Title: Text mining on customer feedbacks of different laptop brands

Step 1:

Data Source: Real time data from reddit social media is scraped using python. Six subreddits communities namely SuggestAlaptop, Dell, AcerOfficial, mac, ASUS and thinkpad were chosen. Two csv files were created. One is with the comments and id.

	A	B	C	D	E	F	G	H	I	J	K
1	ID	COMMENTS									
2	fiwe37	Could we have a budget ballpark? It's quite difficult to suggest anything when we don't know if you can									
3	fiwe37	Not sure of the budget you are working with but on the low end you can look at the [**Lenovo Flex 14 8									
4	fiwe37	Hi, what is your budget? Are you purchasing in the U.S?									
5	fiwe37	I would recommend this [Acer Spin 5](https://www.amazon.com/dp/B07BRGNXB6/?tag=bkadamos_allt									
6	fitxh3	Super crazy idea, maybe consider an iPad? Good for taking notes and annotating papers, they have a gre									
7	fitxh3	Hi, are you open to refurbished or used laptops?									
8	fi2yzl	If you buy immediately (i.e within the next 38 hours) you can get [this HP Pavilion 15z](https://store.hp.									
9	fi2yzl	Hey,Best specs for the money can be translated into the most powerful specs - this means a gaming lapt									
10	fi2yzl	Go for the [**Acer A515-54G-5928**](https://www.amazon.com/Acer-Display-i5-8265U-GeForce-A515-5									
11	fi2yzl	Hi, if you're looking for a good keyboard, I think that calls for a ThinkPad. The [**Lenovo ThinkPad E595*									
12	fi2yzl	I would recommend this [Acer Aspire 5](https://www.amazon.com/dp/B07RF2123Z/?tag=bkadamos_all									
13	fhh9ug	Yes, it will shred CS:GO even on max. What the CPU number means:* **i7:** Intel's second fastest line									
14	fhh9ug	I'd recommend staying away from dell G3 as it has terrible heating issues. I'd rather recommend the Len									
15	fhh9ug	1: Mobile Ryzen 4000 is too close not to wait.https://www.notebookcheck.net/AMD-Ryzen-4000-laptop									
16	fhh9ug	The Gpu is not on par with the cpu and the 512 ssd is bare minimum so more storage would be good									

Figure 1: reddit_comments_hottag.csv

The second one is with 8 columns:

	A	B	C	D	E	F	G	H	I
1	TITLE	SCORE	ID	SUBREDDIT	URL	NUM_COMMENTS	BODY	CREATED	TIMESTAMP
2	READ THIS BEFORE POSTING O	44	6dwp6l	SuggestALaptop	https://w	0		1.5E+09	29/5/2017 14:49
3	/R/SuggestALaptop Stress Tes	16	ekkvox	SuggestALaptop	https://w	2	Laptops t	1.58E+09	6/1/2020 15:23
4	why is everything so thin?	17	fjx4l7	SuggestALaptop	https://w	7	Hi, I am lo	1.58E+09	17/3/2020 18:06
5	Under 1000â,¬ laptop for prog	14	fjsvgw	SuggestALaptop	https://w	4	* **Tot	1.58E+09	17/3/2020 13:33
6	Looking for a Mid-Range Laptc	5	fjy09j	SuggestALaptop	https://w	0	General ir	1.58E+09	17/3/2020 19:07
7	Under 1000	4	fjy7ix	SuggestALaptop	https://w	1	* **Total I	1.58E+09	17/3/2020 19:22
8	Question about a laptop	3	fjy6tm	SuggestALaptop	https://w	0	I'm lookin	1.58E+09	17/3/2020 19:21
9	Recommendations for audio/i	1	fk25ch	SuggestALaptop	https://w	0	* **Total I	1.58E+09	18/3/2020 1:11
10	Help me choose between the	1	fk24k0	SuggestALaptop	https://w	0	[Option 1]	1.58E+09	18/3/2020 1:09
11	Developer Laptop	8	fjsdd9	SuggestALaptop	https://w	3	* **Tot	1.58E+09	17/3/2020 13:03
12	MacBook Air got me through a	3	fjv23n	SuggestALaptop	https://w	1	* **Total I	1.58E+09	17/3/2020 15:47
13	Working from home	3	fjw9n2	SuggestALaptop	https://w	6	* **Total I	1.58E+09	17/3/2020 17:06
14	Cheap laptop to learn coding	1	fk0tp9	SuggestALaptop	https://w	0	Â KB* **T	1.58E+09	17/3/2020 23:01
15	In need of a laptop for college	2	fjxsxl	SuggestALaptop	https://w	1	* **Total I	1.58E+09	17/3/2020 18:53
16	Looking for a Mid-Range lapto	13	fjng02	SuggestALaptop	https://w	7	* **Total I	1.58E+09	17/3/2020 8:12
17	Need laptop with thunderbolt	2	fk0e9e	SuggestALaptop	https://w	1	\	1.58E+09	17/3/2020 22:22
18	Which is a good budget laptop	1	fjzzxb	SuggestALaptop	https://w	0	Hello fellc	1.58E+09	17/3/2020 21:45
19	Lightweight laptop for solidwo	6	fjsbhh	SuggestALaptop	https://w	13	* **Total I	1.58E+09	17/3/2020 13:00
20	Looking for a general purpose	2	fjwn4t	SuggestALaptop	https://w	0	* **Total I	1.58E+09	17/3/2020 17:32
21	Looking for budget student la	5	fjq7zf	SuggestALaptop	https://w	3	\-Around	1.58E+09	17/3/2020 10:56
22	School laptop - $900 USD	7	fjp6bz	SuggestALaptop	https://w	7	* **Total I	1.58E+09	17/3/2020 9:55
23	Recommendations for laptop	1	fjyyu9	SuggestALaptop	https://w	1	Hello All,I	1.58E+09	17/3/2020 20:12

