Practical No. 3

Aim: To clean rediff news articles data set.

Approach Used & Reason for using the Approach:

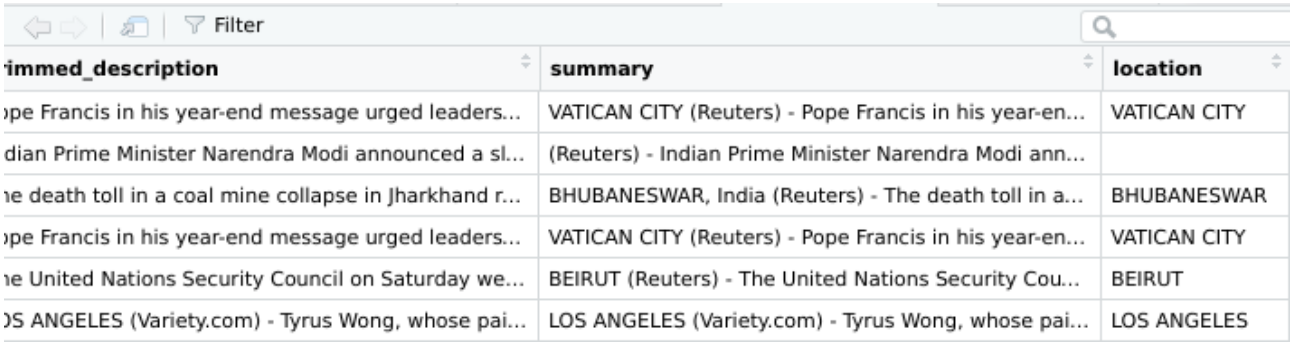
I have used the approach to select different methods from news from different sources. I have used this approach on the basis that data from different sources have different patterns in data. I think the column Crawling Time to be a good approximation of the time when the article was created and published.

Code:

```
a <- rediff_realtime_news_201701_201703[,4, drop=FALSE]

# Script to scrap Reuters
reutersreports <- rediff_realtime_news_201701_201703[grep("Reuters",
rediff_realtime_news_201701_201703$source), ]
reuterscities <- reutersreports
reuterscities$location <- gsub("(\\b(([a-z][a-z]+\\s[a-z]+))|(([A-Z][a-z]+))|(([a-z]+))|(([a-z][A-Z]+))|
(([0-9]+))|(\\W)+|(REUTERS)\\b)|(\\b([A-Z])\\b)", " ", reuterscities$summary)

# Script to scrap The Hindu
thehindureports <- rediff_realtime_news_201701_201703[grep("The Hindu",
rediff_realtime_news_201701_201703$source), ]
newhinducities <- thehindureports
newhinducities$location <- substr(newhinducities$summary, 1, 20)
newhinducities <- newhinducities[grep(":", newhinducities$location), ]
newhinducities$location <- gsub(":(.*)|", "", newhinducities$location)

# Script to scrap DNA
dnareports <- rediff_realtime_news_201701_201703[grep("DNA",
rediff_realtime_news_201701_201703$source), ]


# Script to scrap DeccanHerald
deccanheraldreports <- rediff_realtime_news_201701_201703[grep("Deccan Herald",
rediff_realtime_news_201701_201703$source), ]


# Script to scrap MSN India
msnindiareports <- rediff_realtime_news_201701_201703[grep("MSN India",
rediff_realtime_news_201701_201703$source), ]


# Script to scrap Sify
sifyreports <- rediff_realtime_news_201701_201703[grep("Sify",
rediff_realtime_news_201701_201703$source), ]
sifyreportsreuters <- sifyreports[grep("Reuters", sifyreports$summary), ]
library(dplyr)
```

