Track: B20-AI-01	**Week 4**
Name: Artem Chernitsa	**Introduction to Big Data**
Email: a.chernitsa@innopolis.university	**12.04.2023**

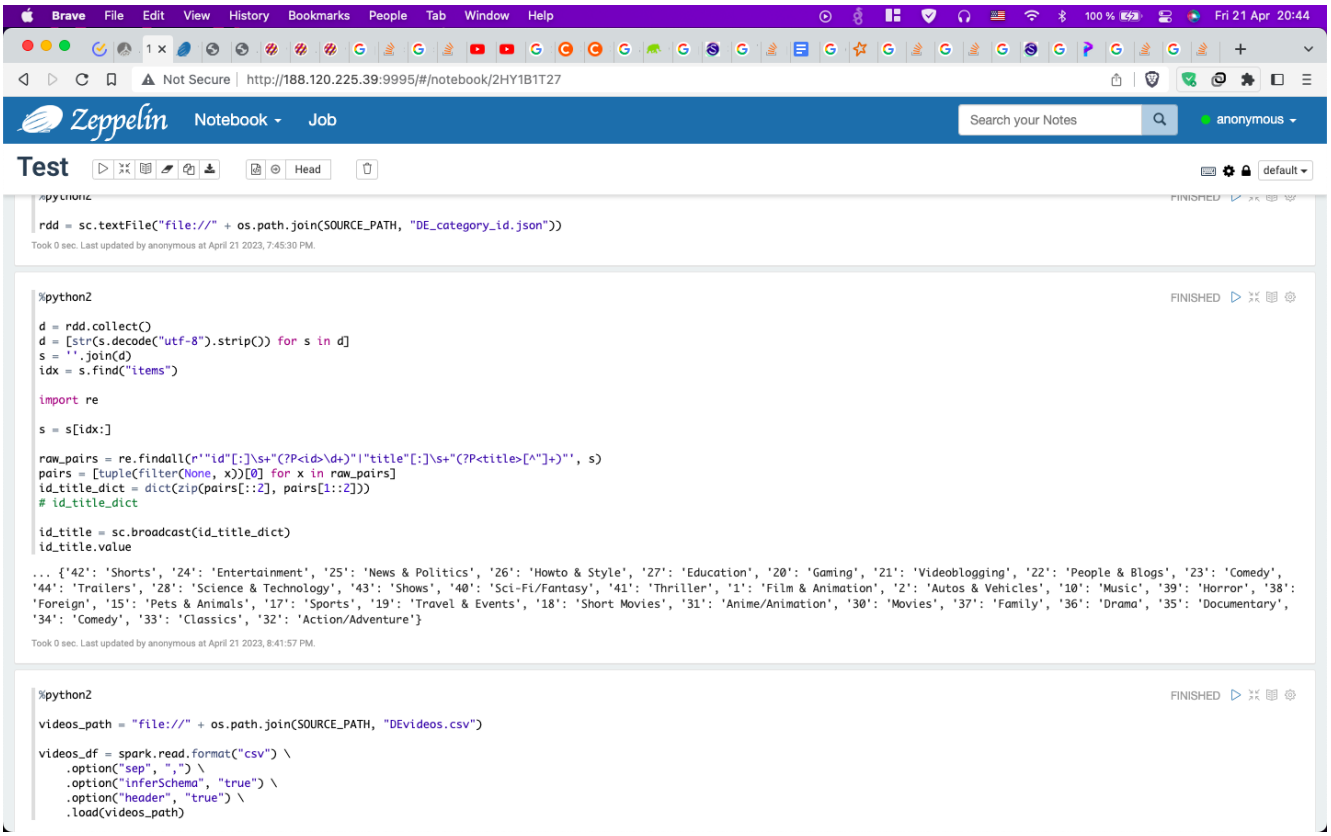
Report

<Spark RDD part>

Task 1

Your query as a text

```
import pyspark
from pyspark.sql import SparkSession

SOURCE_PATH = '/home/zeppelin/ass'

spark = SparkSession.builder \
    .master('local[1]') \
    .appName('Ass Zeppelin') \
    .getOrCreate()
sc = spark.sparkContext

rdd = sc.textFile("file://" + os.path.join(SOURCE_PATH, "DE_category_id.json"))
d = rdd.collect()
d = [str(s.decode("utf-8").strip()) for s in d]
s = ''.join(d)
idx = s.find("items")

import re

s = s[idx:]

raw_pairs = re.findall(r'"id"[:]\s+"(?P<id>\d+)"|"title"[:]\s+"(?P<title>[^"]+)"', s)
pairs = [tuple(filter(None, x))[0] for x in raw_pairs]
id_title_dict = dict(zip(pairs[::2], pairs[1::2]))
# id_title_dict

id_title = sc.broadcast(id_title_dict)
id_title.value
```

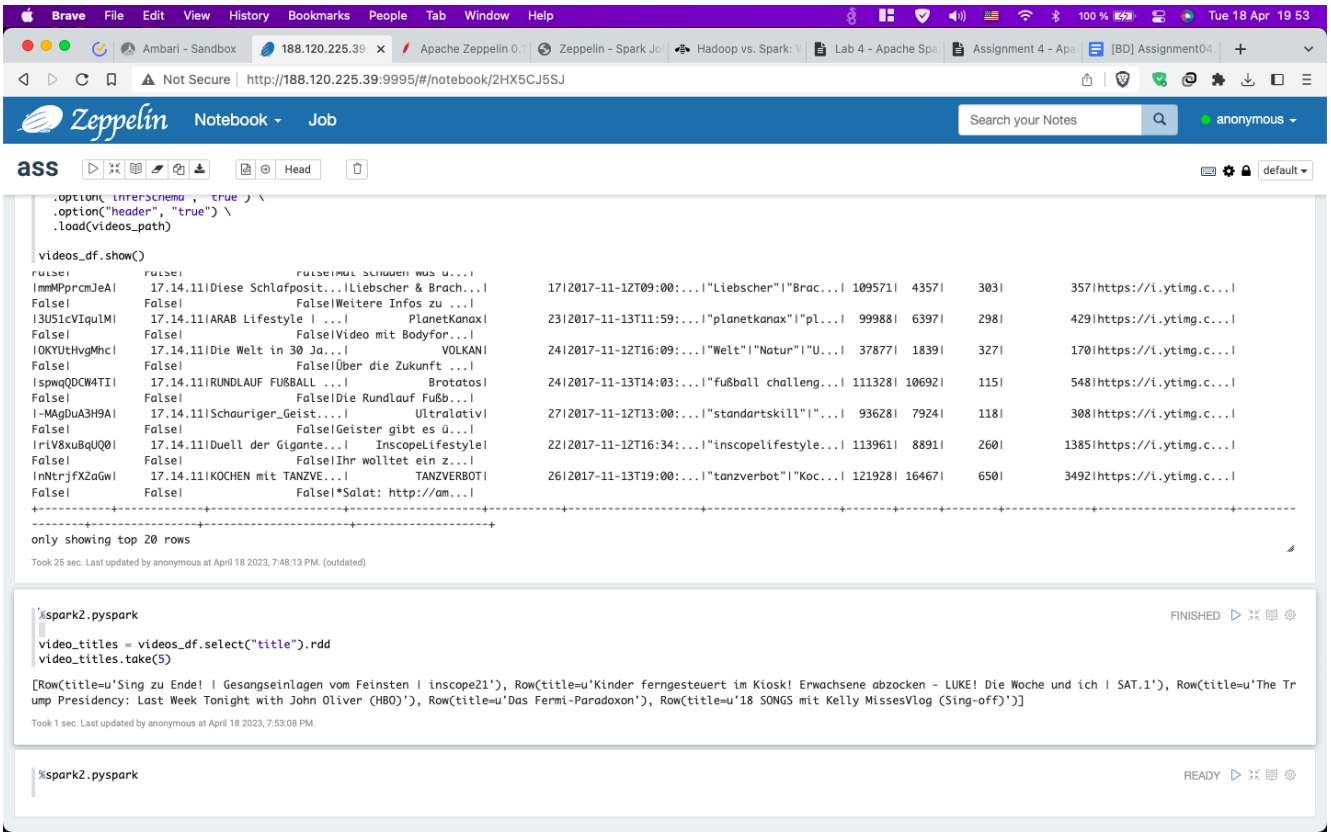


Your comments (optional)

Task 2

Your query as a text

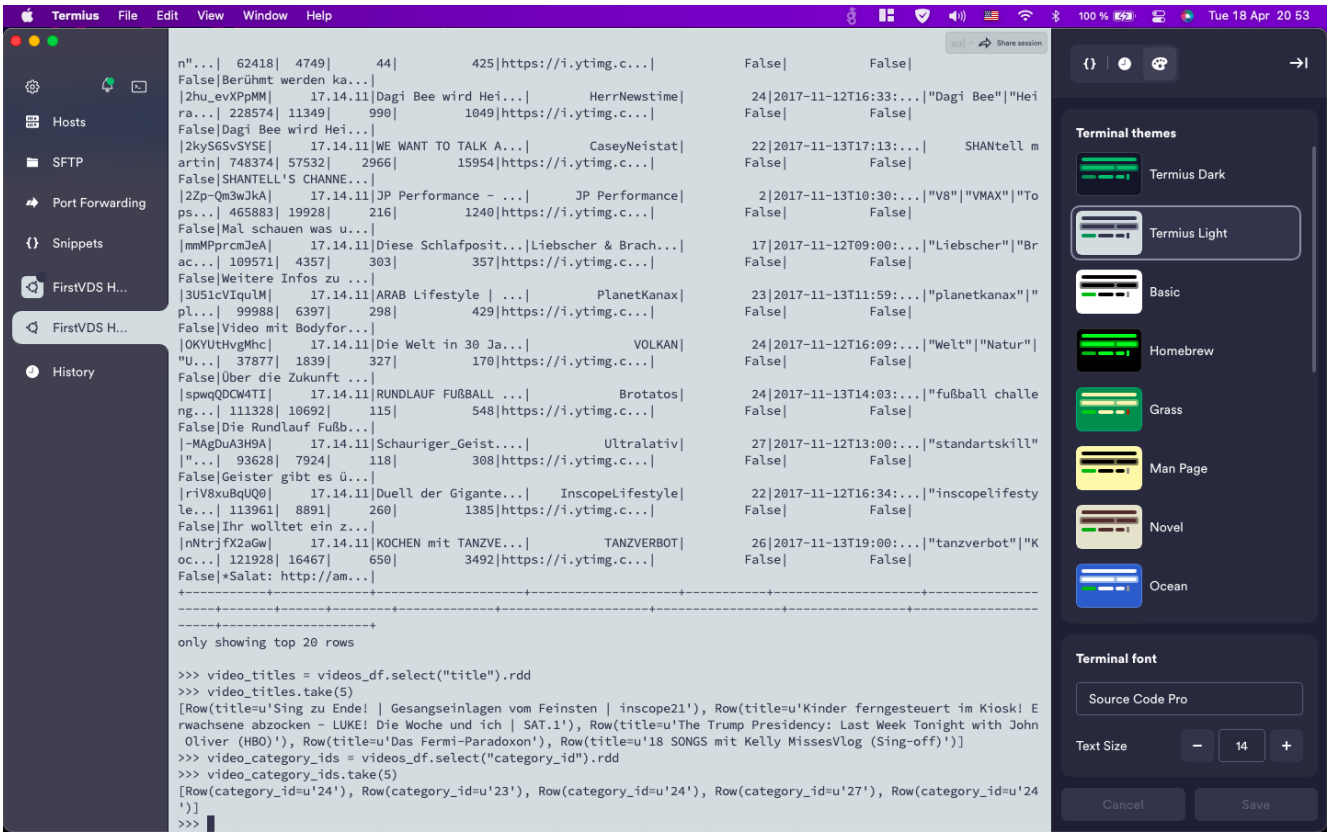
```
videos_path = os.path.join("/ass", "DEvideos.csv")

videos_df = spark.read.format("csv") \
  .option("sep", ",") \
  .option("inferSchema", "true") \
  .option("header", "true") \
  .load(videos_path)

video_titles = videos_df.select("title").rdd
video_titles.take(5)
```



