

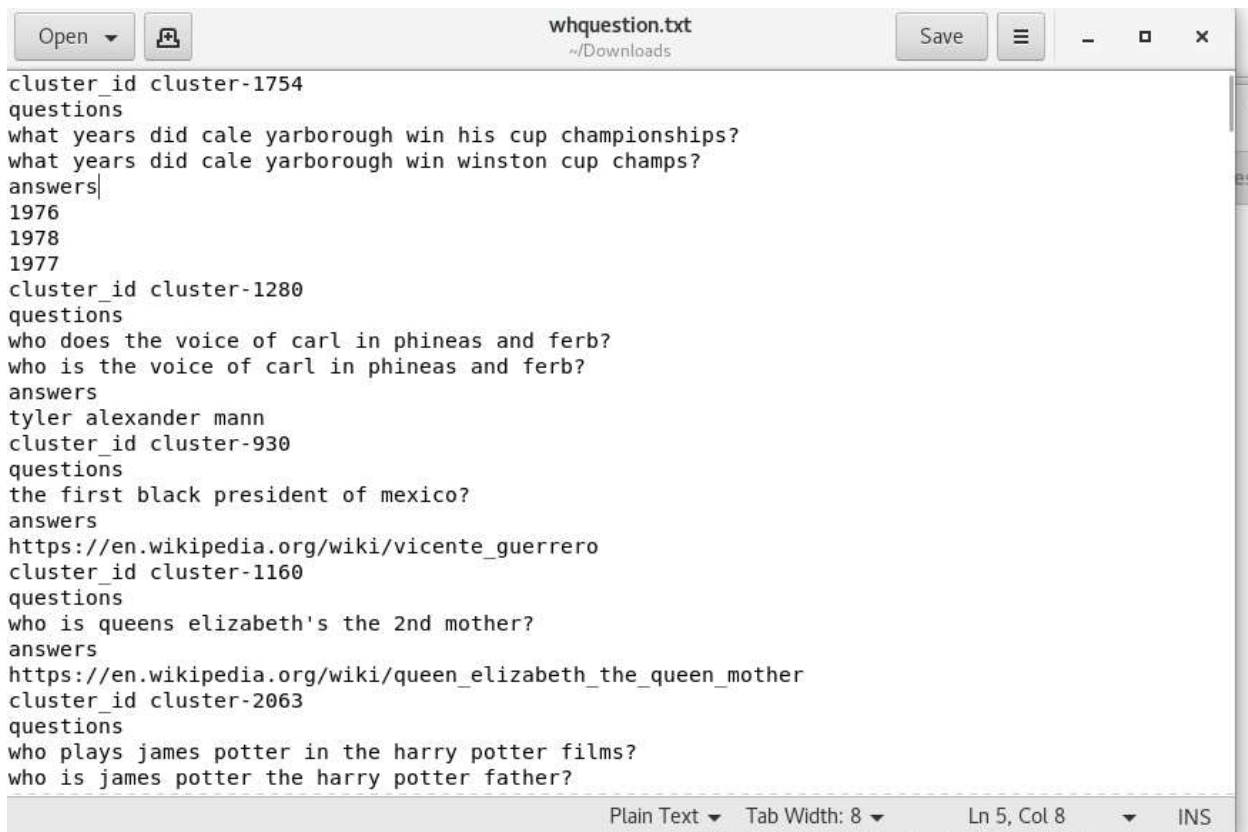
Done By: Esraa Ahmed Fouad Omar # ID: 20399123

Assignment #: 2

Done by: Esraa Khaled Ahmed Fouad Omar # ID: 20399123

January 23, 2023

Get the data from the following link https://qa.mpi-inf.mpg.de/comqa/comqa_train.json. Then convert this file to txt file using this website <https://onlinejsontools.com/convert-json-to-text>. Finally download the text file. As you can see, this is our whquestions.txt file after converting the content.



```
cluster_id cluster-1754
questions
what years did cale yarborough win his cup championships?
what years did cale yarborough win winston cup champs?
answers|
1976
1978
1977
cluster_id cluster-1280
questions
who does the voice of carl in phineas and ferb?
who is the voice of carl in phineas and ferb?
answers
tyler alexander mann
cluster_id cluster-930
questions
the first black president of mexico?
answers
https://en.wikipedia.org/wiki/vicente_guerrero
cluster_id cluster-1160
questions
who is queens elizabeth's the 2nd mother?
answers
https://en.wikipedia.org/wiki/queen_elizabeth_the_queen_mother
cluster_id cluster-2063
questions
who plays james potter in the harry potter films?
who is james potter the harry potter father?
```