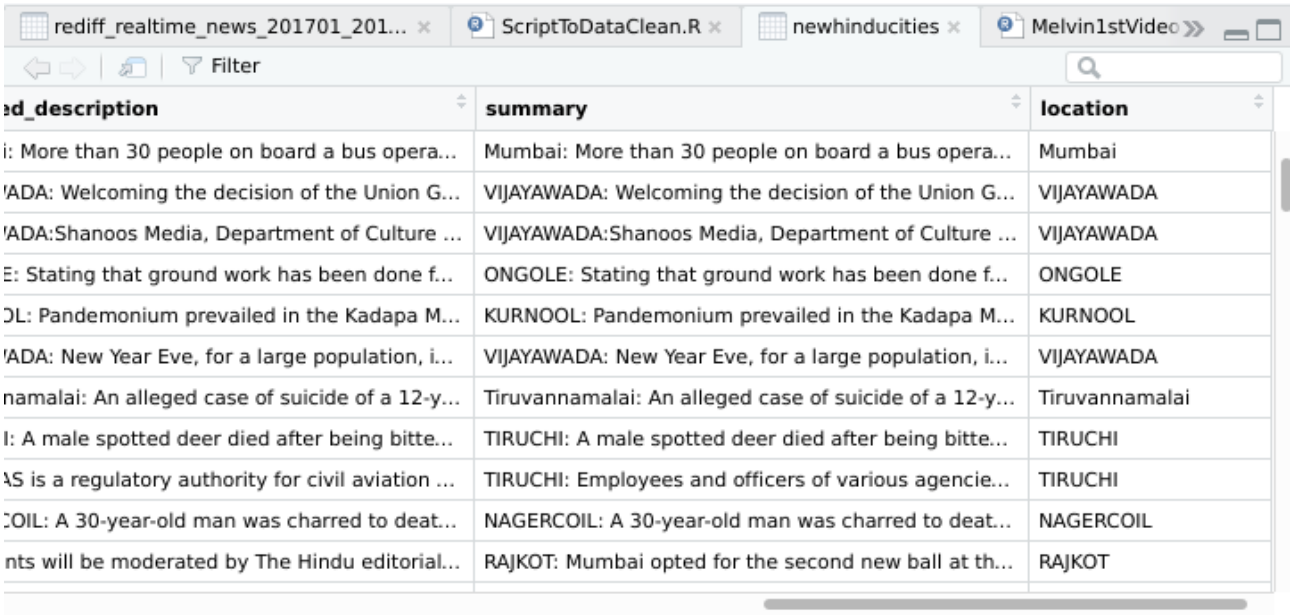
```
sifyreportswithoutreuters <- setdiff(sifyreports, sifyreportsreuters)

sifyreportswithoutreuterswithani <- sifyreportswithoutreuters[grep("ANI",
sifyreportswithoutreuters$summary), ]
#sifyreportswithoutreuterswithani$location <- substr(sifyreportswithoutreuterswithani$summary, 1,
20)

sifyreportswithoutreuterwithoutani <- setdiff(sifyreportswithoutreuters,
sifyreportswithoutreuterswithani)

sifyreportswithoutreuterswithani <- sifyreportswithoutreuters[grep("ANI",
substr(sifyreportswithoutreuters$summary, 1, 20)), ]

sifyreportswithoutreuterswithani$location <- gsub("(.*\\[)|(\\].*)", "",
sifyreportswithoutreuterswithani$summary)
sifyreportswithoutreuterswithani$location <- gsub(":(.*)|", "",
sifyreportswithoutreuterswithani$location)

sifyreportswithoutreuterwithoutani$location <- substr(sifyreportswithoutreuterwithoutani$summary,
1, 20)
sifyreportswithoutreuterwithoutani <- sifyreportswithoutreuterwithoutani[grep(":|-",
sifyreportswithoutreuterwithoutani$location), ]
sifyreportswithoutreuterwithoutani$location <- gsub(":(.*)|-(.*)", "",
sifyreportswithoutreuterwithoutani$location)


# Script to scrap Business Standard
# Business Standard Dates are good
businessstandardreports <- rediff_realtime_news_201701_201703[grep("Business Standard",
rediff_realtime_news_201701_201703$source), ]


# IANS Data
iansreports <- rediff_realtime_news_201701_201703[grep("IANS",
rediff_realtime_news_201701_201703$trimmed_description), ]

#sifyreports <- sifyreports[ !(sifyreports %in% sifyreportsreuters), ]
#english$title <- substr(b, 0, 10)
```