Your comments (optional)
SOURCE_PATH for pyspark instead of /ass

Task 3

Your query as a text
video_category_ids = videos_df.select("category_id").rdd video_category_ids.take(5)

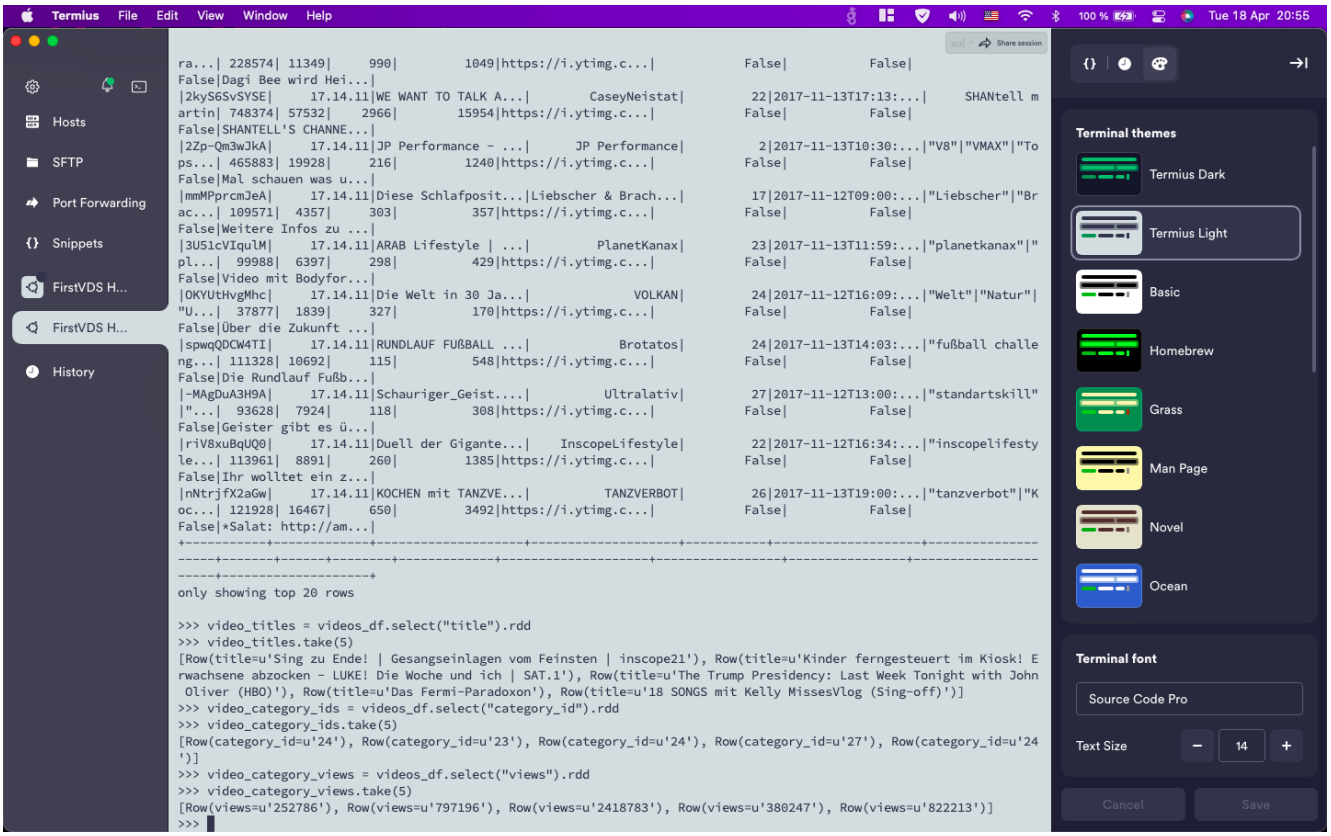



```
n"...|  62418|  4749|      44|          425|https://i.ytimg.c...|           False|           False|
False|Berühmt werden ka...|
|2hu_evXPpMM|     17.14.11|Dagi Bee wird Hei...|        HerrNewstime|         24|2017-11-12T16:33:...|"Dagi Bee"|"Hei
ra...| 228574| 11349|     990|         1049|https://i.ytimg.c...|           False|           False|
False|Dagi Bee wird Hei...|
|2kyS6SvSYSE|     17.14.11|WE WANT TO TALK A...|        CaseyNeistat|         22|2017-11-13T17:13:...|     SHANtell m
artin| 748374| 57532|    2966|        15954|https://i.ytimg.c...|           False|           False|
False|SHANTELL'S CHANNE...|
|2Zp-Qm3wJkA|     17.14.11|JP Performance - ...|      JP Performance|          2|2017-11-13T10:30:...|"V8"|"VMAX"|"To
ps...| 465883| 19928|     216|         1240|https://i.ytimg.c...|           False|           False|
False|Mal schauen was u...|
|mmMPprcmJeA|     17.14.11|Diese Schlafposit...|Liebscher & Brach...|         17|2017-11-12T09:00:...|"Liebscher"|"Br
ac...| 109571|  4357|     303|          357|https://i.ytimg.c...|           False|           False|
False|Weitere Infos zu ...|
|3U51cVIqulM|     17.14.11|ARAB Lifestyle | ...|         PlanetKanax|         23|2017-11-13T11:59:...|"planetkanax"|"
pl...|  99988|  6397|     298|          429|https://i.ytimg.c...|           False|           False|
False|Video mit Bodyfor...|
|OKYUtHvgMhc|     17.14.11|Die Welt in 30 Ja...|              VOLKAN|         24|2017-11-12T16:09:...|"Welt"|"Natur"|
"U...|  37877|  1839|     327|          170|https://i.ytimg.c...|           False|           False|
False|Über die Zukunft ...|
|spwqQDCW4TI|     17.14.11|RUNDLAUF FUßBALL ...|            Brotatos|         24|2017-11-13T14:03:...|"fußball challe
ng...| 111328| 10692|     115|          548|https://i.ytimg.c...|           False|           False|
False|Die Rundlauf Fußb...|
|-MAgDuA3H9A|     17.14.11|Schauriger_Geist....|          Ultralativ|         27|2017-11-12T13:00:...|"standartskill"
|"...|  93628|  7924|     118|          308|https://i.ytimg.c...|           False|           False|
False|Geister gibt es ü...|
|riV8xuBqUQ0|     17.14.11|Duell der Gigante...|    InscopeLifestyle|         22|2017-11-12T16:34:...|"inscopelifesty
le...| 113961|  8891|     260|         1385|https://i.ytimg.c...|           False|           False|
False|Ihr wolltet ein z...|
|nNtrjfX2aGw|     17.14.11|KOCHEN mit TANZVE...|          TANZVERBOT|         26|2017-11-13T19:00:...|"tanzverbot"|"K
oc...| 121928| 16467|     650|         3492|https://i.ytimg.c...|           False|           False|
False|*Salat: http://am...|
+-----------+------------+--------------------+--------------------+-----------+--------------------+----------------
-----+-------+------+--------+-------------+--------------------+----------------+----------------+----------------
-----+--------------------+
only showing top 20 rows

>>> video_titles = videos_df.select("title").rdd
>>> video_titles.take(5)
[Row(title=u'Sing zu Ende! | Gesangseinlagen vom Feinsten | inscope21'), Row(title=u'Kinder ferngesteuert im Kiosk! E
rwachsene abzocken - LUKE! Die Woche und ich | SAT.1'), Row(title=u'The Trump Presidency: Last Week Tonight with John
 Oliver (HBO)'), Row(title=u'Das Fermi-Paradoxon'), Row(title=u'18 SONGS mit Kelly MissesVlog (Sing-off)')]
>>> video_category_ids = videos_df.select("category_id").rdd
>>> video_category_ids.take(5)
[Row(category_id=u'24'), Row(category_id=u'23'), Row(category_id=u'24'), Row(category_id=u'27'), Row(category_id=u'24
')]
>>>
```