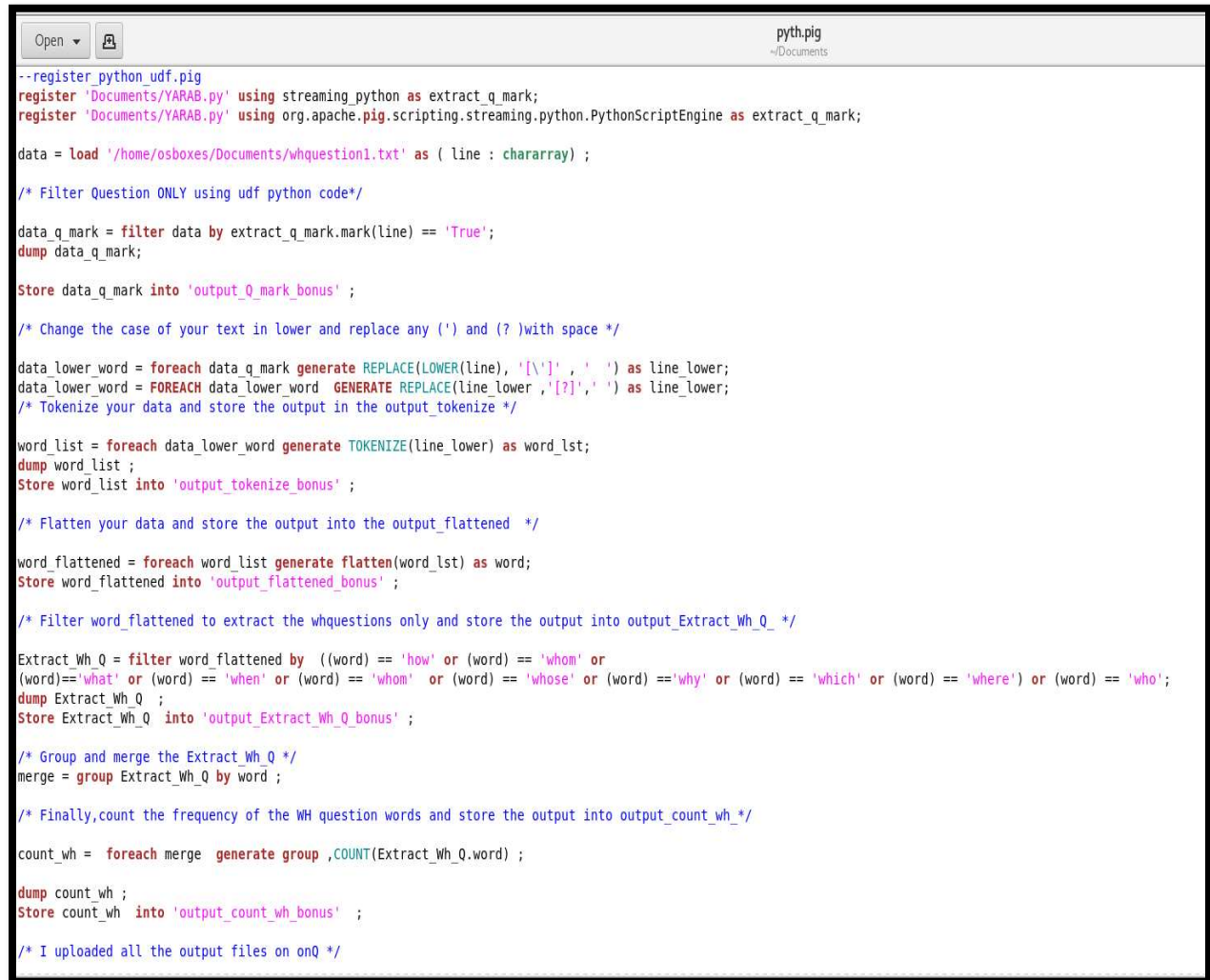
Done By: Esraa Ahmed Fouad Omar # ID: 20399123

