Cloud Set Up Process – I have created the Amazon EC2 instance by the following steps: -

1) Logged in to my AWS account and selected a free tier.[4]
2) Launched a Virtual machine with EC2 instance.
3) Selected Ubuntu Server 18.04 LTS (HVM), SSD Volume Type.
4) Selected the storage I need and created a key pair for ". pem" file to connect to virtual machine.
5) Then Launched the instance.
6) Used putty to generate a .ppk file by loading .pem file and passphrase
7) Using that .ppk and IP address provided my amazon instance I launched the virtual machine.

Data Extraction Process - I have followed the below process for Twitter data extraction

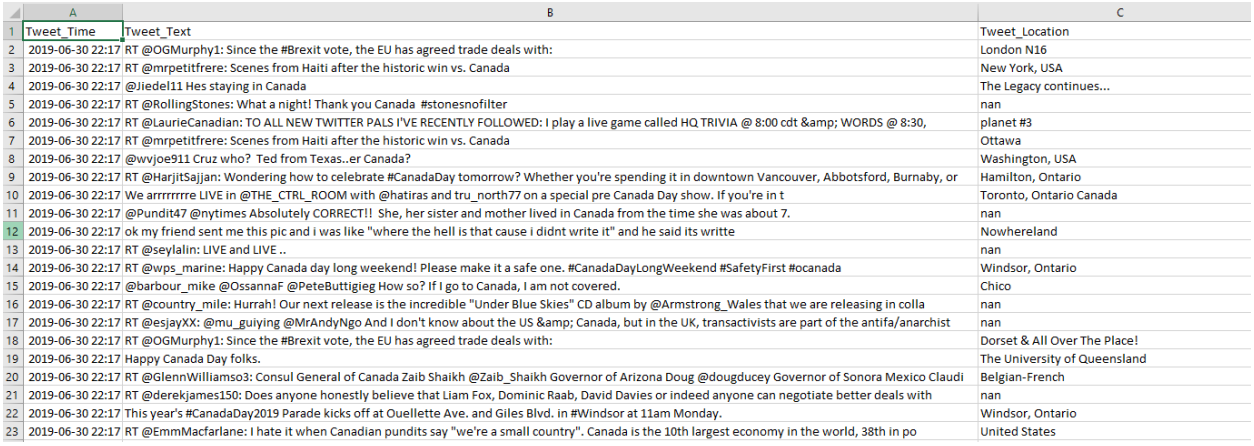
1) I created the twitter developer account and used Tweepy API for all data extraction.[3]
2) I have writer two separate programs namely: - "search_api.py" and "stream_api.py" for extracting around 2200 tweets in total.[5]
3) Search_api.py is using the tweepy search API and Stream_api.py is using the tweepy stream API to retrieves tweets based on keywords provided in assignment.
4) I am passing the keys that I generated from twitter developer account in my python script to establish a connection and then retrieving the data based on keywords. I have also followed PEP 8 standards and given comments in both my python script for easier understandability.

For extracting data from two .SGM files, I have followed the below process: -

1) I have written a python script named: - "multi_files.py"
2) I have used re and os modules.[1][2]
3) First, I am reading the content of both the .sgm files and appending all the contents to an empty string.
4) Then using re package, I have created a regex which helps me extract data between <TEXT></TEXT> tags.
5) The content between these tags are stored as a list element.
6) Then from each element in the list I am creating a new file.

Data Cleaning Process - All cleaning steps are written below: -

1) I am using pandas to load the extracted tweets in csv files into dataframe.
2) I am merging the two dataframe into one so that I can clean both easily.
3) First, I am making sure that the text column is in string format by applying a lambda function.
4) I am using strip() to clean the extra white spaces.
5) Using str.replace() , I am passing a regex to remove the URL from the tweets.
6) I have written a function that removes all non ascii characters.
7) This function helps me in cleaning all the unnecessary special characters and smileys from the tweets.
8) I am having the metadata like tweet_time and tweet_location and this file again I am saving to csv and importing it in mongo db.
9) For spark program to count the keywords I am creating another csv and import it in RDD format in pyspark.



