

1. Preprocessing:

Written a script that will preprocess the file.

- Merge 3 lines of user's info (T, U & W) to 1 line and separate them by “,”.
- Remove empty tweets
- Remove tweets with No Post Title
- Remove special chars from the tweets other than alphanumeric and @# but this happening in mapper

After preprocessing added the preprocessed file to hdfs

```
T 2009-06-07 02:07:42,U zaibatsu,W rt @agunp Twitter to roll out Verified Accounts this summer
T 2009-06-08 21:49:34,U alphaexe,W @lessallan I just played give it away in response to your tweet
T 2009-06-08 21:49:36,U digitalcrimeinv, W Apple Readies Snow Leopard for September, Will Charge 29
T 2009-06-08 21:49:36,U gabrielphb,W @LeOzInHoxP Ai cara vamo faze uma parceria entre nossos blogs Qualquer coisa me add no ms
T 2009-06-08 21:49:36,U gingerkid416,W @msali_sobb how work was today miss me
T 2009-06-08 21:49:36,U itscarol26,W de tanto meu irmão cantar smelly cat, eu fiquei com essa música na cabeça hahahaha meu ir
T 2009-06-08 21:49:36,U momfluential,W Just in case you missed my Celebrity Retail Therapy Du Jour. Go shop yourselves sane
T 2009-06-08 21:49:36,U ps_led_zeppelin,W Me duele la panz
```

2. Part1 - Find the users that tweet the most.

- For this task, I have written a custom writable class TwitterUser which will info of the user like a list of tweets, list of timestamps of that tweets and username and the total count of the tweet
- I have added a mapper and reducer which keep track of the total tweet count for the user and also will give the top 5 users with the most tweets

For running this MapReduce job you need to give the command:

```
yarn jar mr-1.0.jar edu.usfca.cs.mr.part_1_user_withmost_tweets.UserWithMostTweetsJob
/50_06_pre.txt /test01
```

where {mr-1.0.jar → jar name,s_pre6.txt → input file already on hdfs, test01→ output_dir of hdfs}