SAMPLE OUTPUT:



Filter

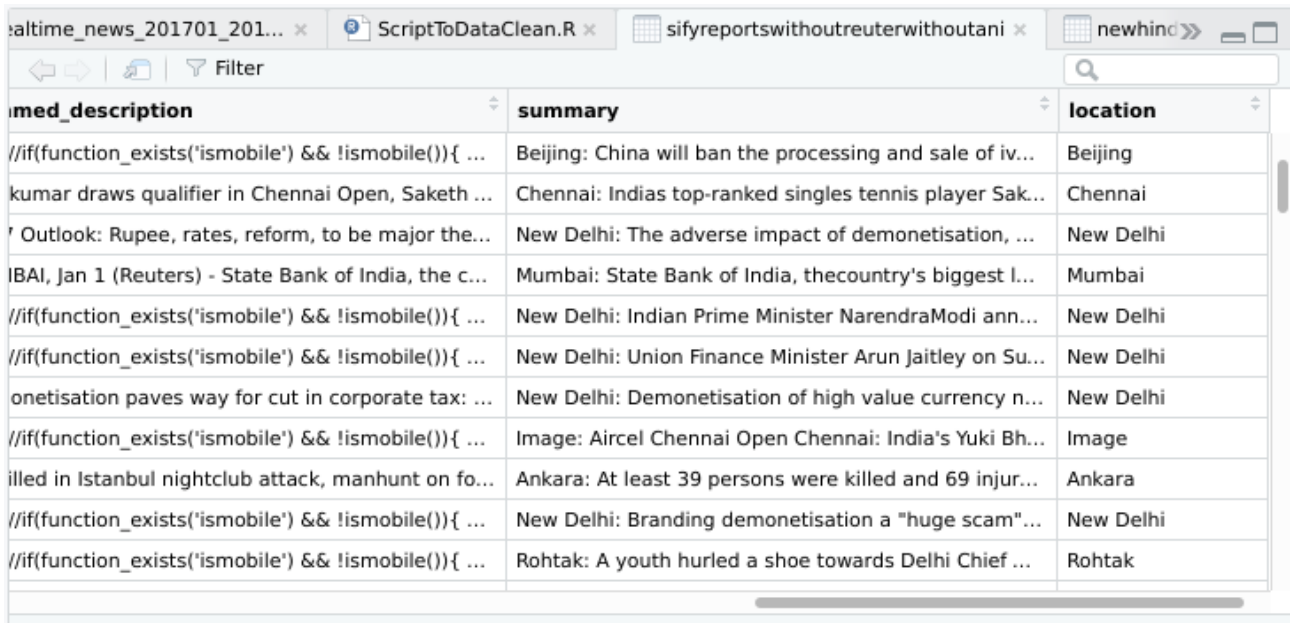
rimmed_description	summary	location
ope Francis in his year-end message urged leaders...	VATICAN CITY (Reuters) - Pope Francis in his year-en...	VATICAN CITY
dian Prime Minister Narendra Modi announced a sl...	(Reuters) - Indian Prime Minister Narendra Modi ann...	
ne death toll in a coal mine collapse in Jharkhand r...	BHUBANESWAR, India (Reuters) - The death toll in a...	BHUBANESWAR
ope Francis in his year-end message urged leaders...	VATICAN CITY (Reuters) - Pope Francis in his year-en...	VATICAN CITY
ne United Nations Security Council on Saturday we...	BEIRUT (Reuters) - The United Nations Security Cou...	BEIRUT
OS ANGELES (Variety.com) - Tyrus Wong, whose pai...	LOS ANGELES (Variety.com) - Tyrus Wong, whose pai...	LOS ANGELES



rediff_realtime_news_201701_201... | ScriptToDataClean.R | newhinducities | Melvin1stVideo