Your comments (optional)
Zeppelin has died

Task 4

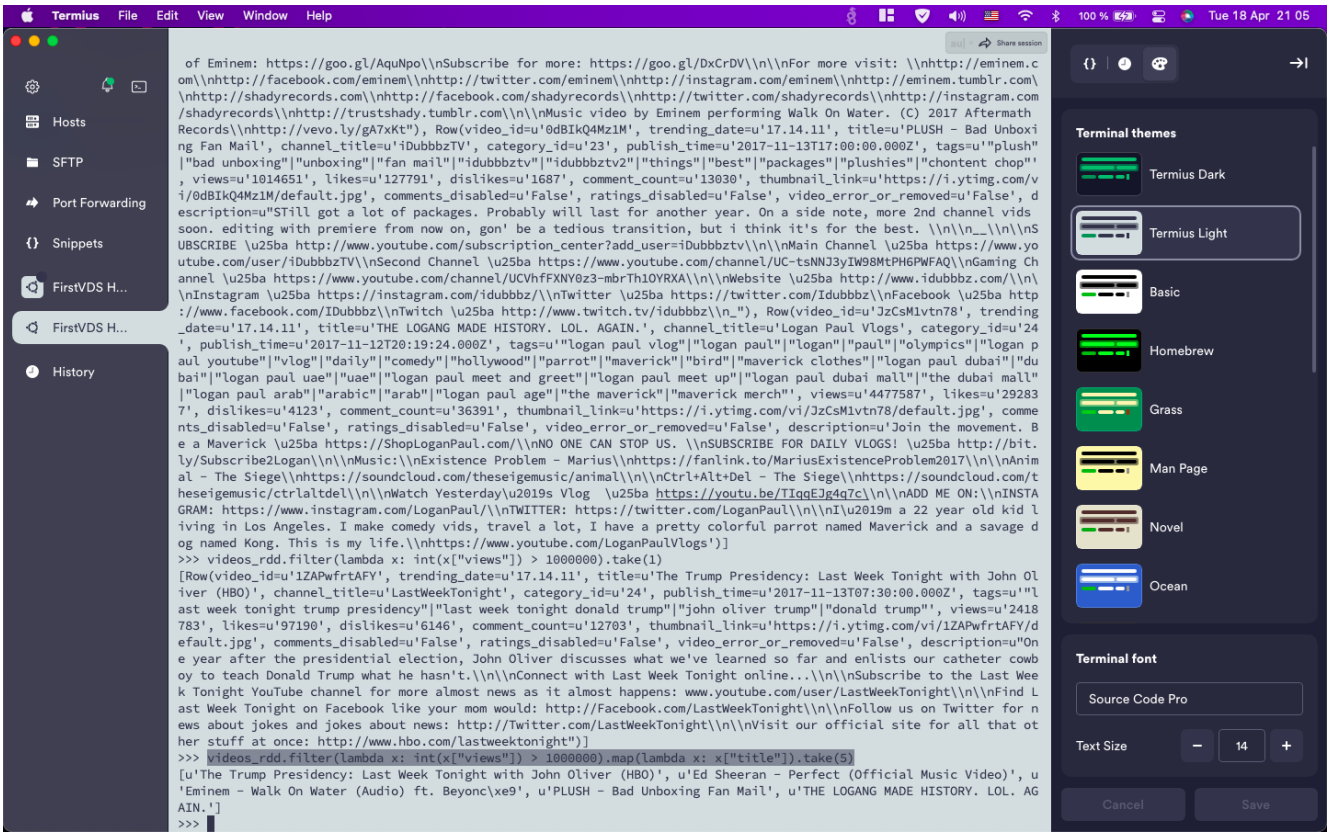
Your query as a text
video_category_views = videos_df.select("views").rdd video_category_views.take(5)



Your comments (optional)

Task 5

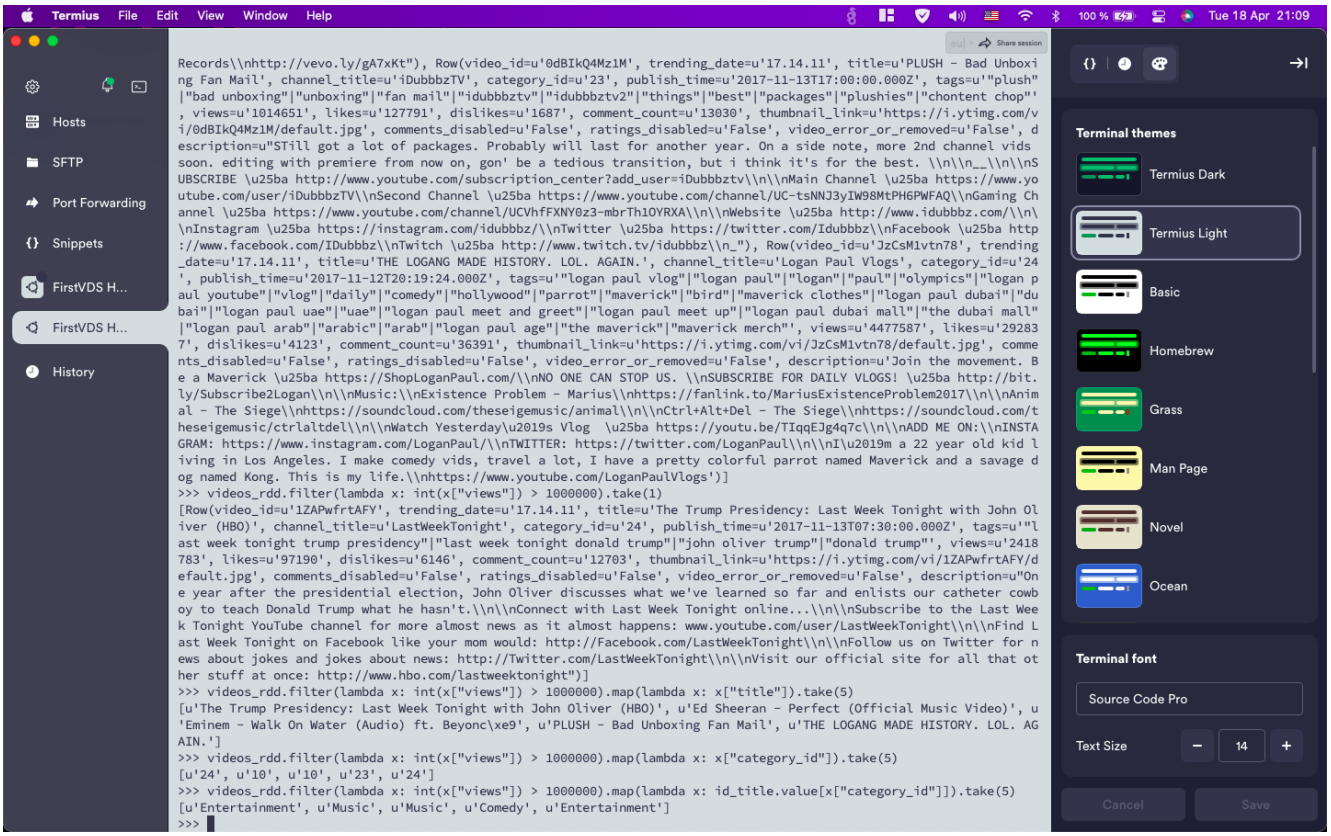
Your query as a text
videos_rdd.filter(lambda x: int(x["views"]) > 1000000).map(lambda x: x["title"]).take(5)



Your comments (optional)
videos_rdd = videos_df.rdd

Task 6

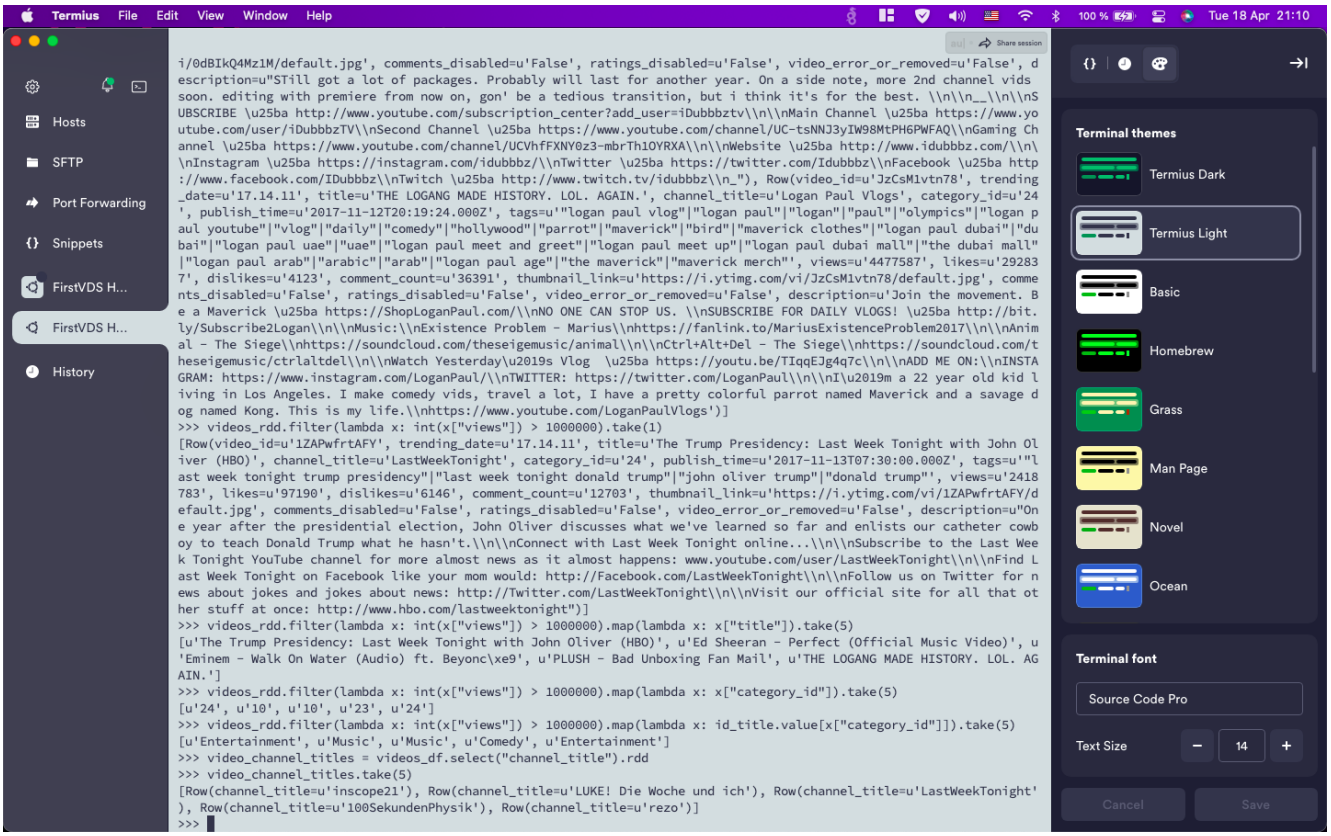
Your query as a text
videos_rdd.filter(lambda x: int(x["views"]) > 1000000).map(lambda x: id_title.value[x["category_id"]]).take(5)



Your comments (optional)