{input → 50k lines from a tweets2009-06.txt file and preprocessed} The output of the top 5 Twitter users with a list of their tweets and timestamp.

```

managerxing UserName --> managerxing Total Tweet count 27
ist of TimeStamps[ 2009-06-11 17:11:13, 2009-06-11 17:11:47, 2009-06-11 17:12:37, 2009-06-11 17:11:39, 2009-06-11 17:11:29, 2009-06-11 17:12:18, 2009-06-11 17:13:02, 2009-06-11 17:12:18, 2009-06-11 17:14:20, 2009-06-11 17:13:08, opeka, 2009-06-11 17:14:57, alladega, 2009-06-11 17:11:09, 2009-06-11 17:12:1, 2009-06-11 17:10:54, 2009-06-11 17:10:53, 2009-06-11 17:14:07, 2009-06-11 17:10:51, 2009-06-11 17:14:01, 2009-06-11 17:10:46, 2009-06-11 17:11:55, 2009-06-11 17:11:55, 2009-06-11 17:13:37, 2009-06-11 17:10:24, 2009-06-11 17:13:43, 2009-06-11 17:10:35, 2009-06-11 17:13:48, 2009-06-11 17:10:28]
ist of Tweets[#jobs#ManagerQualityAssuranceManager, #jobs#ManagerAssociateDirectorofWorshipArts, #jobs#ManagerRestaurantGeneralManagersAssistantManagers, #jobs#ManagerHealthcareFacilityAdministratorProgramManagerDavitaMidtownHomeAtlanhttp://inuricommekjz, #jobs#ManagerSupervisor, #jobs#ManagerMgrRetailSalesStLouis, #jobs#ManagerStoreManagerAnchorage, #jobs#ManagerAssistantMgrRetailSalesStLouis, #jobs#ManagerDirectorofEmergencyServices, #jobs#ManagerProjectManager, #jobs#ManagerAlumniAffairsDirector, #jobs#ManagerSTOREMANAGERKimbaliNE, #jobs#ManagerDirector, #jobs#ManagerStoreManagerPRICE, #jobs#ManagerStoreManagerPRICE, #jobs#ManagerFTBranchManagerGLADESLYONS, #jobs#ManagerMGRMedSurg, #jobs#ManagerSeniorDirectorofConstruction, #jobs#ManagerMGRMedSurg, #jobs#ManagerAssociateDirector, #jobs#ManagerAssociateDirector, #jobs#ManagerDirector, arwick, #jobs#ManagerHUMANRESOURCESMANAGERIIHUMANRESOURCEINFORMATIONSYSTEMSEM01EM03, #jobs#ManagerSupplyChainManager, #jobs#ManagerClinicDirectorII, #jobs#ManagerMgrClaimsLongTermDisabilityTeamLead, #jobs#ManagerSupervisor]

logisticsxing UserName --> logisticsxing Total Tweet count 18
ist of TimeStamps[ 2009-06-11 17:13:05, 2009-06-11 17:13:21, 2009-06-11 17:08:31, 2009-06-11 17:10:53, 2009-06-11 17:12:40, 2009-06-11 17:10:37, 2009-06-11 17:13:16, 2009-06-11 17:12:17, 2009-06-11 17:13:26, 2009-06-11 17:11:30, 2009-06-11 17:13:11, 2009-06-11 17:10:21, 2009-06-11 17:13:10, 2009-06-11 17:12:51, 2009-06-11 17:08:51, 2009-06-11 17:13:02, 2009-06-11 17:07:52, 2009-06-11 17:08:14]
ist of Tweets[#jobs#LogisticsSupervisoryLogisticsManagementSpecialist, #jobs#LogisticsLogisticsSupervisor, illiamsburg, #jobs#LogisticsLogistician2, #jobs#LogisticsFreightAgentsFreightBrokerBusinessOpportu, #jobs#LogisticsLeadLogisticsManagementSpecialist, arren, #jobs#LogisticsSrLogistician, #jobs#LogisticsFreightAgentsFreightBrokerBusinessOpportu, #jobs#LogisticsLogisticsSpecialist, #jobs#LogisticsLogisticsSupervisor, #jobs#LogisticsShippingManagerhighvolumeshippingoperation@globalco, #jobs#LogisticsSeniorLogisticsAnalystFortBelvoir, #jobs#LogisticsFreightAgents, #jobs#LogisticsSeniorLogisticsAnalystFortBelvoir, #jobs#LogisticsDomesticFreightAgentBroker, #jobs#LogisticsFreightAgentsFreightBrokerBusinessOpportunity, #jobs#LogisticsSupervisoryLogisticsManagementSpecialist, #jobs#LogisticsLogisticsSupervisor, #jobs#LogisticsMaterialSpecialistIIContainers]

twack1e_jon UserName --> twack1e_jon Total Tweet count 18
