# Text Classification (Newspaper) using Naive Bayes

*Sheryl Mathew (11627236)*

*20 October, 2018*

### Data Collection

```
library(jsonlite)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(kableExtra)

api = 'https://content.guardianapis.com/search?'
api_key = '0e9a8c4e-1866-4503-a9eb-70e391c499c4'
page_size = '200'
pages = c(1:5)
queries = c('business','sports','entertainment','economy','politics','science','health','art','technolog

news_data = data.frame()
data_count = list()
sum_data_count = list()


for(query in queries){
   for (page in pages){
      url = paste(api, 'q=', query, '&page-size=', page_size, '&page=', page, '&api-key=', api_key,
            '&show-fields=body', sep = "")
      json = fromJSON(url)
      body = as.data.frame(json$response$results$fields)
      data = as.data.frame(json$response$results)
      data = subset(data, select = -c(fields))
      data = cbind(data,body)
      data_count = append(data_count,nrow(data))
      news_data = rbind(news_data,data)
  }


  select_data=select(data,c("sectionName","body"))
  print(kable(head(select_data), format = "latex", booktabs = T,
          caption=paste("Table containing",toupper(query)," News")) %>%
      kable_styling(latex_options = c("striped","hold_position","scale_down")))
  cat("\n")
```