

Data Ingestion GA-2436

Yihao Zhong Larry (yz7654)

Data Source

The data set is located at

[`https://archive.org/download/twitter_cikm_2010/twitter_cikm_2010.zip/`](https://archive.org/download/twitter_cikm_2010/twitter_cikm_2010.zip/)

It is a Twitter content data from September 2009 to January 2010. It is part of the research paper dataset from https://archive.org/details/twitter_cikm_2010. Here is the research paper information:

Z. Cheng, J. Caverlee, and K. Lee. You Are Where You Tweet: A Content-Based Approach to Geo-locating Twitter Users. In Proceeding of the 19th ACM Conference on Information and Knowledge Management (CIKM), Toronto, Oct 2010.

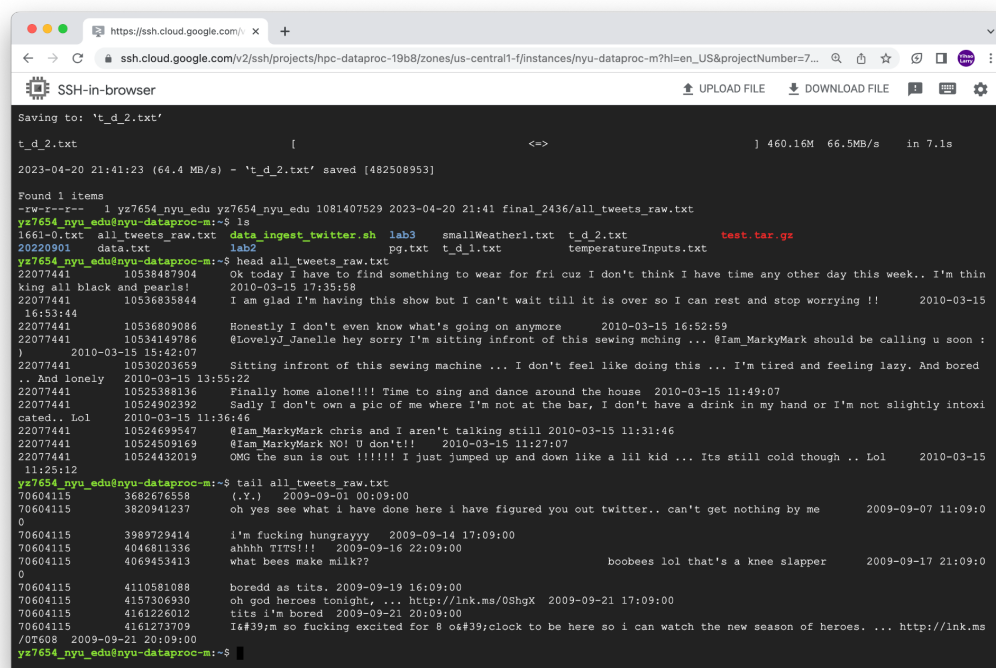
The dataset is around 1GB of size.

I use a shell script to download the dataset (in .txt) onto HDFS by calling wget and hadoop fs -put. It is in the file `data_ingest_twitter.sh`

Data Ingestion Process

It is featured in `ExtractMessageContent.java` mapreduce file and `compile.sh` shell script.

The snapshot of the input file:



```
SSH-in-browser
Saving to: 't_d_2.txt'
t_d_2.txt [ 460.16M 66.5MB/s in 7.1s
2023-04-20 21:41:23 (64.4 MB/s) - 't_d_2.txt' saved [482508953]

Found 1 items
-rw-r--r-- 1 yz7654_nyu_edu yz7654_nyu_edu 1081407529 2023-04-20 21:41 final_2436/all_tweets_raw.txt
yz7654_nyu_edu@nyu-dataproc-m:~$ ls
1661-0.txt all_tweets_raw.txt data_ingest_twitter.sh lab3 smallWeather1.txt t_d_2.txt test.tar.gz
20220901 data.txt lab2 pg.txt t_d_1.txt temperatureInputs.txt
yz7654_nyu_edu@nyu-dataproc-m:~$ head all_tweets_raw.txt
22077441 10538487904 Ok today I have to find something to wear for fri cuz I don't think I have time any other day this week.. I'm thin
king all black and pearls! 2010-03-15 17:35:58
22077441 10536835844 I am glad I'm having this show but I can't wait till it is over so I can rest and stop worrying !! 2010-03-15
16:53:44
22077441 10536809086 Honestly I don't even know what's going on anymore 2010-03-15 16:52:59
22077441 10534149786 @LovelyJanelle hey sorry I'm sitting infront of this sewing mchine ... @IamMarkyMark should be calling u soon :
) 2010-03-15 15:42:07
22077441 10530203659 Sitting infront of this sewing machine ... I don't feel like doing this ... I'm tired and feeling lazy. And bored
.. And lonely 2010-03-15 13:55:22
22077441 10525388136 Finally home alone!!!! Time to sing and dance around the house 2010-03-15 11:49:07
22077441 10524902392 Sadly I don't own a pic of me where I'm not at the bar, I don't have a drink in my hand or I'm not slightly intoxi
cated.. Lol 2010-03-15 11:36:46
22077441 10524699547 @IamMarkyMark chris and I aren't talking still 2010-03-15 11:31:46
22077441 10524509169 @IamMarkyMark NO! U don't!! 2010-03-15 11:27:07
22077441 10524432019 OMG the sun is out !!!!! I just jumped up and down like a lil kid ... Its still cold though .. Lol 2010-03-15
11:25:12
yz7654_nyu_edu@nyu-dataproc-m:~$ tail all_tweets_raw.txt
70604115 3682676558 (.Y.) 2009-09-01 00:09:00
70604115 3820941237 oh yes see what i have done here i have figured you out twitter.. can't get nothing by me 2009-09-07 11:09:0
0
70604115 3989729414 I'm fucking hungryy 2009-09-14 17:09:00
70604115 4046811336 abhhh rrrrr!!! 2009-09-16 22:09:00
70604115 4069453413 what bees make milk?? boobeas lol that's a knee slapper 2009-09-17 21:09:0
0
70604115 4110581088 bored as tits. 2009-09-19 16:09:00
70604115 4157306930 oh god heroes tonight, ... http://lnk.ms/0ShgX 2009-09-21 17:09:00
70604115 4161226012 tits I'm bored 2009-09-21 20:09:00
70604115 4161273709 I#39;m so fucking excited for # os#39;clock to be here so i can watch the new season of heroes. ... http://lnk.ms
/0TE08 2009-09-21 20:09:00
yz7654_nyu_edu@nyu-dataproc-m:~$
```