ist of TimeStamps[ 2009-06-11 17:14:19, 2009-06-11 17:12:44, 2009-06-11 17:05:57, 2009-06-11 17:13:28, 2009-06-11 17:08:28, 2009-06-11 17:11:40, 2009-06-11 17:07:52, 2009-06-11 17:11:39, 2009-06-11 17:09:52, 2009-06-11 17:11:29, 2009-06-11 17:09:12, 2009-06-11 17:06:35, 2009-06-11 17:08:37, 2009-06-11 17:13:03, 2009-06-11 17:12:37, 2009-06-11 17:14:56, 2009-06-11 17:08:50, 2009-06-11 17:13:11]
ist of Tweets[@redwingfan19checkouttweetsandtopstoriesforGame7http://bit.ly/4iEYN, @motleysuchekouttweetsandtopstoriesforGame7http://bit.ly/4iEYN, @Crosby87FansChecktweetsandtopstoriesforGame7http://bit.ly/4iEYN, @RoyBossCheckouttweetsandtopstoriesforGame7http://bit.ly/4iEYN, @tinybubblecheckouttweetsandtopstoriesforGame7http://bit.ly/4iEYN, @Shank721checkouttweetsandtopstoriesforGame7http://bit.ly/4iEYN, @TeddyYorkcheckouttweetsandtopstoriesforGame7http://bit.ly/4iEYN, @Shank721checkouttweetsandtopstoriesforGame7http://bit.ly/4iEYN, @geoffSurrattcheckouttweetsandtopstoriesforGame7http://bit.ly/4iEYN, @supermaniac06checkouttweetsandtopstoriesforGame7http://bit.ly/4iEYN, @twoFacedAngelcheckouttweetsandtopstoriesforGame7http://bit.ly/4iEYN, @harveyharvcheckouttweetsandtopstoriesforGame7http://bit.ly/4iEYN, @deanna8429checkouttweetsandtopstoriesforGame7http://bit.ly/4iEYN, @nebraskawhitcheckouttweetsandtopstoriesforGame7http://bit.ly/4iEYN, @motleysuchekouttweetsandtopstoriesforGame7http://bit.ly/4iEYN, @jhnkafcheckouttweetsandtopstoriesforGame7http://bit.ly/4iEYN, @MosteelerGcheckouttweetsandtopstoriesforGame7http://bit.ly/4iEYN, @KaylakristinecheckouttweetsandtopstoriesforGame7http://bit.ly/4iEYN]

nico_live UserName --> nico_live Total Tweet count 13
ist of TimeStamps[ 2009-06-11 17:10:14, 2009-06-11 17:13:03, 2009-06-11 17:12:51, 2009-06-11 17:12:40, 2009-06-11 17:12:20, 2009-06-11 17:13:15, 2009-06-11 17:11:54, 2009-06-11 17:11:48, 2009-06-11 17:13:21, 2009-06-11 17:11:22, 2009-06-11 17:10:46, 2009-06-11 17:10:35, 2009-06-11 17:14:00]
ist of Tweets[ONAIRSTARhttp://livenicovideojwatch/v1418906, ONAIRSTARhttp://livenicovideojwatch/v1418937, ONAIRSTARhttp://livenicovideojwatch/v1418935, ONAIRSTARhttp://livenicovideojwatch/v1418933, ONAIRSTARhttp://livenicovideojwatch/v1418929, ONAIRSTARhttp://livenicovideojwatch/v1418940, ONAIRSTARhttp://livenicovideojwatch/v1418925, ONAIRSTARhttp://livenicovideojwatch/v1418924, ONAIRSTARhttp://livenicovideojwatch/v1418941, ONAIRSTARhttp://livenicovideojwatch/v1418919, ONAIRSTARhttp://livenicovideojwatch/v1418913, ONAIRSTARhttp://livenicovideojwatch/v1418910, ONAIRSTARhttp://livenicovideojwatch/v1418947]

procurementxing UserName --> procurementxing Total Tweet count 10
ist of TimeStamps[ 2009-06-11 17:07:19, 2009-06-11 17:08:36, 2009-06-11 17:08:04, 2009-06-11 17:06:34, 2009-06-11 17:07:50, 2009-06-11 17:08:59, 2009-06-11 17:08:53, 2009-06-11 17:06:47, 2009-06-11 17:08:51, 2009-06-11 17:06:58]
ist of Tweets[#jobs#ProcurementProcurementCoordinator, estPalmbeach, #jobs#ProcurementProcurementAnalyst, #jobs#ProcurementSrGovernmentProcurementSpecialist, #jobs#ProcurementProcurementAnalyst, #jobs#ProcurementPROCUREMENTSPECIALIST3BatonRouge, #jobs#ProcurementProcurementContractsSpecialist, ashington, #jobs#ProcurementProcurementSubcontractComplianceSpecialist, #jobs#ProcurementContractsManagerProcurementSpartanburg, #jobs#ProcurementProcurementAnalyst, ashington, #jobs#ProcurementCommodityManagerPROLOONS]

jlakshmi@orion05 ~]$

