LETTERKENNY INSTITUTE OF TECHNOLOGY

ASSIGNMENT COVER SHEET

Lecturer's Name: James Connolly				
Assessment Title: Data scraping				
Work to be submitted to: 08/01/19				
Date for submission of work:James Connolly				
Place and time for submitting work:				
To be completed by the Student				
Student's Name: PRATEEK PARASHER				
Class: MSc Big Data Analytics				
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: 08/01/19				

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.

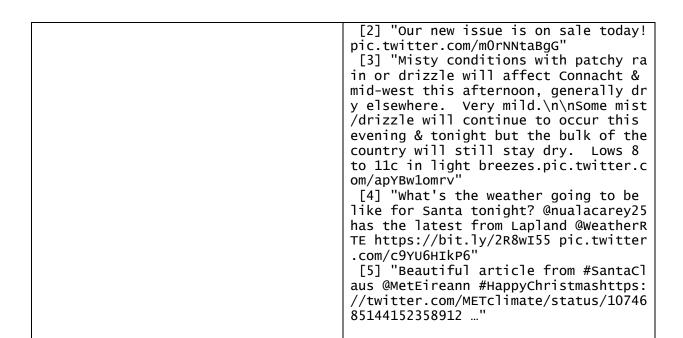
Description: - Twitter is a great source for sentiment data and social media mining furthermore it is quite easy to get significant amounts of data to be able to scrape data from Twitter. Last year we all experienced the different climate conditions and multiple climate warning announced by govt. In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

Step Description: -	install packages
	read link
	head & str data
R code used to perform the scrape: -	<pre>install.packages("rvest") library(rvest) url <- 'https://twitter.com/weatherrte?lang=en' web_page <- read_html(url) head(web_page)</pre>
Structure of scraped data: -	<pre>str(web_page) List of 2 \$ node:<externalptr> \$ doc :<externalptr> - attr(*, "class")= chr [1:2] "xml_document" "xml_node"</externalptr></externalptr></pre>
Result: -	<pre>\$`node` <pointer: 0x00000000d9dd100=""> \$doc <pointer: 0x000000008ef38a0=""></pointer:></pointer:></pre>

Step Description: -	ranking tweet data
R code used to perform the scrape: -	warning <- html_nodes(web_page,'.tweet-text')
Structure of scraped data: -	<pre>str(warning) List of 24 \$:List of 2\$ node:<externalptr>\$ doc :<externalptr> attr(*, "class")= chr "xml_node"</externalptr></externalptr></pre>
	<pre>\$:List of 2\$ node:<externalptr>\$ doc :<externalptr></externalptr></externalptr></pre>

	<pre> attr(*, "class")= chr "xml_node " \$:List of 2\$ node:<externalptr>\$ doc :<externalptr> attr(*, "class")= chr "xml_node "</externalptr></externalptr></pre>
Result: -	[1] <pre>class="TweetTextSize TweetText Sizenormal js-tweet-text tweet-text " lang="en" data-aria-label-part="4"> With details of how to enter the Wint er quarter of the RTE Wea [2] <div ata-aria-label-part="2" class="QuoteTweet-text twee t-text u-dir js-ellipsis" d="" dir="ltr" lang="en">Our new issue is on sale today! <span cla="" data-aria-label-part="0" js-tweet-text="" lang="en" ss="twitter-timeline-link [3] <p class=" t"="" tsizenormal="" tweet-tex="" tweettex="" tweettextsize="">Misty conditions with patchy rain or drizzle will affect Conna [4] What's the weather going to be like for Santa tonight? <a <p="" [5]="" class="TweetTextSize TweetTex tSizenormal js-tweet-text tweet-tex t" data-aria-label-part="4" href="" lang="en">Beautiful article from <a <="" href="/has htag/SantaClaus?src=hash" pre=""></div></pre>

Step Description: -	length of tweet data
R code used to perform the scrape: -	length(warning) data <- html_text(warning) head(data, 30) str(data)
Structure of scraped data: -	chr [1:24] "With details of how to enter the Winter quarter of the RTE Weather Photo competitionso snap to it! @RTEOne @" truncated
Results: -	[1] "with details of how to enter the Winter quarter of the RTE Weather Pho to competitionso snap to it! @RTEO ne @WeatherRTE #Weather #winter #phot ography #competitionhttps://twitter.com/RTE_GUIDE/status/10785990436123729 94"



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