The mapreduce will extract the Twitter content column using Regex.

Here is the snapshot of the output file:

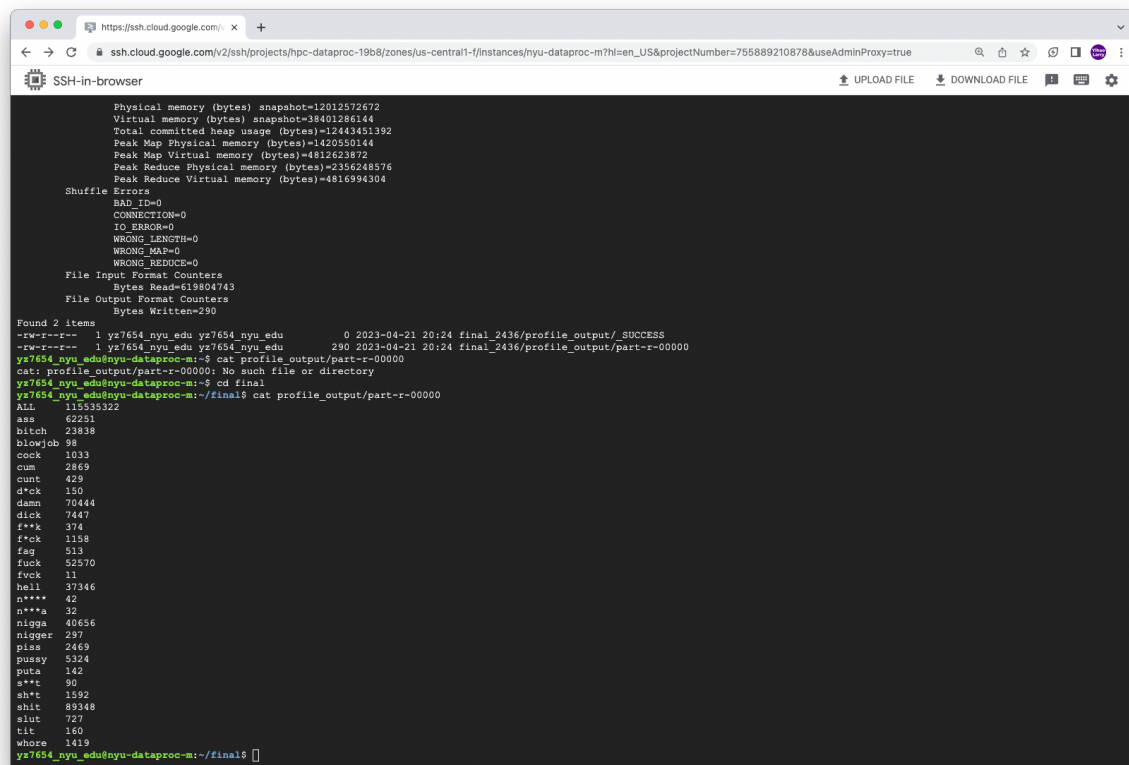
```
yz7654_nyu_edu@nyu-dataproc-m:~/final$ head ingest_output/part-m-00000
New appraisal rules may hurt homebuyers: Both of these groups have been up in arms because they believe the new .. http://bit.ly/6nI2g
One area seeing growth: Trust-and-estate lawyers | Crain&#39;s New ... "We've been seeing a lot of opportunitie.. http://bit.ly/mR848
Welcome to Vancouver 2.0 :: Photo Essay :: thetyee.ca: "There&#39;s a real opportunity to change &#39;busin.. http://bit.ly/16AgoI
Platinum One Destinations: A Home Based Business Thriving Amidst ...: Platinum One Destinations is a home based .. http://bit.ly/QSggI
Wells Fargo&#39;s John Stumpf is CEO of the Year: The company would not have been in a position to seize the opp.. http://bit.ly/kJLH1
Loans and growth | Free exchange | Economist.com: The demand for a loan is determined by an entrepreneur's risk .. http://bit.ly/AdATT
News | The Independent UK - The British Franchise Exhibition: A ...: Home ... offers the opportunity to build yo.. http://bit.ly/15CoJ8
Firm settles claim it banned Spanish - OC Business News ...: The U.S. Equal Employment Opportunity Commission su.. http://bit.ly/Tqby8
April 16, 2009 | next media update: Wireless carriers such as AT&T Corp. are setting their sights on so-call.. http://bit.ly/rRvw5
Find The Best Small Home Based Business: Small Home Based Business - Finding a secure small home based business .. http://bit.ly/3jONr
yz7654_nyu_edu@nyu-dataproc-m:~/final$ tail ingest_output/part-m-00007
Be Your Own Boss - Northeast, & Work From Home : Ecademy Marketplace: Work from home Northeast, is a Teessid.. http://bit.ly/ZS3GD
Grisly slayings brings Mexican drug war to US: That was their home base, and has been for a long time. Now,&quot;.. http://bit.ly/80kvj
A Name You Can Trust In Today's Wintery Economic Climate: I did a quick Google check and found that a leading ad.. http://bit.ly/L7XkI
5 Steps to Overcoming Fear and Getting Rich With Real Estate ...: Despite massive profit opportunities in the re.. http://bit.ly/flmzD
Easiest Way to Start a Home Based Business | BFX Media - Webmaster ...: It can be difficult to sort out all of y.. http://bit.ly/1a6uPX
Work-at-home scams prey on job hunters: Kansas City Star There are some real paycheck opportunities in such busi.. http://bit.ly/15Uu7q
Real Home Business | S-Proprietor.com: When you see this kind of search, it is in regards to an opportunity to w.. http://bit.ly/Urx1j
High-speed Internet means opportunity: Expanded high-speed service would also mean more opportunities for furthe.. http://bit.ly/KzhJB
Profit From Real Work At Home Jobs Easily | Redundancy Solutions: Working from home is one of the most popular a.. http://bit.ly/uAbdu
Maverick Money Makers Celebrates 11 Years of Home Based Business ...: Maverick Money Makers, the leading Home Ba.. http://bit.ly/8Xwt1
yz7654_nyu_edu@nyu-dataproc-m:~/final$
```