```

3. Part2 - Find the top 5 hashtags for each week.

- For this task, I have written two custom writable class HashTag and HashTagByWeek which will have info about hashtags for each week
- I have added a mapper and reducer which keep track of the total hashtag count for the week and also will sort and give the top 5 hashtags for each week. I have added the key of week_year to group them together.

For running this MapReduce job you need to give the command:

```
yarn jar mr-1.0.jar edu.usfca.cs.mr.part_2_top_hashtag.TopHashtagJob /50_06_pre.txt /test02
```

where {mr-1.0.jar → jar name,s_pre6.txt → input file already on hdfs, test02→ output_dir of hdfs}

```

2020-11-05 20:37:47,500 INFO spark.SPARKStringUtils: SASE encryption trust check. Local trust store = false, Remote trust store = true
24_2009 HashTagName -->#durka TotalCount -->2
24_2009 HashTagName -->#spymaster TotalCount -->2
24_2009 HashTagName -->#failliet TotalCount -->2
24_2009 HashTagName -->#niley TotalCount -->2
24_2009 HashTagName -->#jobs TotalCount -->4

```

{input → 50k lines from a tweets2009-06.txt file and preprocessed} The output of the top 5 hashTags for each week.

4. **Part3 a) - Sentiment analysis to calculate the overall sentiment associated with the users**

- For this task, I have used a dataset AFINN.txt which has a list of positive and negative words with sentiment counts attached to each of them
- I have added a mapper and reducer which keep track of the total sentiment count for each tweet.

For running this MapReduce job you need to give the command:

```
yarn jar mr-1.0.jar  
edu.usfca.cs.mr.part_3_sentiment_analysis.TwitterUserSentimentAnalysisJob /AFINN-111.txt  
/s_pre6.txt /test03
```

where {mr-1.0.jar → jar name, AFINN-111.txt → dataset which contains a list of words with their sentiments, s_pre6.txt → input file already on hdfs, test03→ output_dir of hdfs}

{input → 30k lines from a tweets2009-06.txt file and preprocessed} The output contains a list of sentiment counts for each tweet by the user.

Part3 b) -Sentiment analysis to calculate the overall sentiment of hashtags.

- For this task, I have used a dataset AFINN.txt which has a list of positive and negative words with sentiment counts attached to each of them
- I have added a mapper and reducer which keep track of the total sentiment count for each hashTag.

For running this MapReduce job you need to give the command:

```
yarn jar mr-1.0.jar edu.usfca.cs.mr.part_3_sentiment_analysis.HashTagSentimentAnalysisJob  
/AFINN-111.txt /50_06_pre.txt /test03
```

where {mr-1.0.jar → jar name, AFINN-111.txt → dataset which contains a list of words with their sentiments, s_pre6.txt → input file already on hdfs, test03→ output_dir of hdfs}

```

2009-11-05 11:00:00,123.456 0.00
#ass -4
#comedy 1
#crazy -2
#fail -6
#fun 4
#funny 4
#lol 18
#love 6
#odd -4
#rofl 4
#safety 1
#sexy 3
#sunshine 2
#wtf -4

```

{input → 50k lines from a tweets2009-06.txt file and preprocessed} The output contains a list of sentiment counts for each hashtag.

5. Part4 - *Implemented top 5 mentioned user each year*

- For this task, I am getting the list of top 5 users who were most mentioned like @userName.
- I have added a mapper and reducer which keep track of the total count of Top mentioned user for each year.

For running this MapReduce job you need to give the command:

```
yarn jar mr-1.0.jar edu.usfca.cs.mr.part_4_top_mentioned_user.TopMentionedUserJob
/50_06_pre.txt /test04
```

where {mr-1.0.jar → jar name,s_pre6.txt → input file already on hdfs, test04→ output_dir of hdfs}

```

2009  Username: @debruynsdesign -> Total number of Mentions :2
2009  Username: @ituneiphone -> Total number of Mentions :2
2009  Username: @pedrojimenez -> Total number of Mentions :2
2009  Username: @Advocates4Youth -> Total number of Mentions :2
2009  Username: @RuthZ -> Total number of Mentions :2
File: hdfs://.../test04/2009-11-05 11:00:00,123.456 0.00

```

{input → 50k lines from a tweets2009-06.txt file and preprocessed} The output contains a list of the top 5 mentioned users counts for each year.

