UNIVERSITY OF CAPE TOWN
DEPARTMENT OF STATISTICAL SCIENCES
STA5073Z – Data Science for Industry 2023

Assignment 2
A descriptive analysis of SONA speeches 1994–2023

This project must be done in groups of two or three
Due date 16 October 2023

Problem background

The State of the Nation Address of the President of South Africa (SONA) is an annual event in which the President of South Africa reports on the status of the nation, normally to the resumption of a joint sitting of Parliament. You have been provided with full text of all State of the Nation Address (SONA) speeches, from 1994 through to 2023 (sourced from https://www.gov.za/state-nation-address). In years that elections took place, a State of the Nation Address happens twice, once before and again after the election.

Assignment objectives

Your overall goal for this project is to conduct a descriptive analysis of the content of the speeches using sentiment analysis and topic modelling, and to write a short scientific paper (word limit: 4500) that you will host on your own website. Specific learning objectives (with assessment weights in brackets) are:

Analytical skills:

1. To be able to work with (i.e. read in, parse, clean, manipulate) text data. (10)
2. To be able to implement sentiment analysis techniques and correctly interpret results. (20)
3. To be able to implement latent Dirichlet allocation (LDA) for topic modelling and correctly interpret results. (20)
4. To be able to identify trends on these metrics (sentiment, topics) over time in the speeches (10).

Workflow skills:

1. To be able to write up your work in the format of a short scientific paper. Please note the word limit of 4500. (20)
2. To be able to host your paper on a GitHub Pages website, linked to your GitHub repository that you push to programmatically (i.e. not by dragging and dropping files). (10)

3. To be able to work collaboratively on GitHub, for example with different group members working on the same file, using branches, using pull requests. (10)

4. To experiment with the use of a large language model such as ChatGPT to assist with the assignment and to critically assess its ability to do so (in terms of what did it do well or badly; what prompts if any were needed to improve its performance, etc). (20)

Submission guidelines

One member of each group must submit:

1. A link to a GitHub Pages website, where one page contains your scientific paper and another page contains your description of what you used the LLM for, how you used it, and a critical reflection on the performance of the LLM and what, if anything, you have learned about the use of LLMs for assisting with data science work such as this project. You may include a "landing" page and extra webpages for appendices and supplementary material if you have any.

2. A link to a GitHub repository from which your website was built. This should contain the html source files for your website as well as all Quarto markdown (.qmd) files that contain your written text and code.

3. On Vula, please submit three files, using your **STUDENT ID as the files' names** – e.g. `ABCXYZ001.txt`, `ABCXYZ001.html` and `ABCXYZ001.qmd`, where

   (a) `ABCXYZ001.txt` contains the links to the GitHub Pages website and GitHub repository.

   (b) `ABCXYZ001.html` is your rendered scientific paper in html format. This should be the same as what is on your webpage. Do not include appendices, etc.

   (c) `ABCXYZ001.qmd` is the .qmd file used to generate the paper. Use 'embed-resources: true' in your YAML to generate a single .html file from the .qmd.

   Please indicate all students in your group in the author section of .qmd YAML. Please submit the files separately in the Assignment tab on Vula; do not zip.

Note that I primarily use points 1 (written report) and 2 (code) to grade your work. Point 3 (the Vula submission) is to verify the website hasn't changed after the submission date, and for me to be able to rerun your analysis if need be.

Scientific paper writing style

See Assignment 1 for details.

Other points

- On group work: it is perfectly fine and probably desirable to divide the work between group members so that some members focus on certain parts of the project. Your report should include a declaration stating what tasks each person in the group was responsible for/participated in. See `http://journals.plos.org/plosone/s/submission-guidelines#loc-author-contributions` and `http://journals.plos.org/plosone/s/authorship#loc-author-contributionsfordetails`. However, all group members must be familiar with all aspects of the project. If you are working on a topic (say sentiment analysis), you are responsible for making sure that the rest of the group understands that topic, and what you have done, to the extent that they could explain the work to someone else.
- Doing your project in R is **recommended** but Python submissions will be accepted. Your report and website must be rendered from a **Quarto Markdown** document.
- Anyone else should be able to run the code in your .qmd to completion. Use `set.seed()` to set a random seed so your final results do not change.
- You may not share any coding or write-up with any other group. Please sign the plagiarism declaration provided on Vula and append it to your Vula submission.