Figure 2: reddit_hottag.csv

Step 2:

Data Cleaning: The acquired dataset from reddit have been cleaned using R language in RStudio. The screenshot of cleaned dataset is given below:

	A	B	C	D	E	F	G	H	I	J	K	L
1		ID	NEW_COMMENTS									
2	1	fiwe37	Could we have a budget ballpark It s quite difficult to suggest anything when we don t know if you can sp									
3	2	fiwe37	Not sure of the budget you are working with but on the low end you can look at the Lenovo Flex 14 81SS(									
4	3	fiwe37	Hi what is your budget Are you purchasing in the U S									
5	4	fiwe37	I would recommend this Acer Spin 5 since it matches most of your requirements as it has a full HD screer									
6	5	fitxh3	Super crazy idea maybe consider an iPad Good for taking notes and annotating papers they have a great l									
7	6	fitxh3	Hi are you open to refurbished or used laptops									
8	7	fi2yzl	If you buy immediately i e within the next 38 hours you can get this HP Pavilion 15z for 529 It comes with									
9	8	fi2yzl	Hey Best specs for the money can be translated into the most powerful specs this means a gaming laptop									
10	9	fi2yzl	Go for the Acer A515 54G 5928 which is a great value for money laptop for students Screen Size 15 inches									
11	10	fi2yzl	Hi if you re looking for a good keyboard I think that calls for a ThinkPad The Lenovo ThinkPad E595 is with									
12	11	fi2yzl	I would recommend this Acer Aspire 5 because of the following It offers great value for money since it c(									
13	12	fhh9ug	Yes it will shred CS GO even on max What the CPU number means i7 Intel s second fastest line of consum									
14	13	fhh9ug	I d recommend staying away from dell G3 as it has terrible heating issues I d rather recommend the Leno									
15	14	fhh9ug	1 Mobile Ryzen 4000 is too close not to wait 9750H runs hot on cooling much better than the G3 s									
16	15	fhh9ug	The Gpu is not on par with the cpu and the 512 ssd is bare minimum so more storage would be good									

