

LETTERKENNY INSTITUTE OF TECHNOLOGY

ASSIGNMENT COVER SHEET

Lecturer's Name: **James Connolly**

Assessment Title: **Data scraping**

Work to be submitted to: **James Connolly**

Date for submission of work: **29-08-2018**

Place and time for submitting work: **1:00 pm**

To be completed by the Student

Student's Name: **PRATEEK PARASHER**

Class: **Msc in Big Data**

Subject/Module: **Data Science**

Word Count (where applicable): _____

I confirm that the work submitted has been produced solely through my own efforts.

Student's signature: **PRATEEK PARASHER** Date: **28/8/2018**

Notes

Penalties: The total marks available for an assessment is reduced by 15% for work submitted up to one week late. The total marks available are reduced by 30% for work up to two weeks late. Assessment work received more than two weeks late will receive a mark of zero. [Incidents of alleged plagiarism and cheating are dealt with in accordance with the Institute's Assessment Regulations.]

Plagiarism: Presenting the ideas etc. of someone else without proper acknowledgement (see section L1 paragraph 8).

Cheating: The use of unauthorised material in a test, exam etc., unauthorised access to test matter, unauthorised collusion, dishonest behaviour in respect of assessments, and deliberate plagiarism (see section L1 paragraph 8).

Continuous Assessment: For students repeating an examination, marks awarded for continuous assessment, shall normally be carried forward from the original examination to the repeat examination.

Description: - Twitter is a great source for sentiment data and social media mining furthermore it is quite easy to get significant amounts of data to be able to scrape data from Twitter. Last year we all experienced the different climate conditions and multiple climate warning announced by govt. this lab I scrapped the data from twitter regarding climate warning like heatwaves warning, snow warning etc. after that I visualize the wordcloud of scraped data from twitter.

GitHub Link :- https://github.com/prateekparasher/web_scrap

STEP DESCRIPTION AND CODE

#install package

```
install.packages("rvest")
```

```
library(rvest)
```

#read link

```
url <- 'https://twitter.com/weatherrte?lang=en'
```

```
web_page <- read_html(url)
```

#head & str data

```
head(web_page)
```

```
str(web_page)
```

#ranking tweet data

```
warning <- html_nodes(web_page, '.tweet-text')
```

```
head(warning, 30)
```

#length of tweet data

```
length(warning)
```

```
data <- html_text(warning)
```

#show first 30 tweets

```
head(data, 30)
```

#Install and load the required packages

Install

```
install.packages("tm") # for text mining
```

```
install.packages("SnowballC") # for text stemming
```

```
install.packages("wordcloud") # word-cloud generator
```

```
install.packages("RColorBrewer") # color palettes
```

Load

```
library("tm")
```

```
library("SnowballC")
```

```
library("wordcloud")
```

```
library("RColorBrewer")
```

Load the data as a corpus

```
docs <- Corpus(VectorSource(data))
```

```
#Inspect the content of the document
```

```
inspect(docs)
```

#Text transformation

```
toSpace <- content_transformer(function (x , pattern ) gsub(pattern, " ", x))
```

```
docs <- tm_map(docs, toSpace, "/")
```

```
docs <- tm_map(docs, toSpace, "@")
```

```
docs <- tm_map(docs, toSpace, "\\|")
```

Convert the text to lower case

```
docs <- tm_map(docs, content_transformer(tolower))
```

Remove numbers

```
docs <- tm_map(docs, removeNumbers)
```

Remove english common stopwords

```
docs <- tm_map(docs, removeWords, stopwords("english"))
```

Remove your own stop word

stopwords as a character vector

```
docs <- tm_map(docs, removeWords, c("blabla1", "blabla2"))
```

Remove punctuations

```
docs <- tm_map(docs, removePunctuation)
```

Eliminate extra white spaces

```
docs <- tm_map(docs, stripWhitespace)
```

Text stemming

```
docs <- tm_map(docs, stemDocument)
```

#Build a term-document matrix

```
dtm <- TermDocumentMatrix(docs)
```

```
m <- as.matrix(dtm)
```

```
v <- sort(rowSums(m),decreasing=TRUE)
```

```
d <- data.frame(word = names(v),freq=v)
```

```
head(d, 20)
```

Generate the Word cloud

```
set.seed(1234)
```

```
wordcloud(words = d$word, freq = d$freq, min.freq = 1,  
           max.words=200, random.order=FALSE, rot.per=0.35,  
           colors=brewer.pal(8, "Dark2"))
```