	A	B	C
1	Tweet_Time	Tweet_Text	Tweet_Location
2	2019-06-30 22:17	RT @OGMurphy1: Since the #Brexit vote, the EU has agreed trade deals with:	London N16
3	2019-06-30 22:17	RT @mrpetitfrere: Scenes from Haiti after the historic win vs. Canada	New York, USA
4	2019-06-30 22:17	@Jiedel11 Hes staying in Canada	The Legacy continues...
5	2019-06-30 22:17	RT @RollingStones: What a night! Thank you Canada #stonesnofilter	nan
6	2019-06-30 22:17	RT @LaurieCanadian: TO ALL NEW TWITTER PALS I'VE RECENTLY FOLLOWED: I play a live game called HQ TRIVIA @ 8:00 cdt & WORDS @ 8:30,	planet #3
7	2019-06-30 22:17	RT @mrpetitfrere: Scenes from Haiti after the historic win vs. Canada	Ottawa
8	2019-06-30 22:17	@wvjoe911 Cruz who? Ted from Texas..er Canada?	Washington, USA
9	2019-06-30 22:17	RT @HarjitSajjan: Wondering how to celebrate #CanadaDay tomorrow? Whether you're spending it in downtown Vancouver, Abbotsford, Burnaby, or	Hamilton, Ontario
10	2019-06-30 22:17	We arrrrrrrre LIVE in @THE_CTRL_ROOM with @hatiras and tru_north77 on a special pre Canada Day show. If you're in t	Toronto, Ontario Canada
11	2019-06-30 22:17	@Pundit47 @nytimes Absolutely CORRECT!! She, her sister and mother lived in Canada from the time she was about 7.	nan
12	2019-06-30 22:17	ok my friend sent me this pic and i was like "where the hell is that cause i didnt write it" and he said its writte	Nowhereland
13	2019-06-30 22:17	RT @seylalin: LIVE and LIVE ..	nan
14	2019-06-30 22:17	RT @wps_marine: Happy Canada day long weekend! Please make it a safe one. #CanadaDayLongWeekend #SafetyFirst #ocanada	Windsor, Ontario
15	2019-06-30 22:17	@barbour_mike @OssannaF @PeteButtigieg How so? If I go to Canada, I am not covered.	Chico
16	2019-06-30 22:17	RT @country_mile: Hurrah! Our next release is the incredible "Under Blue Skies" CD album by @Armstrong_Wales that we are releasing in colla	nan
17	2019-06-30 22:17	RT @esjayXX: @mu_guiying @MrAndyNgo And I don't know about the US & Canada, but in the UK, transactivists are part of the antifa/anarchist	nan
18	2019-06-30 22:17	RT @OGMurphy1: Since the #Brexit vote, the EU has agreed trade deals with:	Dorset & All Over The Place!
19	2019-06-30 22:17	Happy Canada Day folks.	The University of Queensland
20	2019-06-30 22:17	RT @GlennWilliamso3: Consul General of Canada Zaib Shaikh @Zaib_Shaikh Governor of Arizona Doug @dougducey Governor of Sonora Mexico Claudi	Belgian-French
21	2019-06-30 22:17	RT @derekjames150: Does anyone honestly believe that Liam Fox, Dominic Raab, David Davies or indeed anyone can negotiate better deals with	nan
22	2019-06-30 22:17	This year's #CanadaDay2019 Parade kicks off at Ouellette Ave. and Giles Blvd. in #Windsor at 11am Monday.	Windsor, Ontario
23	2019-06-30 22:17	RT @EmmMacfarlane: I hate it when Canadian pundits say "we're a small country". Canada is the 10th largest economy in the world, 38th in po	United States

(Sample screen capture of Cleaned CSV file)

REFERENCES: -

[1]"re — Regular expression operations — Python 3.7.4rc1 documentation", *Docs.python.org*, 2019. [Online]. Available: https://docs.python.org/3/library/re.html. [Accessed: 02- Jul- 2019].

[2]"os — Miscellaneous operating system interfaces — Python 3.7.4rc1 documentation", *Docs.python.org*, 2019. [Online]. Available: https://docs.python.org/3/library/os.html. [Accessed: 02- Jul- 2019].

[3]"Tweepy Documentation — tweepy 3.7.0 documentation", *Docs.tweepy.org*, 2019. [Online]. Available: http://docs.tweepy.org/en/latest/. [Accessed: 02- Jul- 2019].

[4]*Dal.brightspace.com*, 2019. [Online]. Available: https://dal.brightspace.com/d2l/le/content/98340/viewContent/1333088/View. [Accessed: 02- Jul- 2019].

[5]"Twitter API with Python: Part 1 -- Streaming Live Tweets", *YouTube*, 2019. [Online]. Available: https://www.youtube.com/watch?v=wlnx-7cm4Gg. [Accessed: 02- Jul- 2019].