See, for example, Gonzales et al.[1]) This assignment will investigate whether this concept is present in an online forum, where participants communicate with each other via conversations online. You will investigate whether members who are communicating directly with each other (via threads) use similar language, which may be different to other members of the forum. You are also required to investigate whether the language used changes over time. For example, is the proportion of language expressing optimism different between groups, and/or does it change over time?

Using the metadata and linguistic summary from a real on-line forum you are required to analyse these questions using some or all of the data provided.

The data is contained in the file webforum.csv and consists of the metadata and linguistic analysis of 20,000 posts over the years 2002 to 2011. 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

Data fields are as follows (see the language manual for more detail and examples):

Column	Brief Descriptor
PostID	Unique ID for each post
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

[1] Gonzales, A., Hancock, J. T. and Pennebaker, (2010) J. *Language Style Matching as a Predictor of Social Dynamics in Small Groups.* Communication Research 37 (1), pp 3 – 19.

friend	Words referring to friends/friendship
work	Words referring to work
leisure	Words referring to leisure
home	Words referring to home
money	Words referring to money
relig	Words referring to religion
swear	Swear words
QMark	Question Mark (Punctuation)

You will work in pairs, and are free to choose your partner. Partners don't have to be in the same tutorial. Please email your group details to john.betts@monash.edu (Clayton students) or suleman.khan@monash.edu (Malaysia students) before the 23rd March.

Submission. Due 28th April 2018. Suggested length: 8–10 A4 pages

Submit the results of your analysis, answering the research question and outlining anything else you discover of relevance. If you choose to analyse only a subset of the 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. Give the approximate contribution by each member using the table below if required. Submit your report as *firstname1-firstname2.pdf* on Moodle.

Contribution by each member:

Group members will be awarded a proportion of the final mark based on the estimate of each member's contribution. If you contributed equally then you don't need to submit a table.

Member/Task	**Member A**	**Member B**	**Total**
Preliminary analysis			100%
R research and coding			100%
Preparation of graphics			100%
Analysis of results			100%
Writing up the report			100%

Software

It is expected that you will use R for your data analysis and construction of graphics and tables. You are free to download and use any R packages you need but please document these in your report. You may also want to use a graphics software to annotate and refine your plots.

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: data summary, statistical methods, networks, classification 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.
Graphical devices to communicate your analysis and insights.
Preliminary descriptive analysis.