Copy whquestions.txt from local virtual machine to HDFS and read it. As you can see, we copy the file and read it successfully.

```
#Name: Esraa Ahmed Fouad Omar # ID :20399123
File Edit View Search Terminal Help
[osboxes@quickstart-bigdata ~]$ #Name: Esraa Ahmed Fouad Omar # ID :20399123
[osboxes@quickstart-bigdata ~]$ hdfs dfs -put /home/osboxes/Downloads/whquestion.txt /user/osboxes/inputdata
[osboxes@quickstart-bigdata ~]$ hdfs dfs -ls /user/osboxes/inputdata/
Found 1 items
-rw-r--r-- 3 osboxes osboxes 723331 2023-01-27 02:22 /user/osboxes/inputdata/whquestion.txt
[osboxes@quickstart-bigdata ~]$ hdfs dfs -ls /user/osboxes
Found 8 items
drwx----- - osboxes osboxes 0 2023-01-27 02:20 /user/osboxes/.Trash
drwxr-xr-x - osboxes osboxes 0 2023-01-10 20:39 /user/osboxes/.sparkStaging
drwx----- - osboxes osboxes 0 2023-01-27 02:00 /user/osboxes/.staging
drwxr-xr-x - osboxes osboxes 0 2023-01-27 02:22 /user/osboxes/inputdata
drwxr-xr-x - osboxes osboxes 0 2023-01-25 02:49 /user/osboxes/output
-rw-r--r-- 3 osboxes osboxes 526283 2023-01-26 14:53 /user/osboxes/sales_data_sample.csv
drwxrwxrwx - osboxes osboxes 0 2023-01-25 00:52 /user/osboxes/stocks
-rw-r--r-- 3 osboxes osboxes 502 2023-01-23 18:53 /user/osboxes/whquestions.txt
[osboxes@quickstart-bigdata ~]$ hdfs dfs -cat /user/osboxes/inputdata/whquestion.txt |head
cluster_id cluster-1754
questions
what years did cale yarborough win his cup championships?
what years did cale yarborough win winston cup champs?
answers
1976
1978
1977
cluster_id cluster-1280
questions
cat: Unable to write to output stream.
[osboxes@quickstart-bigdata ~]$
```

Done By: Esraa Ahmed Fouad Omar # ID: 20399123

we create a script and called whquestions.pig that contains all the instructions we need to complete our task.



```
--register python_udf.pig
register 'Documents/YARAB.py' using streaming_python as extract_q_mark;
register 'Documents/YARAB.py' using org.apache.pig.scripting.streaming.python.PythonScriptEngine as extract_q_mark;

data = load '/home/osboxes/Documents/whquestion1.txt' as ( line : chararray ) ;

/* Filter Question ONLY using udf python code*/

data_q_mark = filter data by extract_q_mark.mark(line) == 'True';
dump data_q_mark;

Store data_q_mark into 'output_Q_mark_bonus' ;

/* Change the case of your text in lower and replace any (') and (?) with space */

data_lower_word = foreach data_q_mark generate REPLACE(LOWER(line), '['''] , ' ') as line_lower;
data_lower_word = FOREACH data_lower_word GENERATE REPLACE(line_lower , '['''] , ' ') as line_lower;
/* Tokenize your data and store the output in the output_tokenize */

word_list = foreach data_lower_word generate TOKENIZE(line_lower) as word_list;
dump word_list ;
Store word_list into 'output_tokenize_bonus' ;

/* Flatten your data and store the output into the output_flattened */

word_flattened = foreach word_list generate flatten(word_list) as word;
Store word_flattened into 'output_flattened_bonus' ;

/* Filter word_flattened to extract the whquestions only and store the output into output_Extract_Wh_Q */

Extract_Wh_Q = filter word_flattened by ((word) == 'how' or (word) == 'whom' or
(word)=='what' or (word) == 'when' or (word) == 'whom' or (word) == 'whose' or (word) == 'why' or (word) == 'which' or (word) == 'where' or (word) == 'who';
dump Extract_Wh_Q ;
Store Extract_Wh_Q into 'output_Extract_Wh_Q_bonus' ;

/* Group and merge the Extract_Wh_Q */
merge = group Extract_Wh_Q by word ;

/* Finally, count the frequency of the WH question words and store the output into output_count_wh */

count_wh = foreach merge generate group , COUNT(Extract_Wh_Q.word) ;

dump count_wh ;
Store count_wh into 'output_count_wh_bonus' ;

/* I uploaded all the output files on onQ */
```

Run the previous scrip into the terminal using the following command.



```
Name: Esraa Ahmed Fouad Omar Id : 20399123

File Edit View Search Terminal Help

2023-01-29 09:08:20,093 [main] INFO org.apache.pig.Main - Pig script completed in 27 seconds and 855 milliseconds (27855 ms)
[osboxes@quickstart-bigdata ~]$ #Name: Esraa Ahmed Fouad Omar # ID :20399123
[osboxes@quickstart-bigdata ~]$ pig -x local Documents/pyth.pig
```

