1. Harsha Thulasi
2. I spent about 6 hours on this lab.
3. Most of it felt like trial and error, and I mistakenly didn't set the paths correctly a few times which caused my pc to read all the textcorpora files and crashed my PC. I ended up only reading 5 files (austen-emma, austen-persuasion, austen-sense, austin-persuasion, bible-kjv). It took me some time to do the 4-df reducer but I think I figured it out. Not sure if I am supposed to do more than I did for the hive. It ended up being several hive commands. I hope its enough (wasn't clear in the assignment).
   a. For the 2nd question, I couldn't get the terms ordered with alphabets first. It had all the numbers so it was a little confusing as well.

4. Not sure what you meant for code for question 1. I ended up getting it by sqoop and hive commands:
   a. Commands:

```
sqoop import --connect jdbc:mysql://localhost/airline --username training
--password training --table On_Time_On_Time_Performance_2016_1 --columns
"UniqueCarrier, ArrDelayMinutes" --fields-terminated-by '\t' --bindir
/tmp -m 1 --hive-import

sqoop import --connect jdbc:mysql://localhost/airline --username training
--password training --table L_UNIQUE_CARRIERS  --fields-terminated-by
'\t' --bindir /tmp -m 1 --hive-import


hive << EOF
INSERT OVERWRITE LOCAL DIRECTORY '/code/part1' ROW FORMAT DELIMITED
FIELDS TERMINATED BY '\t' STORED AS TEXTFILE select description,
avg(arrdelayminutes) as avgDelay from On_Time_On_Time_Performance_2016_1
JOIN L_UNIQUE_CARRIERS where L_UNIQUE_CARRIERS.Code =
On_Time_On_Time_Performance_2016_1.UniqueCarrier group by description
order by avgdelay desc;
EOF
```

   b. Output

```
Spirit Air Lines
18.634635101127785
JetBlue Airways
17.283021432779012
Virgin America
15.657723265619012
SkyWest Airlines Inc.
15.167911974333025
ExpressJet Airlines Inc.
10.915249208860759
American Airlines Inc.
10.695254069511659
Frontier Airlines Inc.
10.652892561983471
United Air Lines Inc.
9.674841669055748
Delta Air Lines Inc.
9.46074961033664
```

```
Southwest Airlines Co.
7.5740563477574385
Alaska Airlines Inc.
6.978952625570776
Hawaiian Airlines Inc.
4.184370015948963
```

5. I didn't understand what to put in the "tfidf/hadoop-steraming" file. I just put a generic script that I again used "my_script" to call it with appropriate arguments.
    a. Ouptut;

```
root@2a2a81af2611:/code/tfidf# hdfs dfs -cat /data-output/5-tfidf/* | head
2021-04-24 06:59:00,770 INFO sasl.SaslDataTransferClient: SASL encryption trust
check: localHostTrusted = false, remoteHostTrusted = false
austen-emma 10000   12.647741429574214
austen-emma 1816    3.1619353573935536
austen-emma 23rd    6.323870714787107
austen-emma 24th    6.323870714787107
austen-emma 26th    6.323870714787107
austen-emma 28th    6.323870714787107
austen-emma 28thand 6.323870714787107
austen-emma 7th     6.323870714787107
austen-emma 8th     6.323870714787107
austen-emma a       3886.650941308156
```