STR (Structure of scraped data)

```
str(dtm)
```

```
List of 6
```

```
$ i      : int [1:355] 1 2 3 4 5 6 7 8 9 10 ...  
$ j      : int [1:355] 1 1 1 1 1 1 1 1 1 1 ...  
$ v      : num [1:355] 1 1 1 1 1 1 1 1 1 1 ...  
$ nrow   : int 222  
$ ncol   : int 21  
$ dimnames:List of 2  
..$ Terms: chr [1:222] "can" "changes" "coverage" "forecast" ...  
..$ Docs  : chr [1:21] "1" "2" "3" "4" ...  
- attr(*, "class")= chr [1:2] "TermDocumentMatrix" "simple_triplet_matrix"  
- attr(*, "weighting")= chr [1:2] "term frequency" "tf"
```

```
str(d)
```

```
'data.frame': 222 obs. of 2 variables:
```

```
$ word: Factor w/ 222 levels "across","advisory",...: 153 216 145 208 64 45 5  
2 71 206 154 ...
```

```
$ freq: num 12 12 9 9 7 6 6 5 5 4 ...
```

RESULT & VISUALIZATION

```
content: documents: 21
```

[1] With live coverage on #RTEOne of the #PopeInIreland and Mass from the #PhoenixPark, there may be changes to the schedule today. If so you can see this week's #ISL forecast on #RTEPlayer. Please RT @RTEOne

[2] There's a weather Advisory in place for heavy rain this evening and early tonight - get the forecast for that and the next few days at 1.25 on #RTEOne

[3] Heat warnings remain in operation across Turkey will highs of 35 degrees. @nualacarey25 has more in your #europea forecast at 5.50pm this evening on #RTEOne

[4] Come hail, rain or shine #RTEweather is a constant in the #RTENewSeason schedule. Our office is open 365 days a year! @nualacarey25 <https://twitter.com/helcurran/status/1030049983716892672> ...

[5] Myself and @nualacarey25 at #RTENewSeason pic.twitter.com/yftEkqx9BD

[6] WATCH: After a long dry spell the rain pours down in Dublin #OnThisDay in 1968 <https://bit.ly/2B0vbH7> pic.twitter.com/b6VCokhCBE

[7] Mostly dry with some sunny spells in the east, but cloudier in the west and northwest with patchy rain and drizzle extending slowly eastwards but tending to die away. Top temperatures will be 17 to 22 degrees. pic.twitter.com/cfw4IvcWNR

[8] Dry with some sunshine this morning but cloudier and humid in the southwest with rain gradually spreading northwards, turning persistent and heavy later in the afternoon and evening. Top temperatures of 18 to 21 degrees.

[9] Cloudy with some outbreaks of rain today but generally dry, sunshine will develop in west & north but cooler air moving into west ...highs of 16C there- up to 22C for the east & southeast, join me for an update on @rteone at 11.50am

[10] Very hot weather in Europe now, especially in Iberia. High temperature warnings have been issued as the figures reach over 40 degrees in Spain and Portugal. More details in the #europeanforecast on @RTEOne at 5.50pm approx.

[11] Sunscreen that is spread on the skin too thinly may provide less than half the expected degree of protection, a study has shown <https://www.rte.ie/news/newslens/2018/0725/980981-sun-lotion-application/> ... via @rtenews

[12] Solar UV Index for wednesday: HIGH in all areas.

[13] Dry in most places today with long sunny spells. However there is the chance of a few showers along Atlantic coastal counties. Warm with highs of 20 to 25 degrees. Light to moderate southerly winds will increase strong and gusty along the Atlantic seaboard later in the day. pic.twitter.com/OQy9sd7ojf

[14] See your weekly #ISL weather forecast back on #RTEOne at 5.50pm approx. this evening. Available shortly thereafter on @rteplayer #irishsignlanguage #weather #rteplayer @SenanDunne @electricginger @Carabellew @CarolineMcTweet @RTEOne

[15] Weather warnings have been issued for rain and thunder in eastern parts of Europe, while in southern Italy they are in place for high temperatures. More details this evening #RTEOne at 5.50pm in your European forecast. #weather

[16] Solar UV index: HIGH in sunshine nationwide on Sunday.

[17] A series of weather warnings remain in place across Europe. They concern a mix of rain & thunderstorms for Croatia, Bosnia, Serbia and Italy. In Sardinia and Italy there are also wind warnings in operation. Full details at 5.50pm on #RTEOne with @nualacarey25

[18] Apart from well scattered showers, many places will remain dry this afternoon and evening. Rather cloudy generally but sunny spells will occur too. Highest temperatures 17 to 20 degrees

Fig 1 – warning tweets scraped from twitter using R