Done By: Esraa Ahmed Fouad Omar # ID: 20399123

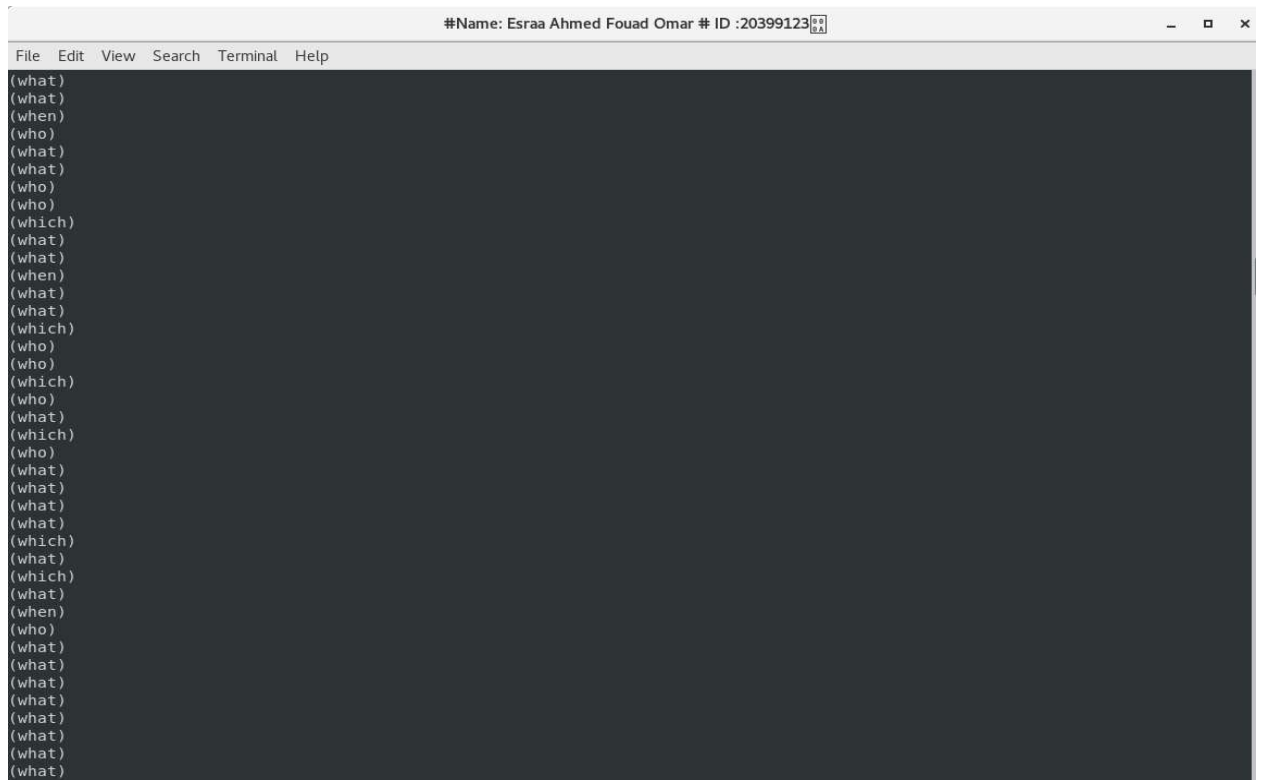
Let's see sample of our output.

