Data Science for Industry Project 2: Predict the President
Due date: 13 September 2018

This project can be done in groups of 4 or fewer.

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 2018 (sourced from https://www.kaggle.com/allank/state-of-the-nation-1990-2017/home). In years that elections took place, a State of the Nation Address happens twice, once before and again after the election.

The objectives of this project are:

1. To construct a neural network that, given a sentence of text, predicts which president was the source of that sentence.
2. To assess the out-of-sample performance of your classifier.
3. To conduct a descriptive analysis of the text in the speeches. This is a relatively open-ended question but could include, for example, an analysis of word/token frequencies, sentiment analysis – look at the course notebooks for ideas, or see "Tidy text mining with R".

Items 1 & 2 count 60% towards the project mark, item 3 counts 40%.

Write up your work in the form of a report **in the form of R Markdown document or a Jupyter notebook** (if not using R). The report should contain a description of the problem, the approach you took, and your results. Your code should be integrated into the document, and this code should be clearly described and commented (see the class .Rmd notebooks for examples).

Note on group work:

You can cover a lot more ground working in groups than on your own, and group work can be fun, but there is always the risk that some group members do the bulk of the work while others do very little. To this end, some ground rules:

- It is perfectly fine to divide the work between group members, so that some members focus on certain parts of the project. I would suggest that in a group of 4, 2 people work on the predictive model and 2 on the descriptive part, with further division of tasks as you see fit.
- 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-contributions for details.
- 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.
- At the end of the project, each group member will anonymously rate the contribution of each of member of the group, including themselves. These ratings will be used to adjust the group mark to obtain individual marks where necessary (the group mark will remain the mean of the individual marks).
- In extreme circumstances where I suspect a group is not playing by the rules I will ask the group in for an oral question-and-answer-style assessment.
- +5-10% for visible evidence of group collaboration on GitHub (this could be a shared repo, pull requests from group members, etc – and should be described in the report).