Task 7

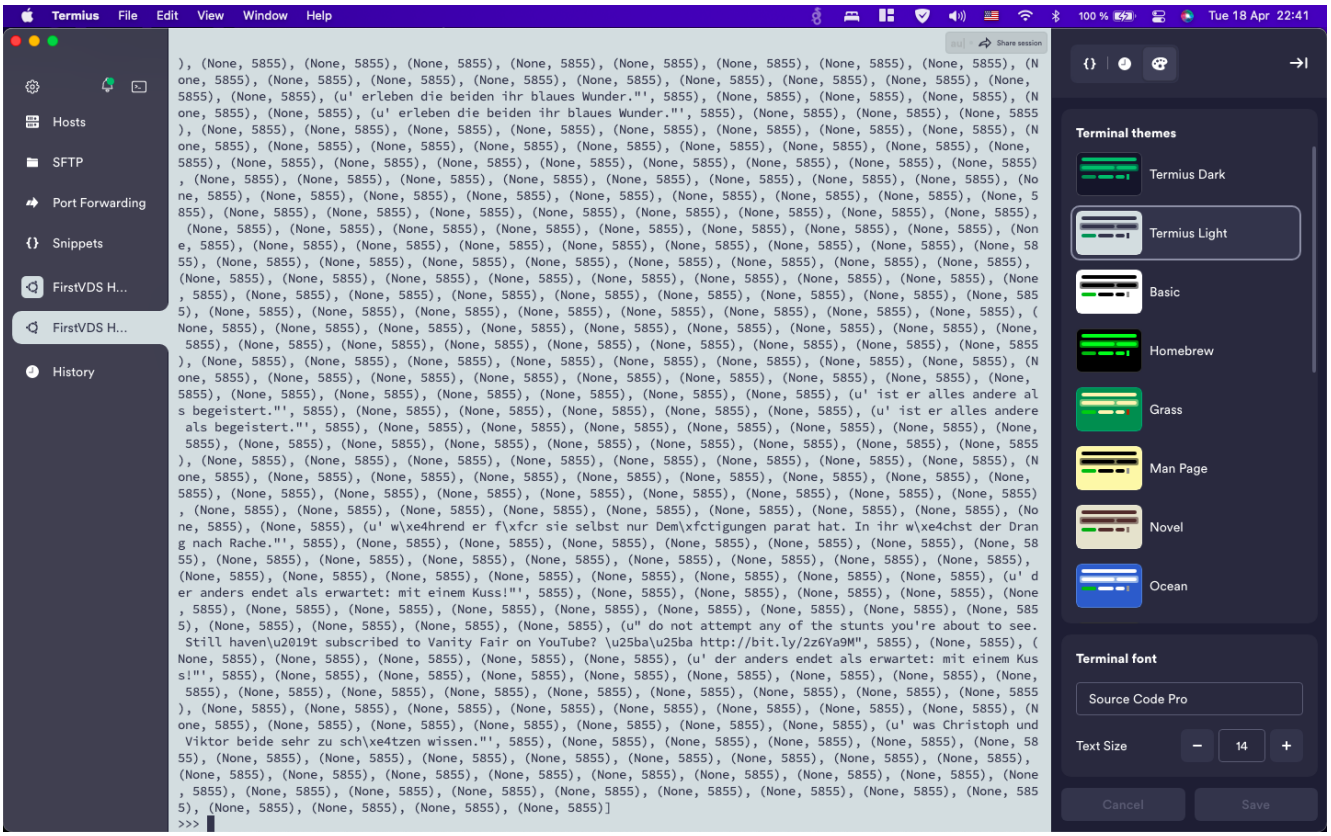
Your query as a text
video_channel_titles = videos_df.select("channel_title").rdd video_channel_titles.take(5)

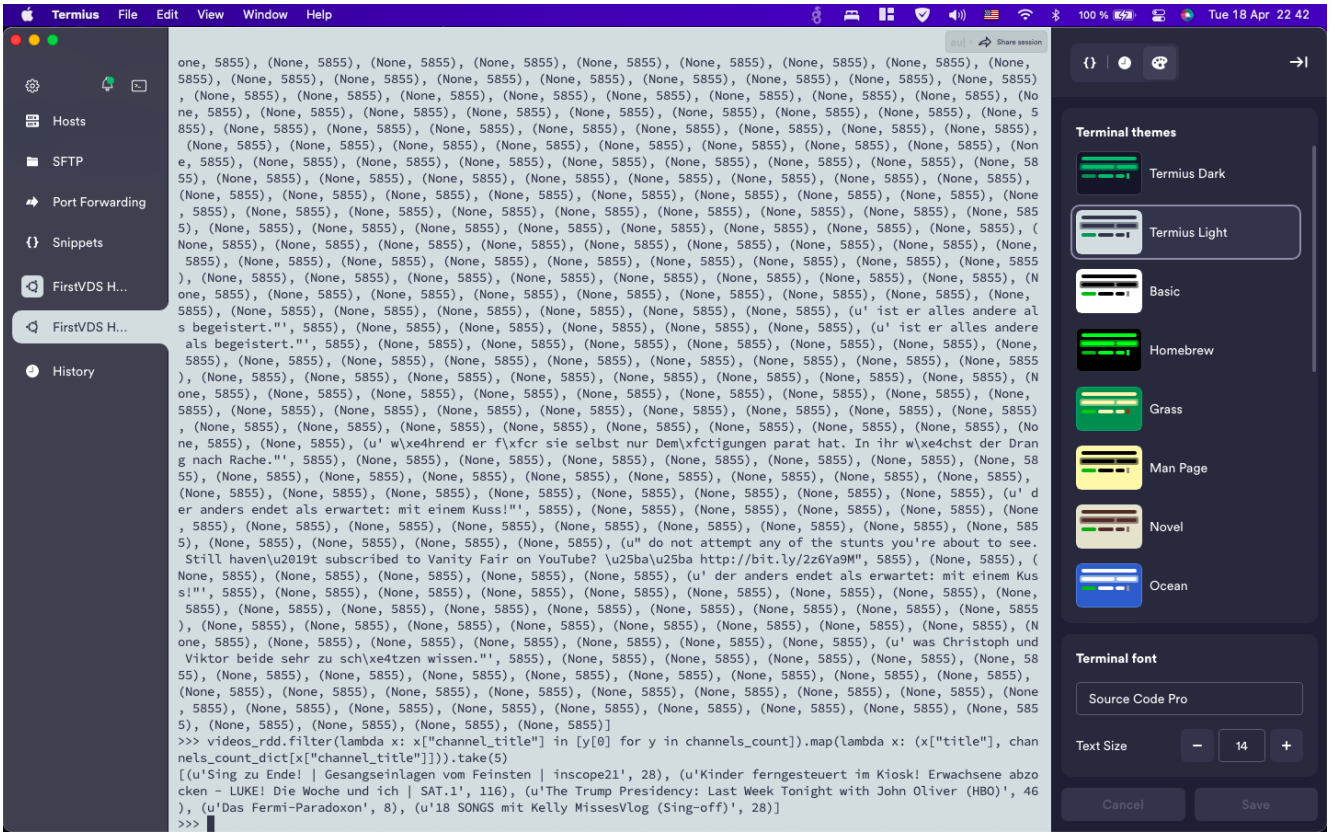


Your comments (optional)

Task 8

Your query as a text
mean = videos_rdd.map(lambda x: (x["channel_title"], 1)).reduceByKey(lambda x, y: x + y).map(lambda x: x[1]).reduce(lambda x, y: x+ y) / videos_rdd.map(lambda x: (x["channel_title"], 1)).reduceByKey(lambda x, y: x + y).count() channels_count = videos_rdd.map(lambda x: (x["channel_title"], 1)).reduceByKey(lambda x, y: x + y).filter(lambda x: x[1] > mean).map(lambda x: (x[0], x[1])).collect() channels_count_dict = dict(channels_count) videos_rdd.filter(lambda x: x["channel_title"] in [y[0] for y in channels_count]).map(lambda x: (x["title"], channels_count_dict[x["channel_title"]])).sortBy(lambda y: y[1], ascending=True).collect()





Your comments (optional)

I'm not sure if this title is just a problem with the dataset. But I've checked there are empty video titles

```
from pyspark.sql.functions import col,isnan,when,count
>>> df2 = df.select([count(when(col(c).contains('None') | \
...                     col(c).contains('NULL') | \
```

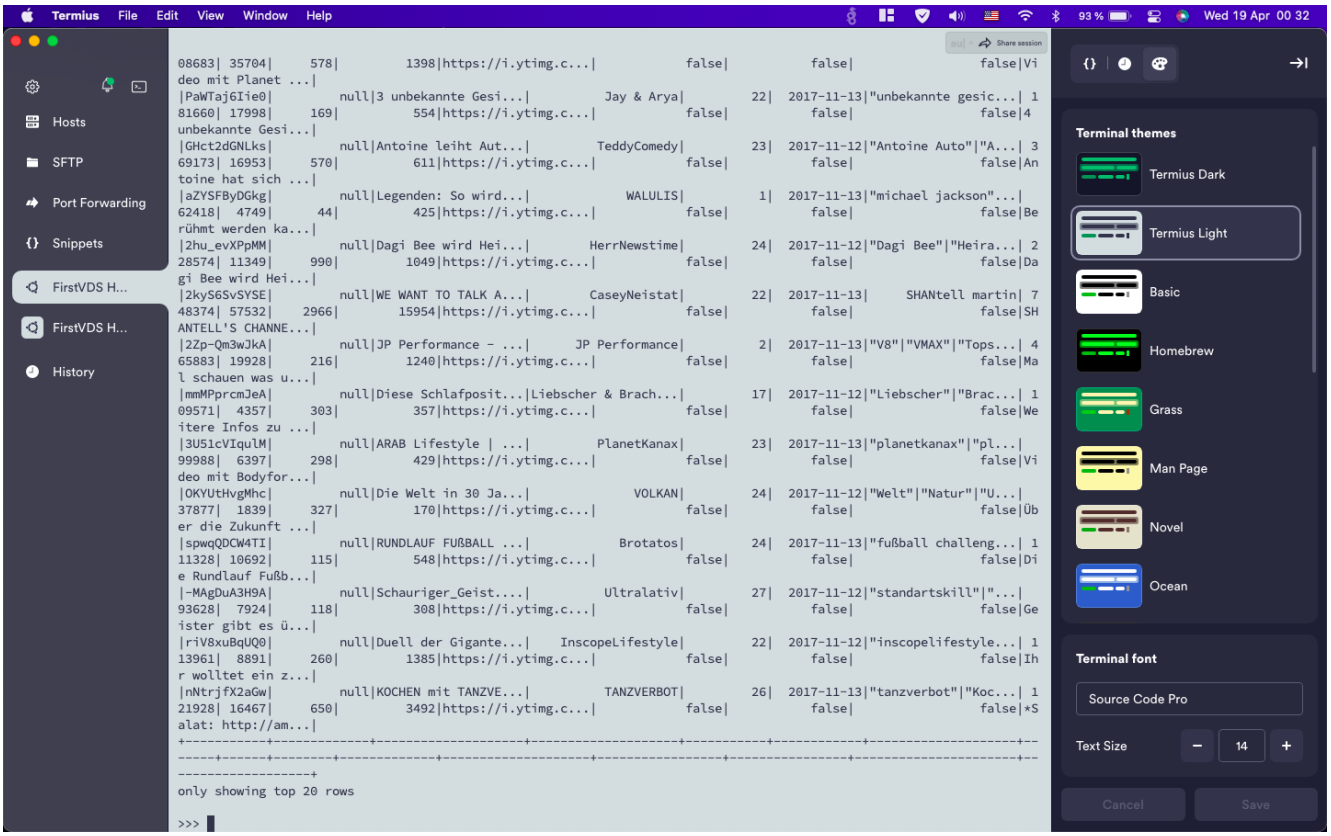
```
...                         (col(c) == '' ) | \
...                         col(c).isNull() | \
...                         isnan(c), c
...                        )).alias(c)
...                    for c in df.columns])
>>> df2.show()
```

