SI618: Lab5

Name: Junqi Chen

Uniqname: junqich

Date: Oct. 3rd, 2021

- **Question 1**

```
[junqich@cavium-thunderx-login01 si618CaviumSetup]$ ls
ngram-job.py  spark-run.sh
```

Two files called `ngram-job.py` and `spark-run.sh` are in the directory.

- **Question 2**

```
[junqich@cavium-thunderx-login01 ~]$ hadoop fs -ls /var/umsi618f21/lab5/ngrams/data/
Found 39 items
-rw-r-----   1 dmal umsi618f21  192403080 2021-09-27 22:10 /var/umsi618f21/lab5/ngrams/data/googlebooks-eng-all-1gram-20120701-0
-rw-r-----   1 dmal umsi618f21 1479686629 2021-09-27 22:11 /var/umsi618f21/lab5/ngrams/data/googlebooks-eng-all-1gram-20120701-1
-rw-r-----   1 dmal umsi618f21  605024382 2021-09-27 22:11 /var/umsi618f21/lab5/ngrams/data/googlebooks-eng-all-1gram-20120701-2
-rw-r-----   1 dmal umsi618f21  425135677 2021-09-27 22:11 /var/umsi618f21/lab5/ngrams/data/googlebooks-eng-all-1gram-20120701-3
-rw-r-----   1 dmal umsi618f21  326914094 2021-09-27 22:11 /var/umsi618f21/lab5/ngrams/data/googlebooks-eng-all-1gram-20120701-4
-rw-r-----   1 dmal umsi618f21  272533631 2021-09-27 22:11 /var/umsi618f21/lab5/ngrams/data/googlebooks-eng-all-1gram-20120701-5
-rw-r-----   1 dmal umsi618f21  231213011 2021-09-27 22:11 /var/umsi618f21/lab5/ngrams/data/googlebooks-eng-all-1gram-20120701-6
-rw-r-----   1 dmal umsi618f21  202691966 2021-09-27 22:11 /var/umsi618f21/lab5/ngrams/data/googlebooks-eng-all-1gram-20120701-7
-rw-r-----   1 dmal umsi618f21  182906620 2021-09-27 22:11 /var/umsi618f21/lab5/ngrams/data/googlebooks-eng-all-1gram-20120701-8
-rw-r-----   1 dmal umsi618f21  165073935 2021-09-27 22:11 /var/umsi618f21/lab5/ngrams/data/googlebooks-eng-all-1gram-20120701-9
-rw-r-----   1 dmal umsi618f21 1801526075 2021-09-27 22:12 /var/umsi618f21/lab5/ngrams/data/googlebooks-eng-all-1gram-20120701-a
-rw-r-----   1 dmal umsi618f21 1268392934 2021-09-27 22:12 /var/umsi618f21/lab5/ngrams/data/googlebooks-eng-all-1gram-20120701-b
-rw-r-----   1 dmal umsi618f21 2090710388 2021-09-27 22:12 /var/umsi618f21/lab5/ngrams/data/googlebooks-eng-all-1gram-20120701-c
-rw-r-----   1 dmal umsi618f21 1252213884 2021-09-27 22:12 /var/umsi618f21/lab5/ngrams/data/googlebooks-eng-all-1gram-20120701-d
-rw-r-----   1 dmal umsi618f21 1085415448 2021-09-27 22:13 /var/umsi618f21/lab5/ngrams/data/googlebooks-eng-all-1gram-20120701-e
-rw-r-----   1 dmal umsi618f21  959470924 2021-09-27 22:13 /var/umsi618f21/lab5/ngrams/data/googlebooks-eng-all-1gram-20120701-f
-rw-r-----   1 dmal umsi618f21  823166881 2021-09-27 22:13 /var/umsi618f21/lab5/ngrams/data/googlebooks-eng-all-1gram-20120701-g
-rw-r-----   1 dmal umsi618f21  948615440 2021-09-27 22:13 /var/umsi618f21/lab5/ngrams/data/googlebooks-eng-all-1gram-20120701-h
-rw-r-----   1 dmal umsi618f21 1093823911 2021-09-27 22:13 /var/umsi618f21/lab5/ngrams/data/googlebooks-eng-all-1gram-20120701-i