Figure 3: clean_reddit_comments.csv

	SCORE	ID	SUBREDDIT	URL	NUM_COM	CREATED	TIMESTAMP	NEW_TITL	NEW_BODY
1	44	6dwp6l	SuggestALaptop	https://w	0	1.5E+09	29/5/2017 14:49	READ THIS	BEFORE POSTING OR COMMME
2	16	ekkvox	SuggestALaptop	https://w	2	1.58E+09	6/1/2020 15:23	R Suggest	Laptops this gen often have the
3	17	fjx4l7	SuggestALaptop	https://w	7	1.58E+09	17/3/2020 18:06	why is eve	Hi I am looking at laptops speci
4	14	fjsvgw	SuggestALaptop	https://w	4	1.58E+09	17/3/2020 13:33	Under 100	Total budget and country of pur
5	5	fjy09j	SuggestALaptop	https://w	0	1.58E+09	17/3/2020 19:07	Looking fo	General info I m a working adul
6	4	fjy7ix	SuggestALaptop	https://w	1	1.58E+09	17/3/2020 19:22	Under 100	Total budget and country of pur
7	3	fjy6tm	SuggestALaptop	https://w	0	1.58E+09	17/3/2020 19:21	Question	I m looking for an opinion on w
8	1	fk25ch	SuggestALaptop	https://w	0	1.58E+09	18/3/2020 1:11	Recomme	Total budget and country of pur
9	1	fk24k0	SuggestALaptop	https://w	0	1.58E+09	18/3/2020 1:09	Help me c	Option 1 2 3 The titles a bit misl
10	8	fjsdd9	SuggestALaptop	https://w	3	1.58E+09	17/3/2020 13:03	Develope	Total budget and country of pur
11	3	fjv23n	SuggestALaptop	https://w	1	1.58E+09	17/3/2020 15:47	MacBook /	Total budget and country of pur
12	3	fjw9n2	SuggestALaptop	https://w	6	1.58E+09	17/3/2020 17:06	Working f	Total budget and country of pur
13	1	fk0tp9	SuggestALaptop	https://w	0	1.58E+09	17/3/2020 23:01	Cheap lap	A KB Total budget and country (
14	2	fjxsxl	SuggestALaptop	https://w	1	1.58E+09	17/3/2020 18:53	In need of	Total budget and country of pur
15	13	fjng02	SuggestALaptop	https://w	7	1.58E+09	17/3/2020 8:12	Looking fo	Total budget and country of pur
16	2	fk0e9e	SuggestALaptop	https://w	1	1.58E+09	17/3/2020 22:22	Need lapt	x200B Total budget and country
17	1	fjzzxb	SuggestALaptop	https://w	0	1.58E+09	17/3/2020 21:45	Which is a	Hello fellow redditors Basically
18	6	fjsbhh	SuggestALaptop	https://w	13	1.58E+09	17/3/2020 13:00	Lightweig	Total budget and country of pur

Figure 4: clean_Reddit.csv

Step 3:

Store Data in Hive data warehouse:

Apache Hive is a data warehouse infrastructure that facilitates querying and managing large data sets which resides in distributed storage system. It is built on top of Hadoop and developed by Facebook. **Hive** provides a way to query the data using a SQL-like query language called **HiveQL (Hive query Language).** Internally, a compiler translates **HiveQL** statements into **MapReduce** jobs, which are then submitted to **Hadoop framework** for execution.

1) In order to store data in hive, Hadoop environment is needed. Through Virtual machine, ubuntu operating system has been installed and then in ubuntu Hadoop and hive is installed respectably.

2) First, with start-all.sh command the Hadoop environment is set up, and then 'jps' command which is a tool to check, whether expected Hadoop processes are up and in running state or not.

```
sidratul@sidratul-VirtualBox:~$ start-all.sh
This script is Deprecated. Instead use start-dfs.sh and start-yarn.sh
Starting namenodes on [localhost]
localhost: starting namenode, logging to /usr/local/hadoop/hadoop-2.10.0/logs/h
adoop-sidratul-namenode-sidratul-VirtualBox.out
localhost: starting datanode, logging to /usr/local/hadoop/hadoop-2.10.0/logs/h
adoop-sidratul-datanode-sidratul-VirtualBox.out
Starting secondary namenodes [0.0.0.0]
0.0.0.0: starting secondarynamenode, logging to /usr/local/hadoop/hadoop-2.10.0
/logs/hadoop-sidratul-secondarynamenode-sidratul-VirtualBox.out
starting yarn daemons
starting resourcemanager, logging to /usr/local/hadoop/hadoop-2.10.0/logs/yarn-
sidratul-resourcemanager-sidratul-VirtualBox.out
localhost: starting nodemanager, logging to /usr/local/hadoop/hadoop-2.10.0/log
s/yarn-sidratul-nodemanager-sidratul-VirtualBox.out
sidratul@sidratul-VirtualBox:~$ jps
2578 NameNode
6965 Jps
3255 ResourceManager
2968 SecondaryNameNode
2760 DataNode
3422 NodeManager
sidratul@sidratul-VirtualBox:~$
```