<Spark DataFrame part>

Task 1

Your query as a text

```
import pyspark.sql.functions as f
from pyspark.sql.types import *
schema = StructType([ \
    StructField("video_id", StringType(), True), \
    StructField("trending_date", DateType(), True), \
    StructField("title", StringType(), True), \
    StructField("channel_title", StringType(), True), \
    StructField("category_id", IntegerType(), True), \
    StructField("publish_time", DateType(), True), \
    StructField("tags", StringType(), True), \
    StructField("views", IntegerType(), True), \
    StructField("likes", IntegerType(), True), \
    StructField("dislikes", IntegerType(), True), \
    StructField("comment_count", IntegerType(), True), \
    StructField("thumbnail_link", StringType(), True), \
    StructField("comments_disabled", BooleanType(), True), \
    StructField("ratings_disabled", BooleanType(), True), \
    StructField("video_error_or_removed", BooleanType(), True), \
    StructField("description", StringType(), True) \
])
path = "/home/zeppelin/ass/DEvideos.csv"
df = spark.read.format("csv").option("header", True).schema(schema).load(path)
```

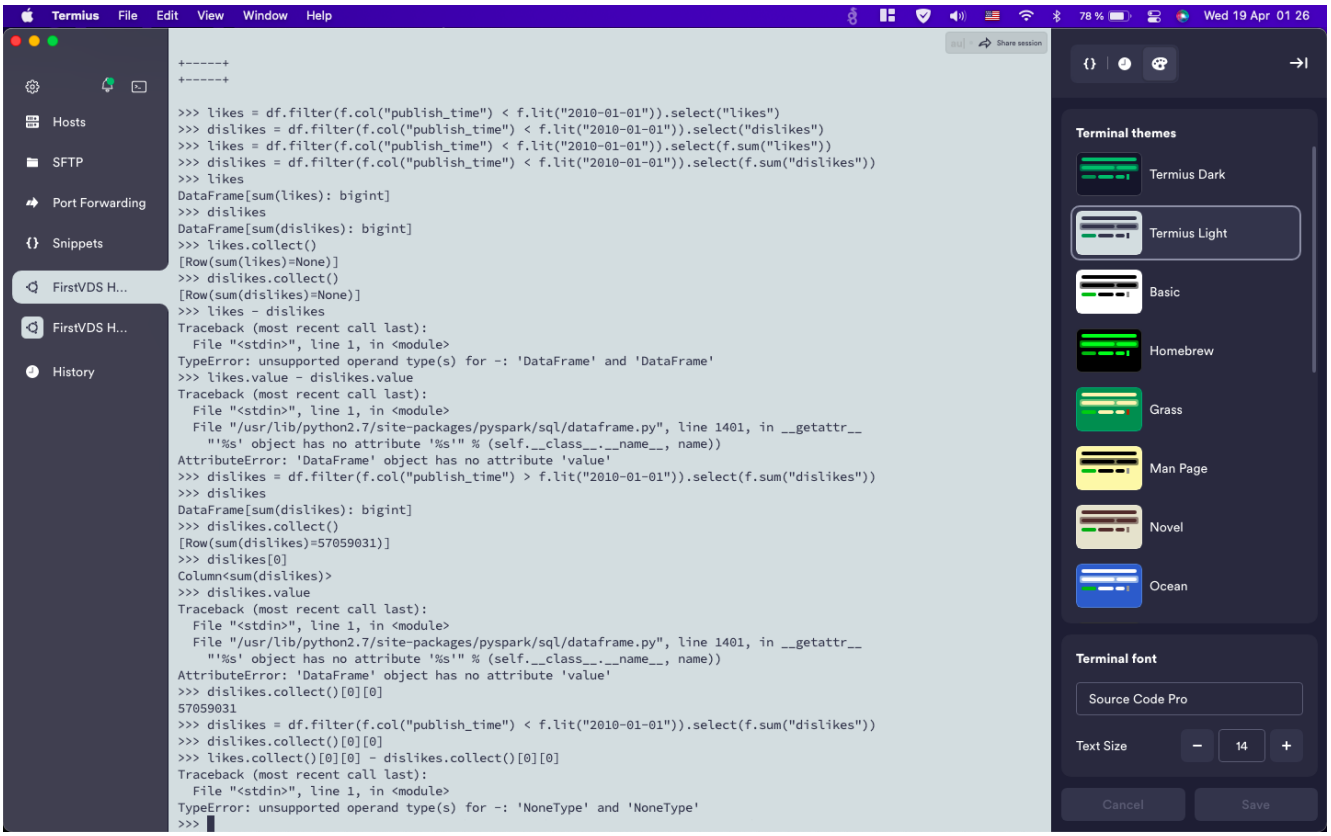



```
08683| 35704|     578|        1398|https://i.ytimg.c...|           false|          false|               false|Vi
deo mit Planet ...|
|PaWTaj6Iie0|        null|3 unbekannte Gesi...|          Jay & Arya|        22|  2017-11-13|"unbekannte gesic...| 1
81660| 17998|     169|         554|https://i.ytimg.c...|           false|          false|               false|4
unbekannte Gesi...|
|GHct2dGNLks|        null|Antoine leiht Aut...|        TeddyComedy|        23|  2017-11-12|"Antoine Auto"|"A...| 3
69173| 16953|     570|         611|https://i.ytimg.c...|           false|          false|               false|An
toine hat sich ...|
|aZYSFByDGkg|        null|Legenden: So wird...|             WALULIS|         1|  2017-11-13|"michael jackson"...|
62418|  4749|      44|         425|https://i.ytimg.c...|           false|          false|               false|Be
rühmt werden ka...|
|2hu_evXPpMM|        null|Dagi Bee wird Hei...|       HerrNewstime|        24|  2017-11-12|"Dagi Bee"|"Heira...| 2
28574| 11349|     990|        1049|https://i.ytimg.c...|           false|          false|               false|Da
gi Bee wird Hei...|
|2kyS6SvSYSE|        null|WE WANT TO TALK A...|       CaseyNeistat|        22|  2017-11-13|     SHANtell martin| 7
48374| 57532|    2966|       15954|https://i.ytimg.c...|           false|          false|               false|SH
ANTELL'S CHANNE...|
|2Zp-Qm3wJkA|        null|JP Performance - ...|     JP Performance|         2|  2017-11-13|"V8"|"VMAX"|"Tops...| 4
65883| 19928|     216|        1240|https://i.ytimg.c...|           false|          false|               false|Ma
l schauen was u...|
|mmMPprcmJeA|        null|Diese Schlafposit...|Liebscher & Brach...|        17|  2017-11-12|"Liebscher"|"Brac...| 1
09571|  4357|     303|         357|https://i.ytimg.c...|           false|          false|               false|We
itere Infos zu ...|
|3U51cVIqulM|        null|ARAB Lifestyle | ...|         PlanetKanax|        23|  2017-11-13|"planetkanax"|"pl...|
99988|  6397|     298|         429|https://i.ytimg.c...|           false|          false|               false|Vi
deo mit Bodyfor...|
|OKYUtHvgMhc|        null|Die Welt in 30 Ja...|              VOLKAN|        24|  2017-11-12|"Welt"|"Natur"|"U...|
37877|  1839|     327|         170|https://i.ytimg.c...|           false|          false|               false|Üb
er die Zukunft ...|
|spwqQDCW4TI|        null|RUNDLAUF FUßBALL ...|            Brotatos|        24|  2017-11-13|"fußball challeng...| 1
11328| 10692|     115|         548|https://i.ytimg.c...|           false|          false|               false|Di
e Rundlauf Fußb...|
|-MAgDuA3H9A|        null|Schauriger_Geist....|          Ultralativ|        27|  2017-11-12|"standartskill"|"...|
93628|  7924|     118|         308|https://i.ytimg.c...|           false|          false|               false|Ge
ister gibt es ü...|
|riV8xuBqUQ0|        null|Duell der Gigante...|    InscopeLifestyle|        22|  2017-11-12|"inscopelifestyle...| 1
13961|  8891|     260|        1385|https://i.ytimg.c...|           false|          false|               false|Ih
r wolltet ein z...|
|nNtrjfX2aGw|        null|KOCHEN mit TANZVE...|          TANZVERBOT|        26|  2017-11-13|"tanzverbot"|"Koc...| 1
21928| 16467|     650|        3492|https://i.ytimg.c...|           false|          false|               false|*S
alat: http://am...|
+-----------+-------------+--------------------+--------------------+----------+-----------+-----------+--------------------+--
-----+------+--------+-------------+--------------------+----------+---------------+---------------+--------------------+--
------------------+
only showing top 20 rows

>>>
```

Your comments (optional)
video_id,trending_date,title,channel_title,category_id,publish_time,tags,views,likes,dislikes,comment_count,thumbnail_link,comments_disabled,ratings_disabled,video_error_or_removed,description It can't read array type, so: df2 = df.withColumn("tags", f.split(df.tags, "[|]").alias('tags'))

