# Introduction to the use of R for applying content/sentiment anlaysis for OMNeoHealth's policy analysis component

*16/05/2018*

For this note we will use the text data of Papua New Guinea's Alotau Accord of 2014 available in the `papuanewguinea` data package of the OMNeoHealth GitHub project.

First, install the `papuanewguinea` data package:

```
#
# Install devtools package in R
# (allows for installation of packages from GitHub)
#
install.packages("devtools")
#
# Install papuanewguinea data package in R via GitHub
#
devtools::install_github("OMNeoHealth/papuanewguinea")
```

Once installed, the Alotau Accord of 2014 is now available for us to use. the object name for the text data is `alotau_accord_2014`

You can examine the data by simply typing in the name of the data in R. In the code snippet below, I use a function from the `dplyr` package called `as_tibble()` which prints out the dataframe in a `tibble` format. A tibble is a modern class of data frame within R, available in the dplyr and tibble packages, that has a convenient print method, will not convert strings to factors, and does not use row names. I use it here for convenience and to make the printing of the dataframe tidier. To use `as_tibble`, we need to install and load the `dplyr` package first and then view the dataframe as tibble:

```
# install.packages("dplyr")
library(dplyr)
as_tibble(papuanewguinea::alotau_accord_2014)
```

```
## # A tibble: 477 x 2
##     line text
##    <int> <chr>
##  1     1 "                                    CONTENTS"
##  2     2 Glossary of Terms                                                 ~
##  3     3 Message from the Minister of Health and HIV/AIDS                  ~
##  4     4 Foreword from the Secretary for the National Department of Healt~
##  5     5 What is the Alotau Accord?                                        ~
##  6     6 Why is there a need for Free Primary Health Care and Subsidized ~
##  7     7 Specialist Services?
##  8     8 What is Primary Health Care and Subsidized Specialist Services? ~
##  9     9 What is in this Policy?                                           ~
## 10    10 What Are Facility Levels?                                         ~
## # ... with 467 more rows
```

You will notice that the `alotau_accord_2014` dataset is structured as a dataframe with each row corresponding to each of the lines of text in the actual Alotau Accord of 2014 document.

To be able to work with this text data further using content analysis or sentiment analysis, we need to `tokenise` the dataset. This means we need to break up the text data into `tokens` - a meaningful unit of text, most often a word, that we are interested in using for further analysis. To do this, we will use the `tidytext` package. Install and load the package in R by issuing the following commands:

```
install.packages("tidytext")
library(tidytext)
```

The `tidytext` package has the `unnest_tokens()` function that basically tokenises the dataset. This can be applied to the `alotau_accord_2014` dataset as follows:

```
alotauDF <- unnest_tokens(tbl = papuanewguinea::alotau_accord_2014,
                          output = word,
                          input = text)
```

Let's now examine what the resulting `alotauDF` data object looks like:

```
as_tibble(alotauDF)
```

```
## # A tibble: 4,652 x 2
##     line word
##    <int> <chr>
##  1     1 contents
##  2     2 glossary
##  3     2 of
##  4     2 terms
##  5     2 2
##  6     3 message
##  7     3 from
##  8     3 the
##  9     3 minister
## 10     3 of
## # ... with 4,642 more rows
```

The `alotauDF` data object is now a dataframe of all the words in the actual Alotau Accord of 2014.

Given such a data structure, we can now potentially perform various types of analysis based on the word content of the accord. This analysis should be guided by the research questions that need to be answered by the project.

The example above treats the policy document as a collection of words. Depending on the research question/s, you may want to tokenise the policy document/s into sentences, paragraphs, n-grams (collection of consecutive words) on which you can perform various types of analysis.

## Further reading

I would recommend the book ***Text Mining with R*** as our main reference guide for doing the content/sentiment analysis in R. It is available to read online as an online book at https://www.tidytextmining.com.