-rw-r-----   1 dmal umsi618f21  327435021 2021-09-27 22:14 /var/umsi618f21/lab5/ngrams/data/googlebooks-eng-all-1gram-20120701-j
-rw-r-----   1 dmal umsi618f21  547335615 2021-09-27 22:14 /var/umsi618f21/lab5/ngrams/data/googlebooks-eng-all-1gram-20120701-k
-rw-r-----   1 dmal umsi618f21  959686094 2021-09-27 22:14 /var/umsi618f21/lab5/ngrams/data/googlebooks-eng-all-1gram-20120701-l
-rw-r-----   1 dmal umsi618f21 1501649198 2021-09-27 22:14 /var/umsi618f21/lab5/ngrams/data/googlebooks-eng-all-1gram-20120701-m
-rw-r-----   1 dmal umsi618f21  730118203 2021-09-27 22:14 /var/umsi618f21/lab5/ngrams/data/googlebooks-eng-all-1gram-20120701-n
-rw-r-----   1 dmal umsi618f21  732438650 2021-09-27 22:14 /var/umsi618f21/lab5/ngrams/data/googlebooks-eng-all-1gram-20120701-o
-rw-r-----   1 dmal umsi618f21    1828173 2021-09-27 22:14 /var/umsi618f21/lab5/ngrams/data/googlebooks-eng-all-1gram-20120701-other
-rw-r-----   1 dmal umsi618f21 1900898850 2021-09-27 22:15 /var/umsi618f21/lab5/ngrams/data/googlebooks-eng-all-1gram-20120701-p
-rw-r-----   1 dmal umsi618f21      99864 2021-09-27 22:15 /var/umsi618f21/lab5/ngrams/data/googlebooks-eng-all-1gram-20120701-pos
-rw-r-----   1 dmal umsi618f21  178176943 2021-09-27 22:15 /var/umsi618f21/lab5/ngrams/data/googlebooks-eng-all-1gram-20120701-punctuation

-rw-r-----   1 dmal umsi618f21  136197316 2021-09-27 22:15 /var/umsi618f21/lab5/ngrams/data/googlebooks-eng-all-1gram-20120701-q
-rw-r-----   1 dmal umsi618f21 1137454640 2021-09-27 22:15 /var/umsi618f21/lab5/ngrams/data/googlebooks-eng-all-1gram-20120701-r
-rw-r-----   1 dmal umsi618f21 2316331839 2021-09-27 22:16 /var/umsi618f21/lab5/ngrams/data/googlebooks-eng-all-1gram-20120701-s
-rw-r-----   1 dmal umsi618f21 1383305366 2021-09-27 22:16 /var/umsi618f21/lab5/ngrams/data/googlebooks-eng-all-1gram-20120701-t
-rw-r-----   1 dmal umsi618f21  466747550 2021-09-27 22:16 /var/umsi618f21/lab5/ngrams/data/googlebooks-eng-all-1gram-20120701-u
-rw-r-----   1 dmal umsi618f21  560636053 2021-09-27 22:16 /var/umsi618f21/lab5/ngrams/data/googlebooks-eng-all-1gram-20120701-v
-rw-r-----   1 dmal umsi618f21  612797788 2021-09-27 22:16 /var/umsi618f21/lab5/ngrams/data/googlebooks-eng-all-1gram-20120701-w
-rw-r-----   1 dmal umsi618f21   70121491 2021-09-27 22:16 /var/umsi618f21/lab5/ngrams/data/googlebooks-eng-all-1gram-20120701-x
-rw-r-----   1 dmal umsi618f21  129575526 2021-09-27 22:16 /var/umsi618f21/lab5/ngrams/data/googlebooks-eng-all-1gram-20120701-y
-rw-r-----   1 dmal umsi618f21  135433431 2021-09-27 22:16 /var/umsi618f21/lab5/ngrams/data/googlebooks-eng-all-1gram-20120701-z
```

The name of the last file in the listing for HFS folder `/var/umsi618f21/lab5/ngrams/data/` is `googlebooks-eng-all-1gram-20120701-z`.