- Tokenization step

```
Name: Esraa Ahmed Fouad Omar Id : 20399123
File Edit View Search Terminal Help
{{(what),(is),(trey),(songz),(first),(love),(name)}}
{{(what),(are),(george),(bush),(pets)}}
{{(when),(did),(the),(2002),(world),(cup),(take),(place)}}
{{(when),(did),(pope),(pius),(xi),(die)}}
{{(who),(was),(the),(leader),(for),(graet),(britain),(in),(ww1)}}
{{(tamela),(j),(mann),(and),(david),(mann),(are),(married)}}
{{(who),(is),(david),(mann),(and),(tamela),(mann),(married)}}
{{(what),(are),(hans),(crisian),(andersen),(s),(famous),(story),(books)}}
{{(who),(was),(mexico),(s),(first),(indigenous),(president)}}
{{(who),(did),(john),(wilkes),(booth)}}
{{(what),(is),(the),(largest),(city),(in),(romania)}}
{{(what),(is),(biggest),(city),(in),(romania)}}
{{(what),(is),(romania),(largest),(city)}}
{{(largest),(city),(in),(romania)}}
{{(romania),(largest),(city)}}
{{(what),(river),(is),(located),(south),(of),(tigris),(river),(and),(flows),(eat),(into),(the),(persian),(gulf)}}
{{(what),(is),(harvard),(university),(population)}}
{{(what),(is),(the),(population),(for),(harvard),(university)}}
{{(what),(country),(separates),(russia),(and),(pakistan)}}
{{(this),(country),(separates),(pakistan),(and),(the),(soviet),(union)}}
{{(this),(country),(lies),(between),(pakistan),(and),(russia)}}
{{(what),(country),(borders),(both),(pakistan),(and),(russia)}}
{{(what),(is),(the),(first),(action),(film),(made)}}
{{(who),(is),(the),(husband),(of),(mariah),(carey)}}
{{(who),(is),(mariah),(careys),(husband)}}
{{(what),(is),(mariah),(carey),(husband),(name)}}
{{(mariah),(carey),(husband)}}
{{(mariah),(carey),(spouse)}}
{{(who),(was),(the),(captain),(for),(france),(soccer),(team)}}
{{(who),(is),(the),(captain),(of),(the),(french),(national),(soccer),(team)}}
{{(who),(is),(the),(captain),(of),(the),(french),(soccer),(team)}}
{{(who),(is),(the),(french),(soccer),(captain)}}
{{(what),(day),(did),(harry),(s),(truman),(get),(married)}}
{{(when),(did),(harry),(truman),(get),(married)}}
{{(who),(were),(janet),(jackson),(s),(husbands)}}
{{(who),(did),(the),(miami),(dolphins),(beat),(to),(get),(their),(second),(superbowl),(win)}}
{{(first),(human),(in),(mars)}}
{{(the),(first),(human),(to),(go),(to),(the),(planet),(mars)}}
{{(who),(was),(the),(first),(actor),(to),(play),(superman),(on),(tv)}}
{{(who),(is),(older),(hillary),(or),(haylie),(duff)}}
{{(what),(is),(canadas),(major),(language)}}
{{(what),(movies),(are),(michael),(angarano),(on)}}
{{(every),(movie),(michael),(angarano),(has),(been),(in)}}
{{(what),(films),(did),(micheal),(angarano),(star),(in)}}
{{(what),(films),(did),(michael),(angarano),(play),(in)}}
{{(what),(movies),(has),(michael),(angarano),(starred),(in)}}
{{(what),(are),(the),(movies),(michael),(angarano),(has),(been),(in)}}
```

- Filter step

Done By: Esraa Ahmed Fouad Omar # ID: 20399123



A screenshot of a terminal window with a title bar that reads "#Name: Esraa Ahmed Fouad Omar # ID :20399123". The window has a menu bar with "File", "Edit", "View", "Search", "Terminal", and "Help". The terminal content consists of a vertical list of words in parentheses: (what), (what), (when), (who), (what), (what), (who), (who), (which), (what), (what), (when), (what), (what), (which), (who), (who), (which), (who), (what), (which), (who), (what), (what), (what), (what), (which), (what), (which), (what), (when), (who), (what), (what), (what), (what), (what), (what), (what), (what).

Done By: Esraa Ahmed Fouad Omar # ID: 20399123

Our Final output (The count of the frequency of the wh Questions)

```

Name: Esraa Ahmed Fouad Omar Id : 20399123
File Edit View Search Terminal Help
Successfully read 25508 records from: "/home/osboxes/Documents/whquestion1.txt"

Output(s):
Successfully stored 9 records in: "file:/tmp/temp-1890671666/tmp1152099865"

Counters:
Total records written : 9
Total bytes written : 0
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_local320168179_0007

2023-01-29 09:08:16,453 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialize
!
2023-01-29 09:08:16,455 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialize
!
2023-01-29 09:08:16,456 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTracker metrics system already initialize
!
2023-01-29 09:08:16,457 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2023-01-29 09:08:16,458 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2023-01-29 09:08:16,477 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input files to process : 1
2023-01-29 09:08:16,477 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(how,24)
(who,1721)
(why,1)
(what,3736)
(when,964)
(whom,8)
(where,361)
(which,613)
(whose,8)

```