Task 2

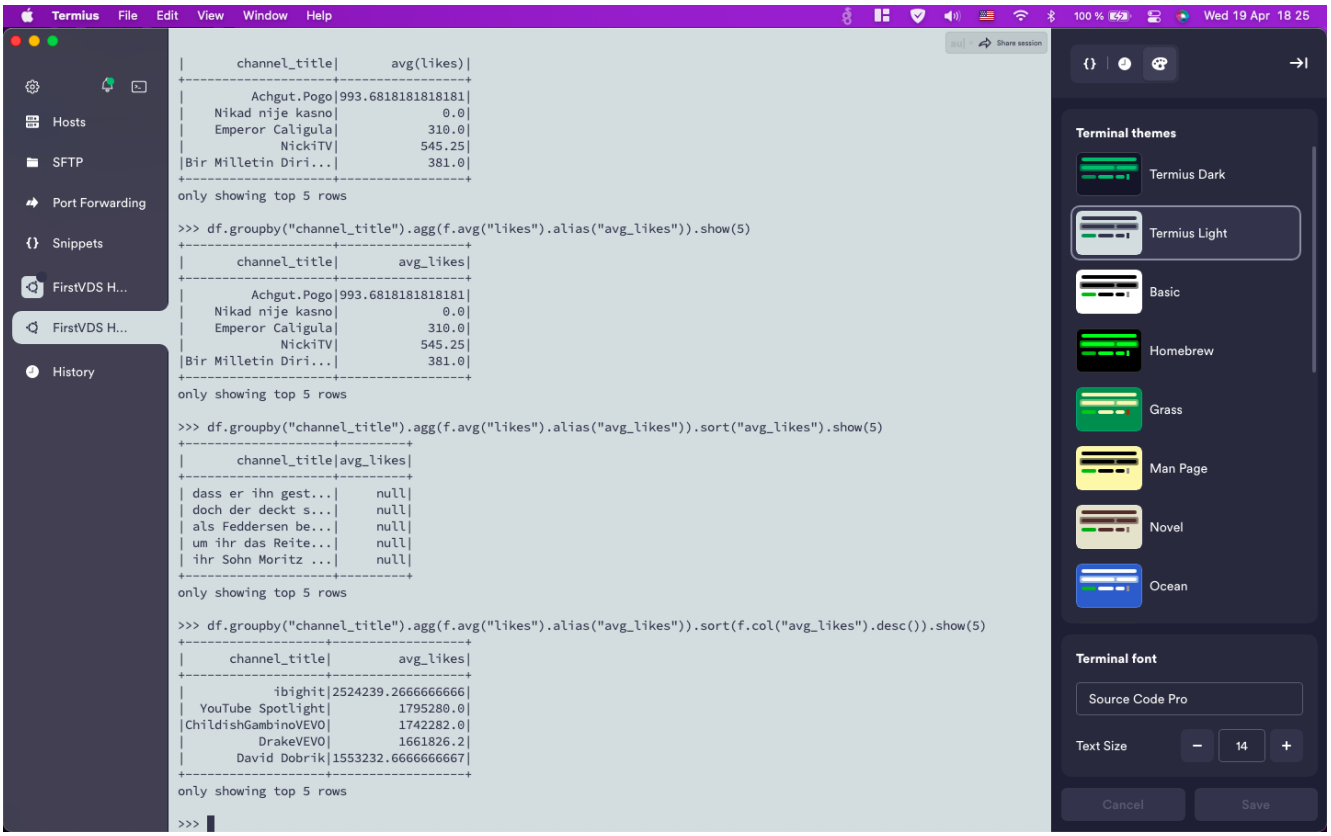
Your query as a text
likes = df.filter(f.col("publish_time") < f.lit("2010-01-01")).select(f.sum("likes")) dislikes = df.filter(f.col("publish_time") < f.lit("2010-01-01")).select(f.sum("dislikes")) likes.collect()[0][0] - dislikes.collect()[0][0]



Your comments (optional)
df.filter(f.col("publish_time").isNotNull()).sort(f.col("publish_time")).show()

Task 3

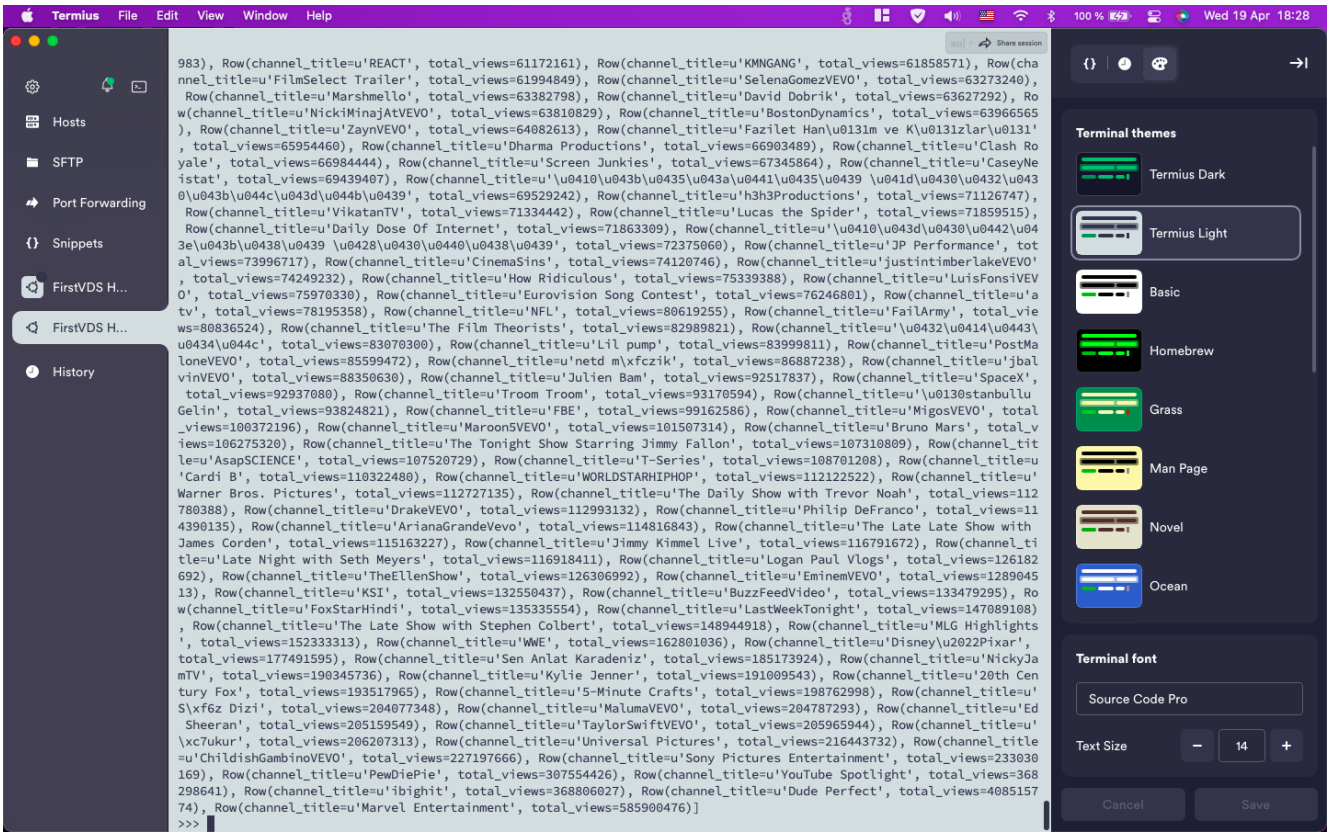
Your query as a text
df.groupby("channel_title").agg(f.avg("likes").alias("avg_likes")).sort(f.col("avg_likes").desc()).show(5)



Your comments (optional)

Task 4

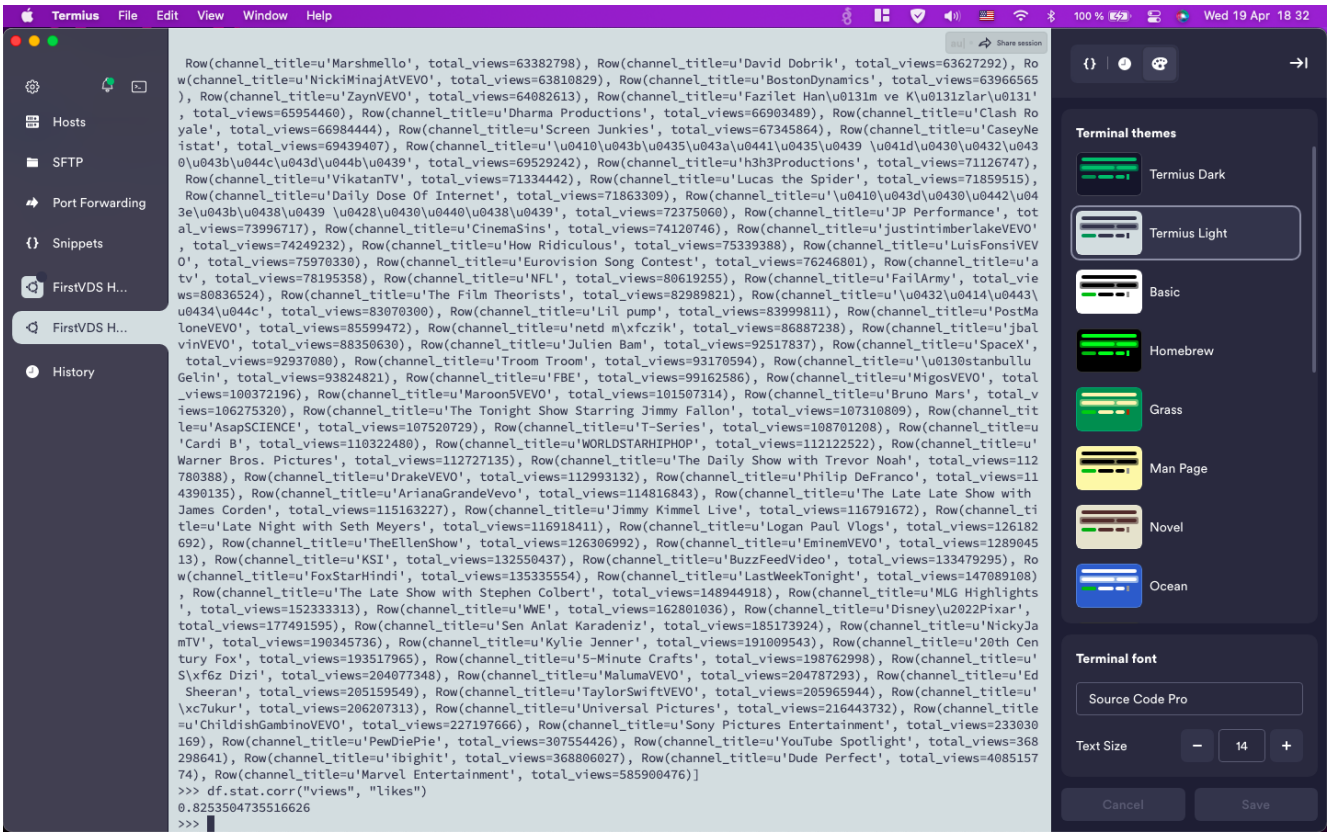
Your query as a text
df.groupby("channel_title").agg(f.sum("views").alias("total_views")).where(f.col("total_views") > 1000000).sort(f.col("total_views")).collect()



Your comments (optional)