6. Part5 - *Final project (Amazon users reviews dataset) 12GB*

Dataset:

- Dataset: Amazon all product review dataset is around 12 GB (<https://snap.stanford.edu/data/web-Amazon.html>)
- But for this assignment, I am working on a subset of the complete dataset, “amazon_reviews_us_Personal_Care_Appliances_v1_00” which is 50 MB. I am analyzing reviews related to the product that belongs to the personal care appliances category.
- The dataset contains the following important info.
Customer_Id: Customer who has given the review
Review_id: Review Id for each review
Product_id: product id of that product
Product title: Short title of the product
Star_rating: Rating of that product

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	marketpla	customer_	review_id	product_id	product_p	product_t	product_c	star_rating	helpful_vo	total_vote	vine	verified_p	review_he	review_bo	review_date	
2	US	32114233	R1QX6706	B00OYRW	2.24E+08	Elite Sport	Personal_c	5	0	0	N	Y	Good qual	Exactly as	#####	
3	US	18125776	R3QWMLJ	B0000537	8.2E+08	Ezy Dose V	Personal_c	5	0	0	N	Y	Five Stars	It is great	#####	
4	US	19917519	R14Z1VR1	B00HXX03	8.49E+08	Pulse Oxim	Personal_c	5	1	1	N	Y	It's really r	It's really r	#####	
5	US	18277171	R25ZJL0C	B00EOB0J	7.01E+08	SE Tools T	Personal_c	2	0	0	N	Y	Two Stars	The kit wo	#####	
6	US	2593270	R3837KYH	B00OC2O1	7.94E+08	doTERRA I	Personal_c	4	0	1	N	Y	Four Stars	It works be	#####	
7	US	2592955	R2MN0QY	B00HES9C	3.19E+08	Viva Natur	Personal_c	5	0	0	N	Y	not bad at	I added to	#####	
8	US	15168265	R3AN2UJ1	B0016BFR	8.87E+08	Uncle Lee'	Personal_c	5	0	0	N	Y	Mild, enjoy	Husband d	#####	
9	US	13761624	R3U29ZLU	B00K504U	4.59E+08	Syrteny El	Personal_c	5	0	0	N	Y	Five Stars	Good qual	#####	
10	US	37070734	R16ZDMJJ	B00HES9C	3.19E+08	Viva Natur	Personal_c	5	0	0	N	N	High Quali	This is high	#####	
11	US	29615023	RRRDOEJZ	B00P6TUC	1.7E+08	Viva Natur	Personal_c	4	0	0	N	Y	we like it	Buying mo	#####	
12	US	47893062	R2KR5ZEA	B0006VJ6	4.13E+08	Body Back	Personal_c	5	0	0	N	Y	Five Stars	Their best	#####	
13	US	2582596	RR7PGQY7	B00H9L7V	8.51E+08	boostULTI	Personal_c	5	0	0	N	Y	Great proc	Great proc	#####	
14	US	21969415	RN37YYZB	B00P1JNZ	2.74E+08	doTERRA \	Personal_c	5	4	4	N	Y	Love the p	Love the p	#####	
15	US	43153609	R1UXGB7C	B00N5HD3	9.56E+08	Straight R	Personal_c	5	0	0	N	Y	Five Stars	great buy.	#####	
16	US	17782951	R1OC5ZNX	B0007DHN	78719480	BONGER M	Personal_c	5	0	0	N	Y	Five stars	These are	#####	
17	US	13710264	R1W4ZN8I	B0002JG2	9.01E+08	Home Hea	Personal_c	5	0	0	N	Y	It works fc	It works fc	#####	
18	US	30720884	R2KE33CN	B00OYRW	2.24E+08	Elite Sport	Personal_c	5	0	0	N	Y	These wor	These wor	#####	

Preprocessing:

- Dataset was provided in the TSV format so the first thing I did convert it to CSV comma-separated file for easy processing.
- Also, I tried removing reviews with empty product title
- Most of the dataset was clean so not much needed in the preprocessing step

Map Reduce 1: *Amazon Most Reviewed Product [personal_appliances category]*

- This MapReduce job will display the top 5 most reviewed product that belongs to personal appliances category from this dataset