3) Creating a directory called 'milestone_2/input/'.

```
sidratul@sidratul-VirtualBox:~$ hadoop fs -mkdir /milestone2/input
sidratul@sidratul-VirtualBox:~$
```

4) The datasets are placed at the desired directory.

```
sidratul@sidratul-VirtualBox:~$ hadoop fs -put /home/sidratul/Downloads/clean_r
eddit_hottag.csv /milestone_2/input
sidratul@sidratul-VirtualBox:~$ hadoop fs -ls /milestone_2/input
```

```
sidratul@sidratul-VirtualBox:~$ hadoop fs -put /home/sidratul/Downloads/clean_r
eddit_hottag.csv /milestone_2/input
sidratul@sidratul-VirtualBox:~$ hadoop fs -ls /milestone_2/input
Found 2 items
-rw-r--r--   4 sidratul supergroup     267204 2020-03-19 21:54 /milestone_2/inp
ut/clean_reddit_comments_hottag.csv
-rw-r--r--   4 sidratul supergroup    4260523 2020-03-19 23:31 /milestone_2/inp
ut/clean_reddit_hottag.csv
sidratul@sidratul-VirtualBox:~$
```

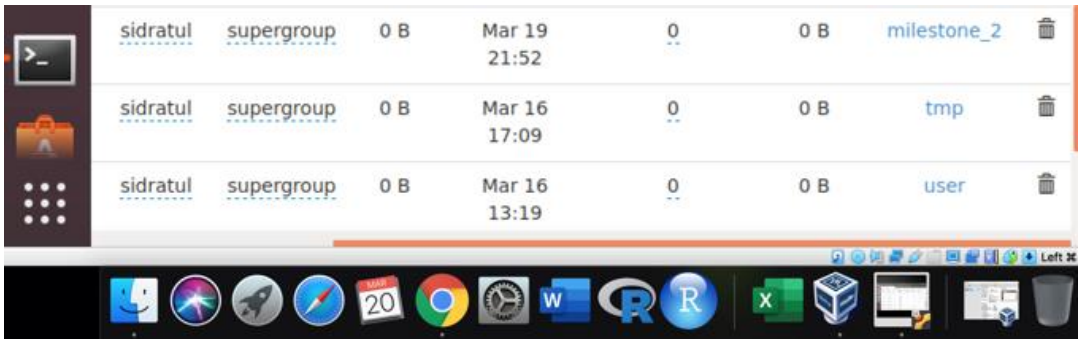
5) Initializing hive.

```
sidratul@sidratul-VirtualBox:~$ cd /usr/local/apache-hive-3.1.2-bin/bin
sidratul@sidratul-VirtualBox:/usr/local/apache-hive-3.1.2-bin/bin$ hive
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/local/apache-hive-3.1.2-bin/lib/log4j-sl
f4j-impl-2.10.0.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/local/hadoop/hadoop-2.10.0/share/hadoop/
common/lib/slf4j-log4j12-1.7.25.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation
```

6) Creating Table to put datasets on hive.

```
hive> CREATE SCHEMA IF NOT EXISTS milestone_2;
OK
Time taken: 1.103 seconds
hive> CREATE EXTERNAL TABLE IF NOT EXISTS milestone_2.clean_reddit_hottag_table
(score STRING,id STRING,subreddit STRING,url STRING,num_comments STRING,created
 STRING,`timestamp` STRING,new_title STRING,new_body STRING) ROW FORMAT DELIMIT
ED FIELDS TERMINATED BY ',' STORED AS TEXTFILE LOCATION '/milestone_2/input';
OK
Time taken: 1.643 seconds
hive>
```

7) Created directory on hdfs:



	Size	Last Modified	Replication	Block Size	Name	
oup	260.94 KB	Mar 19 21:54	4	128 MB	clean_reddit_comments_hottag.csv	
oup	4.06 MB	Mar 19 23:31	4	128 MB	clean_reddit_hottag.csv	

to 2 of 2 entries

1 Next