- **Question 3**

```
[junqich@cavium-thunderx-login01 ~]$ hadoop fs -cat /var/umsi618f21/lab5/ngrams/data/* | grep "^information_NOUN"
information_NOUN        1505    1       1
information_NOUN        1507    5       1
information_NOUN        1515    105     1
information_NOUN        1524    69      1
information_NOUN        1525    2       1
information_NOUN        1563    28      1
information_NOUN        1564    8       1
information_NOUN        1568    1       1
information_NOUN        1572    2       1
information_NOUN        1574    10      1
information_NOUN        1575    32      1
information_NOUN        1579    7       2
information_NOUN        1581    14      2
information_NOUN        1582    1       1
```

It was in the year **1505** that “information” first mentioned (as a noun) in Google Books data.

- **Question 4**

```
[junqich@cavium-thunderx-login01 si618CaviumSetup]$ hadoop fs -ls ./ngrams-out
Found 3 items
-rw-r-----   3 junqich hadoop          0 2021-10-03 19:53 ngrams-out/_SUCCESS
-rw-r-----   3 junqich hadoop       5363 2021-10-03 19:53 ngrams-out/part-00000
-rw-r-----   3 junqich hadoop       5304 2021-10-03 19:53 ngrams-out/part-00001
```

The top two files are `_SUCCESS` and `part-00000` as shown above.

- **Question 5**

```
[junqich@cavium-thunderx-login01 si618CaviumSetup]$ grep 1810 ngrams-output.txt
(1810, 5.388058732511713)
[junqich@cavium-thunderx-login01 si618CaviumSetup]$ grep 1846 ngrams-output.txt
(1846, 5.469332533823033)
[junqich@cavium-thunderx-login01 si618CaviumSetup]$ grep 2002 ngrams-output.txt
(2002, 4.959304448361536)
(1965, 4.672820081235415)
```

The average word lengths of words starting with x observed in books from the years 1810, 1946, and 2002 are **5.388058732511713, 4.911973636754396, 4.959304448361536.**

Useful Commands and results:

```
# upload
scp -r D:\CScode\UMCode\si618\week5\SI618_Lab5\si618CaviumSetup junqich@cavium-thunderx.arc-ts.umich.edu:/home/junqich
```

```
./spark-run.sh ngram-job.py /var/umsi618f21/lab5/ngrams/data/googlebooks-eng-all-1gram-20120701-x ./ngrams-out
```

```
(serviceOption=None,
 services=List(),
 started=false)
21/10/03 19:53:48 INFO cluster.YarnClientSchedulerBackend: Stopped
21/10/03 19:53:48 INFO spark.MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
21/10/03 19:53:48 INFO memory.MemoryStore: MemoryStore cleared
21/10/03 19:53:48 INFO storage.BlockManager: BlockManager stopped
21/10/03 19:53:48 INFO storage.BlockManagerMaster: BlockManagerMaster stopped
21/10/03 19:53:48 INFO scheduler.OutputCommitCoordinator$OutputCommitCoordinatorEndpoint: OutputCommitCoordinator stopped!
21/10/03 19:53:48 INFO spark.SparkContext: Successfully stopped SparkContext
21/10/03 19:53:49 INFO util.ShutdownHookManager: Shutdown hook called
21/10/03 19:53:49 INFO util.ShutdownHookManager: Deleting directory /tmp/spark-1cf0f105-c455-49e5-8ca5-1df31f382a46/pyspark-7d0e4b7a-4b01-46ac-8b6a-374b83b57a14
21/10/03 19:53:49 INFO util.ShutdownHookManager: Deleting directory /tmp/spark-1cf0f105-c455-49e5-8ca5-1df31f382a46
21/10/03 19:53:49 INFO util.ShutdownHookManager: Deleting directory /tmp/spark-9eaff06a-5557-4006-adf9-c9e47131227c
[junqich@cavium-thunderx-login01 si618CaviumSetup]$
```

```
# download
scp uniqname@cavium-thunderx.arc-ts.umich.edu:remotefile localfile
```