ISM Honours
Information and Knowledge in Organisations
Data Analysis
Assignment 2: Data Collection and Text Mining

For this assignment, building on the first, you are required to produce a complete and professional data analysis report on the spread of the Coronavirus in South Africa over the past 6 months. In contrast to the first report, this second report needs to focus on the responses by various media agencies in South Africa. This analysis should come in the form of an analysis of social media posts on Twitter. Elements to consider include: topics covered, sentiment in general (either with specific lexicons for relevant keywords or other sentiment lexicons/dictionaries), sentiment of specific topics, frequency of posts, network interaction metrics, etc.

You are encouraged to identify (with justification) a set of relevant media agencies and collect a sample of their tweets and the interactions with these tweets and perform the analyses with this data, comparing these media agencies. For resources on data collection in this regard see:

- https://compsocialscience.github.io/summer-institute/2019/materials/day2-digital-trace-data/apis/rmarkdown/Application_Programming_interfaces.html

A note on the Twitter API: there are 2 versions, API V1 and API V2. At present V2 (which provides full historical search data) is restricted to academic purposes and requires an extensive application and evaluation process, beyond what it possible in this assignment. For this reason, you will need to use V1 through a wrapper such as rtweet (see https://cran.r-project.org/web/packages/rtweet/vignettes/intro.html for more information). For topic searching, this limits you to the last two weeks. To circumvent this, to identify older data, you will need to get the timelines of the accounts you identify. This will enable you to collect their latest 3200 tweets. To collect data beyond this (i.e., older data or interactions) you'll need to use some creativity and consider other possibilities.

Requirements:

Minimum requirements:
Identify a set of twitter accounts for media agencies in South Africa, extract their latest 3200 tweets, identify relevant tweets from this set, for those tweets collect relevant interaction data (likes, retweets, and accounts retweeted etc.). With this dataset, provide a descriptive account of 1) how reporting has progressed over time, 2) what topics have been reported on, 3) the tonality/sentiment of this reporting (choosing the most appropriate lexicon in your view), 4) the tonality/sentiment for specific topics, 5) how each of these differ by a) time and b) media agency.

Additional requirements:
In addition to the minimum requirements, to provide further insight, you are expected to identify additional relevant sources of data that might further contribute to our understanding of the twitter conversations in South Africa surrounding the pandemic over the last 6 months and conduct appropriate analyses with the sources. These analyses may be similar to the primary analyses or, optimally, they may go beyond the minimum requirements. For data, you might consider alternative data from twitter that is relevant, or you might consider if there is another

platform or set of platforms that provide relevant data (for instance media reporting or other discussion fora).

Resources for text analysis:

- Text Mining with R https://www.tidytextmining.com
- Text as Data https://cbail.github.io/textasdata/Text_as_Data.html
- sentimentr package: https://cran.r-project.org/web/packages/sentimentr/sentimentr.pdf
- vader package https://cran.r-project.org/web/packages/vader/vader.pdf
- udpipe package: https://cran.r-project.org/web/packages/udpipe/udpipe.pdf
- LDA Tuning: https://cran.r-project.org/web/packages/ldatuning/vignettes/topics.html

Expectations:

This is simultaneously an exercise in data collection, text mining, data wrangling, visualisation, and reporting. It is expected that your reports will include both quantitative numeric analyses as well as text analysis. Emphasis is placed equally on the description, appropriateness, accuracy, sophistication, and presentation of the analyses, as well as the description and presentation of the outputs of these analyses (both visual and quantitative).

Your reports need to include all necessary background information, figures, analyses, and statistics. Alongside the reports, you need to provide access to your analysis scripts or Rmds, either by uploading these alongside the reports or by providing access to a repository (e.g., gitlab, github). A PDF file of your report (with an optional zip file or link to a repository) should be uploaded to the supplied link on Sunlearn.

Assessment:

You will be assessed on:
- The general quality of your R code (15%)
- The data collection and cleaning for the base requirements (20%)
- The data analysis for the base requirements (30%)
- The presentation (structure, writeup, visualisation) of the report (15%)

- The data collection and cleaning for the additional requirements (10%)
- The data analysis for the additional requirements (10%)

The assignment is to be completed in pairs.

Due Date: Friday 9 July 2021: 17h00.