Filter

ed_description	summary	location
i: More than 30 people on board a bus opera...	Mumbai: More than 30 people on board a bus opera...	Mumbai
'ADA: Welcoming the decision of the Union G...	VIJAYAWADA: Welcoming the decision of the Union G...	VIJAYAWADA
'ADA:Shanoos Media, Department of Culture ...	VIJAYAWADA:Shanoos Media, Department of Culture ...	VIJAYAWADA
E: Stating that ground work has been done f...	ONGOLE: Stating that ground work has been done f...	ONGOLE
OL: Pandemonium prevailed in the Kadapa M...	KURNOOL: Pandemonium prevailed in the Kadapa M...	KURNOOL
'ADA: New Year Eve, for a large population, i...	VIJAYAWADA: New Year Eve, for a large population, i...	VIJAYAWADA
namalai: An alleged case of suicide of a 12-y...	Tiruvannamalai: An alleged case of suicide of a 12-y...	Tiruvannamalai
I: A male spotted deer died after being bitte...	TIRUCHI: A male spotted deer died after being bitte...	TIRUCHI
AS is a regulatory authority for civil aviation ...	TIRUCHI: Employees and officers of various agencie...	TIRUCHI
COIL: A 30-year-old man was charred to deat...	NAGERCOIL: A 30-year-old man was charred to deat...	NAGERCOIL
nts will be moderated by The Hindu editorial...	RAJKOT: Mumbai opted for the second new ball at th...	RAJKOT



altime_news_201701_201... | ScriptToDataClean.R | sifyreportswithoutreuterwithoutani | newhind

Filter

med_description	summary	location
//if(function_exists('ismobile') && !ismobile()){ ...	Beijing: China will ban the processing and sale of iv...	Beijing
kumar draws qualifier in Chennai Open, Saketh ...	Chennai: Indias top-ranked singles tennis player Sak...	Chennai
' Outlook: Rupee, rates, reform, to be major the...	New Delhi: The adverse impact of demonetisation, ...	New Delhi
IBAI, Jan 1 (Reuters) - State Bank of India, the c...	Mumbai: State Bank of India, thecountry's biggest l...	Mumbai
//if(function_exists('ismobile') && !ismobile()){ ...	New Delhi: Indian Prime Minister NarendraModi ann...	New Delhi
//if(function_exists('ismobile') && !ismobile()){ ...	New Delhi: Union Finance Minister Arun Jaitley on Su...	New Delhi
onetisation paves way for cut in corporate tax: ...	New Delhi: Demonetisation of high value currency n...	New Delhi
//if(function_exists('ismobile') && !ismobile()){ ...	Image: Aircel Chennai Open Chennai: India's Yuki Bh...	Image
illed in Istanbul nightclub attack, manhunt on fo...	Ankara: At least 39 persons were killed and 69 injur...	Ankara
//if(function_exists('ismobile') && !ismobile()){ ...	New Delhi: Branding demonetisation a "huge scam"...	New Delhi
//if(function_exists('ismobile') && !ismobile()){ ...	Rohtak: A youth hurled a shoe towards Delhi Chief ...	Rohtak

Conclusion:

I noticed that each news agency uses a specific way of formatting and printing information. In short this should be exploited in a way considering there are not all agencies producing quality diverse information and agencies not producing quality data could be excluded.

In order to do that the challenges faced were getting dirty with the data and understanding and researching about the ‘Good’ News Agencies in the world and in India.

This method also solved the Big Data problem using conventional Divide and Rule Strategy where data frames size were filtered and limited by the news agency name.

Another advantages was the ease of debugging and finding faults within the cleaning process as well as the ability to handle cleaning on the laptop itself.