For running this MapReduce job you need to give the command:

```
yarn jar mr-1.0.jar
```

```
edu.usfca.cs.mr.part_5_final_project_amazon_reviews.AmazonMostReviewedProductJob  
/amazon_reviews_us_Personal_Care_Appliances_v1_00.csv /test05
```

where {mr-1.0.jar → jar name,amazon_reviews_us_Personal_Care_Appliances_v1_00.csv → input file already on hdfs, test05→ output_dir of hdfs}

```
B00H9L7VIV Product Title: "boostULTIMATE - 60 Capsules - Increase Workout Stamina  
Total Reviews count:3918  
B0006VJ6TO Product Title: Body Back Company's Body Back Buddy Trigger Point Therapy Self Massage Tool - PARENT  
Total Reviews count:1821  
B00HES9CMS Product Title: "Viva Naturals #1 Best Selling Certified Organic Cacao Powder from Superior Criollo Beans  
Total Reviews count:1223  
B000SOQ30E Product Title: "MedMobile's BATHTUB TRANSFER BENCH / BATH CHAIR WITH BACK  
Total Reviews count:618  
B00HXX0332 Product Title: "Pulse Oximeter  
Total Reviews count:578  
B0073TX6IO Product Title: Pedi Spin As Seen on TV Pedispin  
Total Reviews count:553  
B002PL33AQ Product Title: "Zephyr HxM BT Wireless Heart Rate Sensor  
Total Reviews count:497  
B002ONHBBW Product Title: Parker SRW Stainless Steel Straight Edge Barber Razor & 100 Shark Super Stainless Blades  
Total Reviews count:458  
B001GGVKFG Product Title: Cervical Neck Traction  
Total Reviews count:435  
B003CDXJUK Product Title: Hearing Aid Battery Powerone size 10 made in Germany Genuine 60 Pack  
Total Reviews count:421
```

{input →Amazon reviews CSV file and preprocessed} The output contains a list of the top 5 amazon products reviewed by the customer.

Map Reduce 2: *Amazon Top rated Product [personal_appliances category]*

- This MapReduce job will display the top 5 rated product that belongs to the personal care appliances category from this dataset.

For running this MapReduce job you need to give the command:

yarn jar **mr-1.0.jar**

edu.usfca.cs.mr.part_5_final_project_amazon_reviews.TopRatedAmazonProductJob
/amazon_reviews_us_Personal_Care_Appliances_v1_00.csv /test05rated

```
98005      Product Title: Nannini SOS Reading Glasses +1.00
Page Rating: 5

98021      Product Title: Nannini SOS Reading Glasses +2.00
Page Rating: 5

02632      Product Title: CRYSTAL CLEAR GREEN BEADS With SILVER PLATED ROSARY CROSS & HOLY LAND SOIL MARIA ICON
Page Rating: 5

02829      Product Title: Real Olive Wood Beads from Jerusalem and Jesus Cross Crucifix Rosary & 2 Sides Center (1. Maria and Baby Jesus / 2. Jesus) - Brand New in Gift Box
Page Rating: 5

01066      Product Title: "Holy water 4 in1 (Water
Page Rating: 5

01244      Product Title: Dark Copper Mezuzah for House Door - Copper Decorations Jerusalem Western Wall and S.D.I. (Guardian of Doors)
Page Rating: 5

02079      Product Title: GOLDFILLED NECKLACE JEWELRY GOLD JERUSALEM CROSS
Page Rating: 5

03776      Product Title: Shema Israel With Magen David Kabbalah Evil Eye Red Leather Cord Bracelet
Page Rating: 5

14247      Product Title: "23cm/9"" Kabbalah Metal & Glass Flipping Diamond Rhombus Evil Eye Lucky Charm Wall/Car/Window Decor Protection"
Page Rating: 5

14271      Product Title: "20cm/8"" Kabbalah Metal & Glass Heart and Horse Evil Eye Lucky Charm Wall/Car/Window Decor Protection"
Page Rating: 5
```

where {mr-1.0.jar → jar name,amazon_reviews_us_Personal_Care_Appliances_v1_00.csv →
input file already on hdfs, test05rated→ output_dir of hdfs}