You can see the date time and user id and identifiers are all removed.

Data Profiling Process

This profiling is featured in `WordCountPercentage.java` and `compile_profile.sh`

It is a wordcount mapreduce program to capture all the count of our customized curse word list. The input is the output from the data ingest progress. The output is as follow:



```
Physical memory (bytes) snapshot=12012572672
Virtual memory (bytes) snapshot=38401286144
Total committed heap usage (bytes)=12443451392
Peak Map Physical memory (bytes)=1420550144
Peak Map Virtual memory (bytes)=4032628872
Peak Reduce Physical memory (bytes)=2356248576
Peak Reduce Virtual memory (bytes)=4816994304

Shuffle Errors
BAD_ID=0
CONNECTION=0
IO_ERROR=0
WRONG_LENGTH=0
WRONG_MAP=0
WRONG_REDUCE=0

File Input Format Counters
Bytes Read=619804743
File Output Format Counters
Bytes Written=290

Found 2 items
-rw-r--r-- 1 yz7654_nyu_edu yz7654_nyu_edu 0 2023-04-21 20:24 final_2436/profile_output/ SUCCESS
-rw-r--r-- 1 yz7654_nyu_edu yz7654_nyu_edu 290 2023-04-21 20:24 final_2436/profile_output/part-r-00000
yz7654_nyu_edu@nyu-dataproc-m:~$ cat profile_output/part-r-00000
cat: profile_output/part-r-00000: No such file or directory
yz7654_nyu_edu@nyu-dataproc-m:~$ cd final
yz7654_nyu_edu@nyu-dataproc-m:~/final$ cat profile_output/part-r-00000
ALL 115535322
ass 62251
bitch 23838
blowjob 98
cock 1033
cum 2869
cunt 429
dick 150
damn 70444
dick 7447
f**k 374
f**k 1158
fag 513
fuck 52570
fuck 11
hell 37346
n**** 42
n***a 32
nigga 40656
nigger 297
piss 2469
pussy 5324
puta 142
s**t 90
sh*t 1592
shit 89348
slut 727
tit 150
whore 1419
yz7654_nyu_edu@nyu-dataproc-m:~/final$
```

For example, there are 115535322 words in all tweets, and among them, 89348 are sh*t word, and we have a percentage of 0.0008%.