Task 5

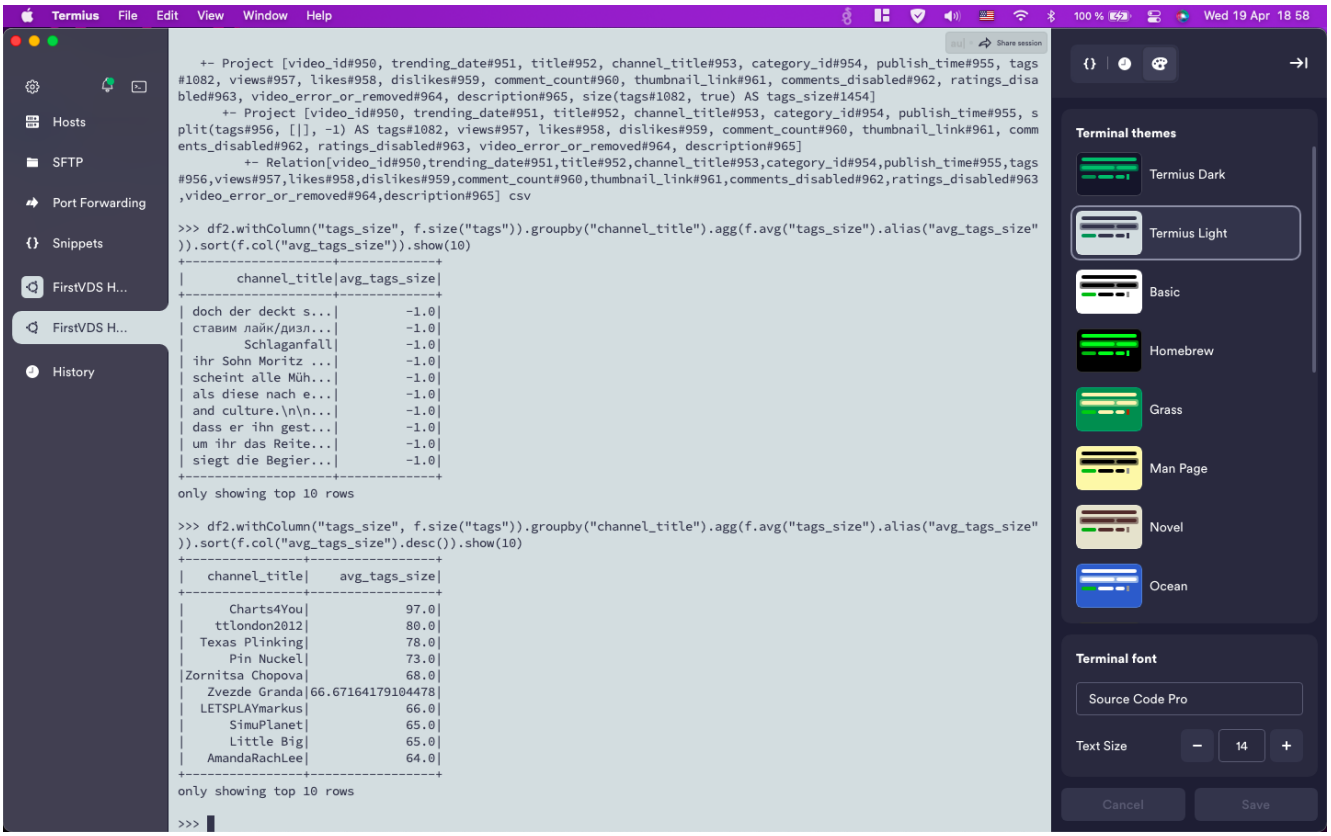
Your query as a text
df.stat.corr("views", "likes")



Your comments (optional)
It lies between 0.5 and 1.0, hence it has strong correlation

Task 6

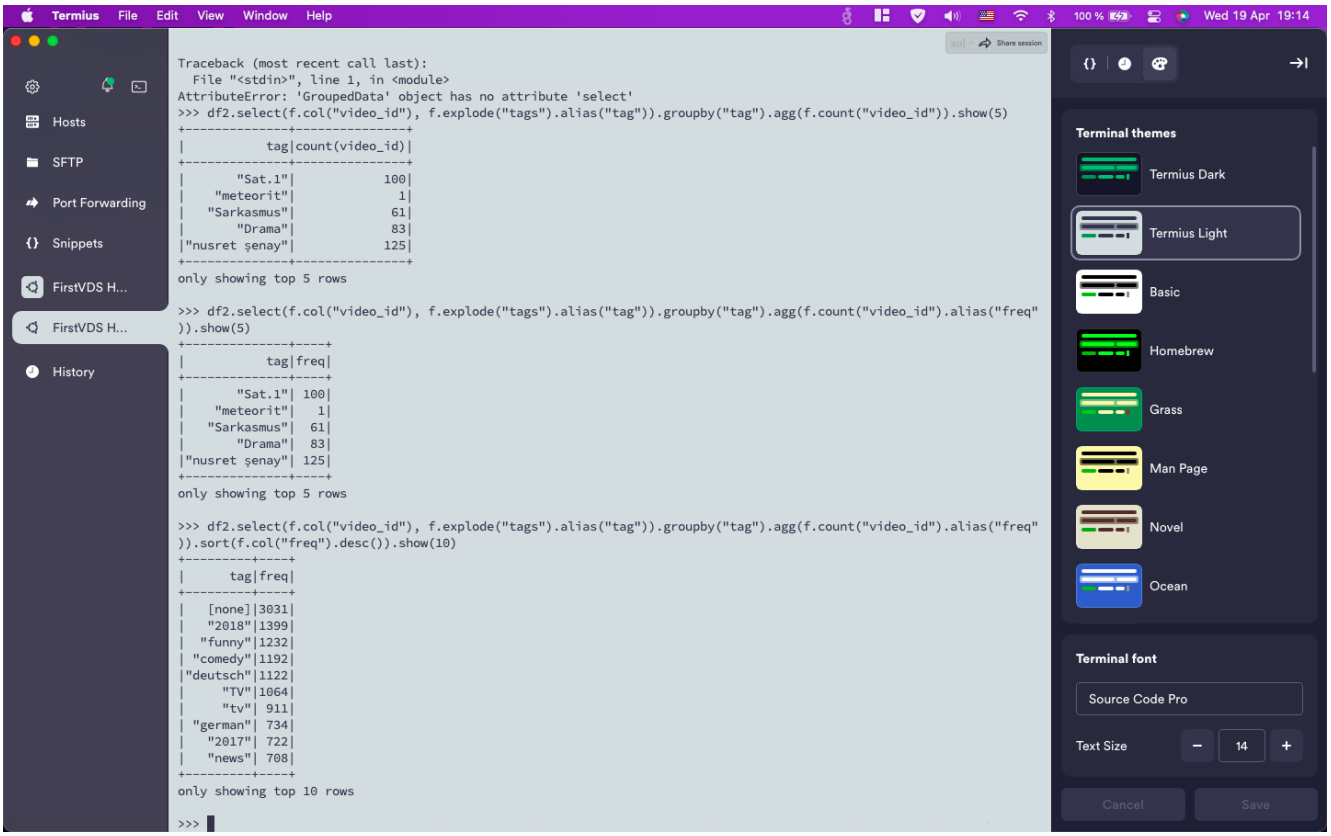
Your query as a text
df2.withColumn("tags_size", f.size("tags")).groupby("channel_title").agg(f.avg("tags_size").alias("avg_tags_size")).sort(f.col("avg_tags_size").desc()).show(10)



Your comments (optional)
df2 = df.withColumn("tags", f.split(df.tags, "[|]").alias('tags'))

Task 7

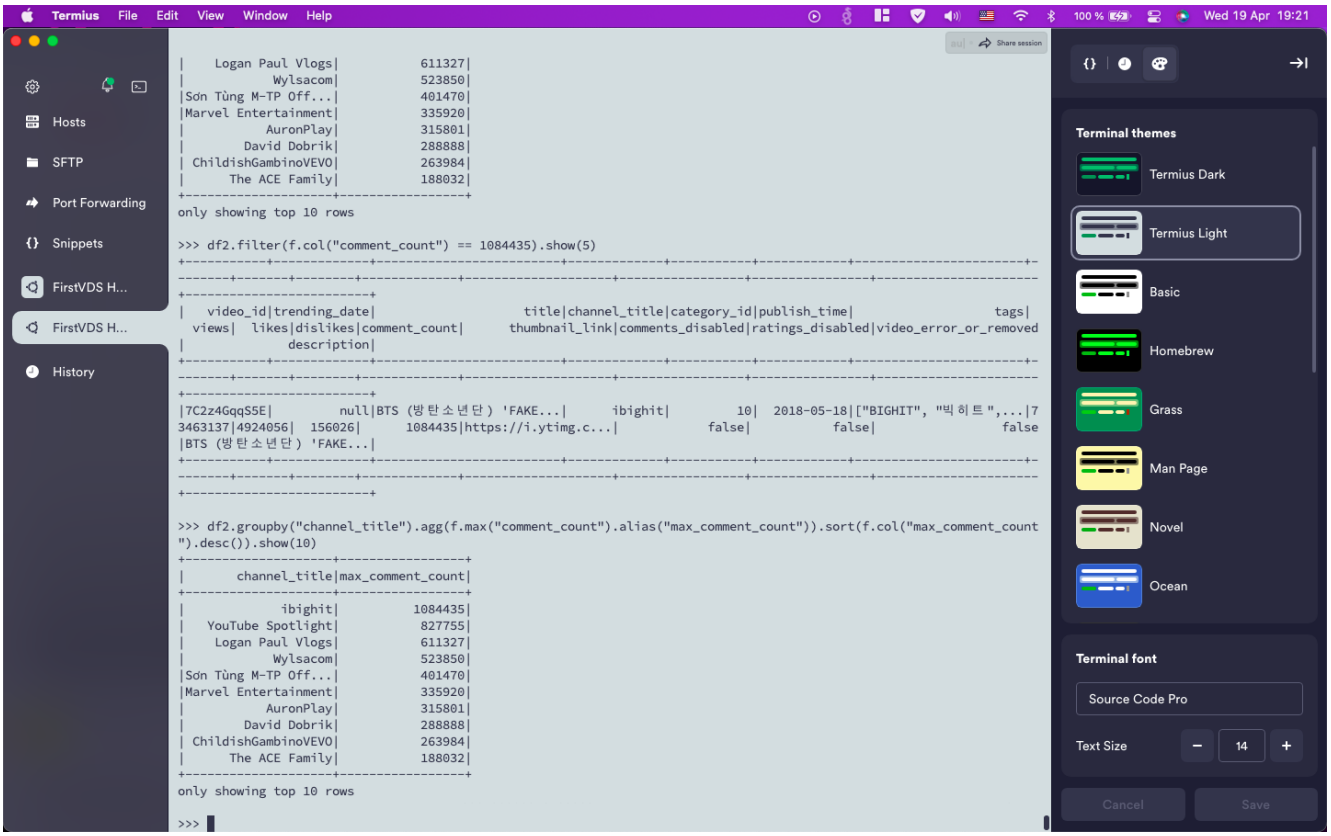
Your query as a text
df2.select(f.col("video_id"), f.explode("tags").alias("tag")).groupby("tag").agg(f.count("video_id").alias("freq")).sort(f.col("freq").desc()).show(10)



Your comments (optional)

Task 8

Your query as a text
df2.groupby("channel_title").agg(f.max("comment_count").alias("max_comment_count")).sort(f.col("max_comment_count").desc()).show(10)

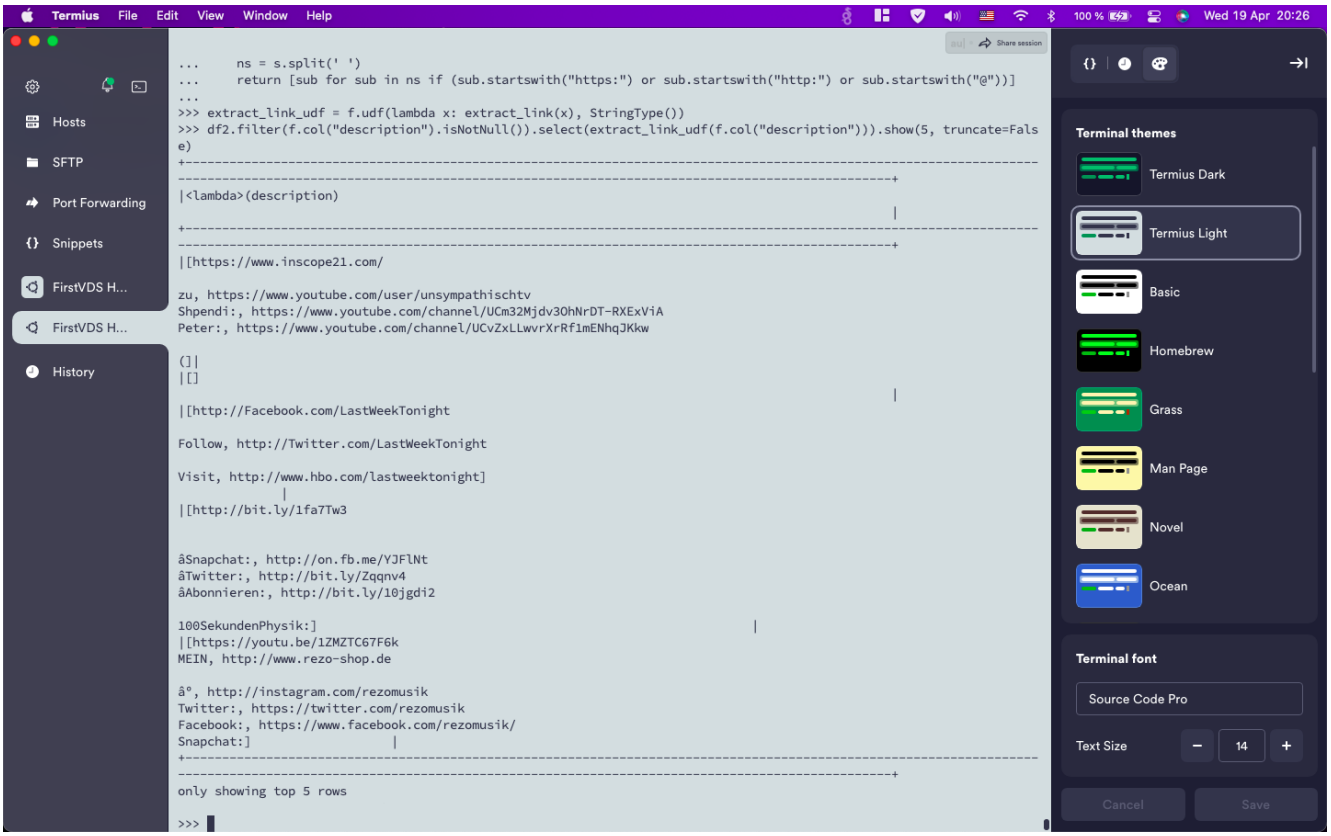


Your comments (optional)

Task 9

Your query as a text

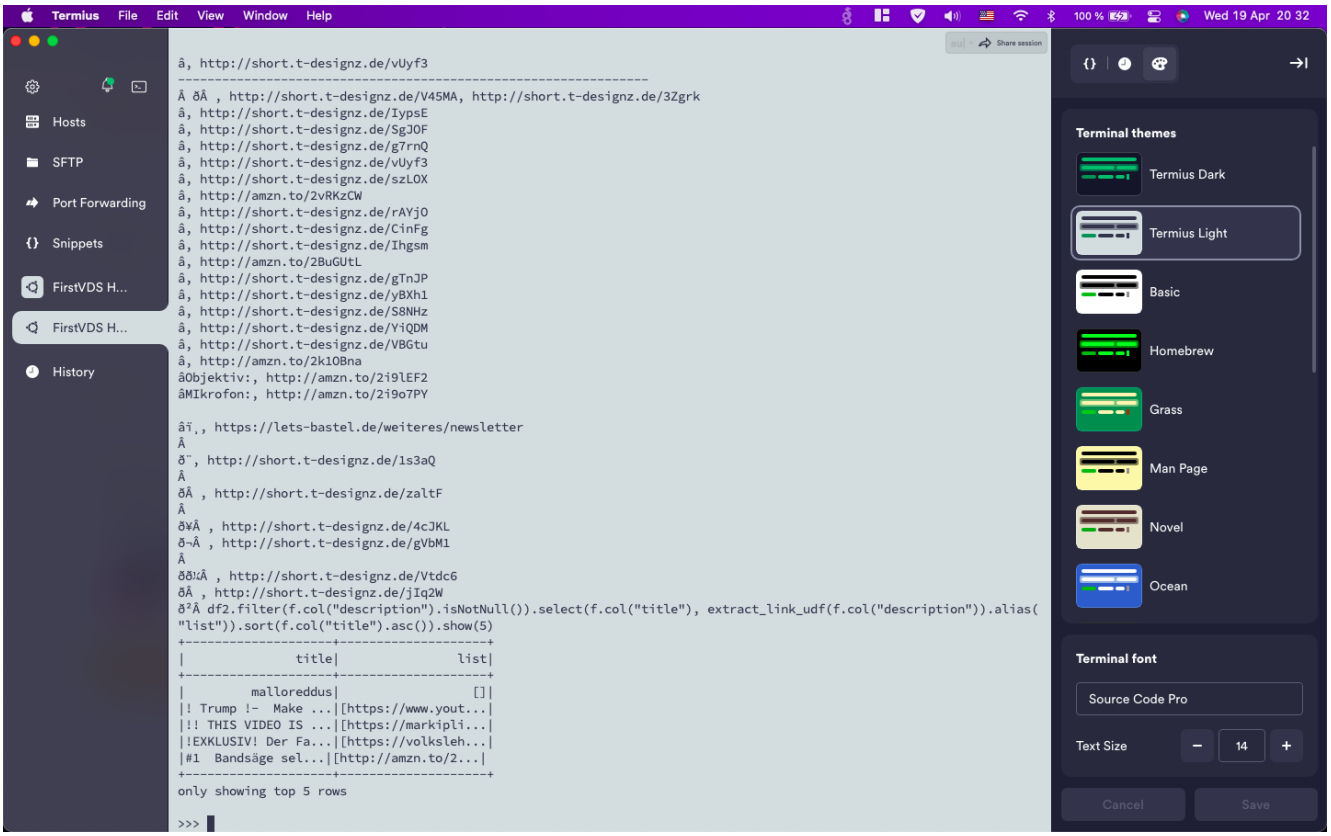
def extract_link(s):
  s = s.encode('utf-8').decode('unicode-escape')
  ns = s.split(' ')
  return [sub for sub in ns if (sub.startswith("https:") or sub.startswith("http:") or sub.startswith("@"))]
extract_link_udf = f.udf(lambda x: extract_link(x), StringType())
df2.filter(f.col("description").isNotNull()).select(extract_link_udf(f.col("description"))).show(5, truncate=False)

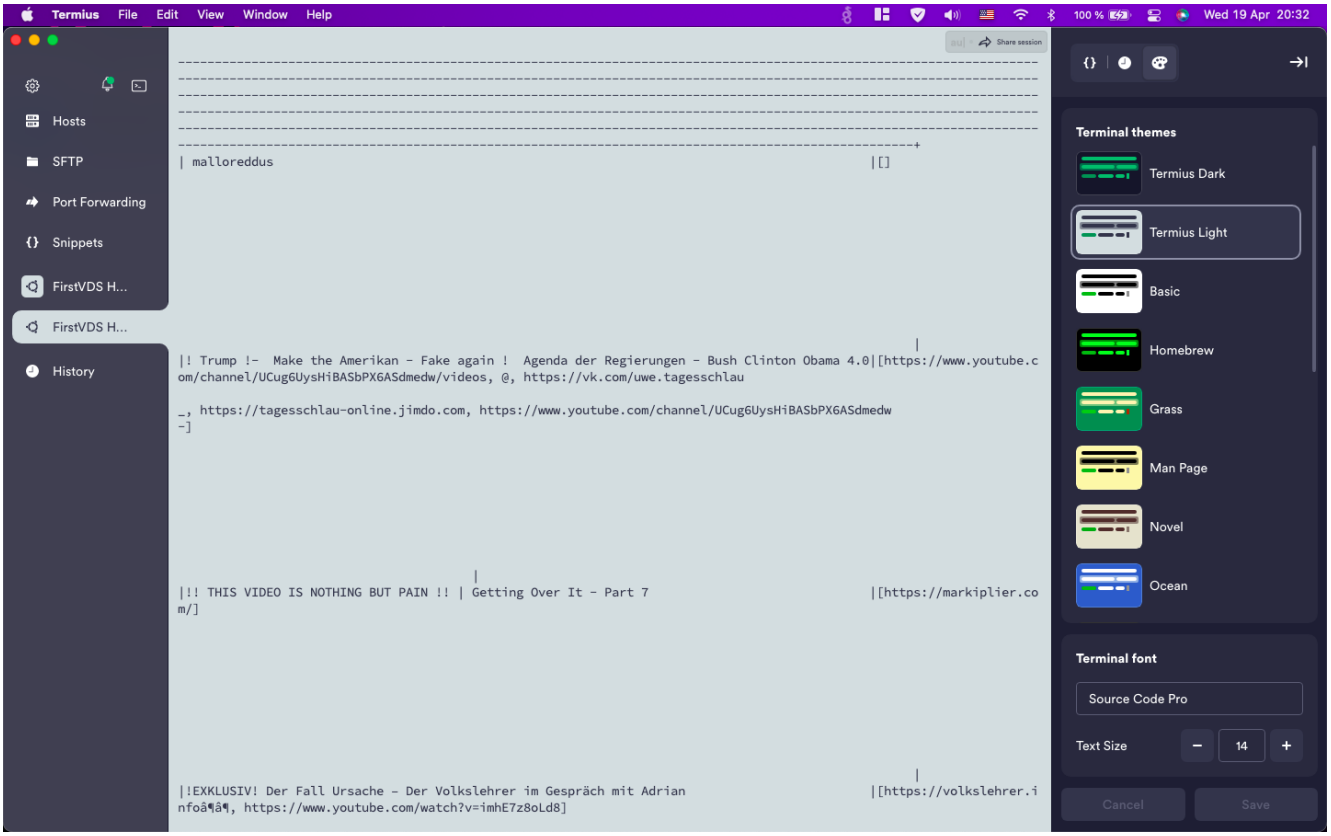


Your comments (optional)

Task 10

Your query as a text
df2.filter(f.col("description").isNotNull()).select(f.col("title"), extract_link_udf(f.col("description")).alias("list")).sort(f.col("title").asc()).show(5)



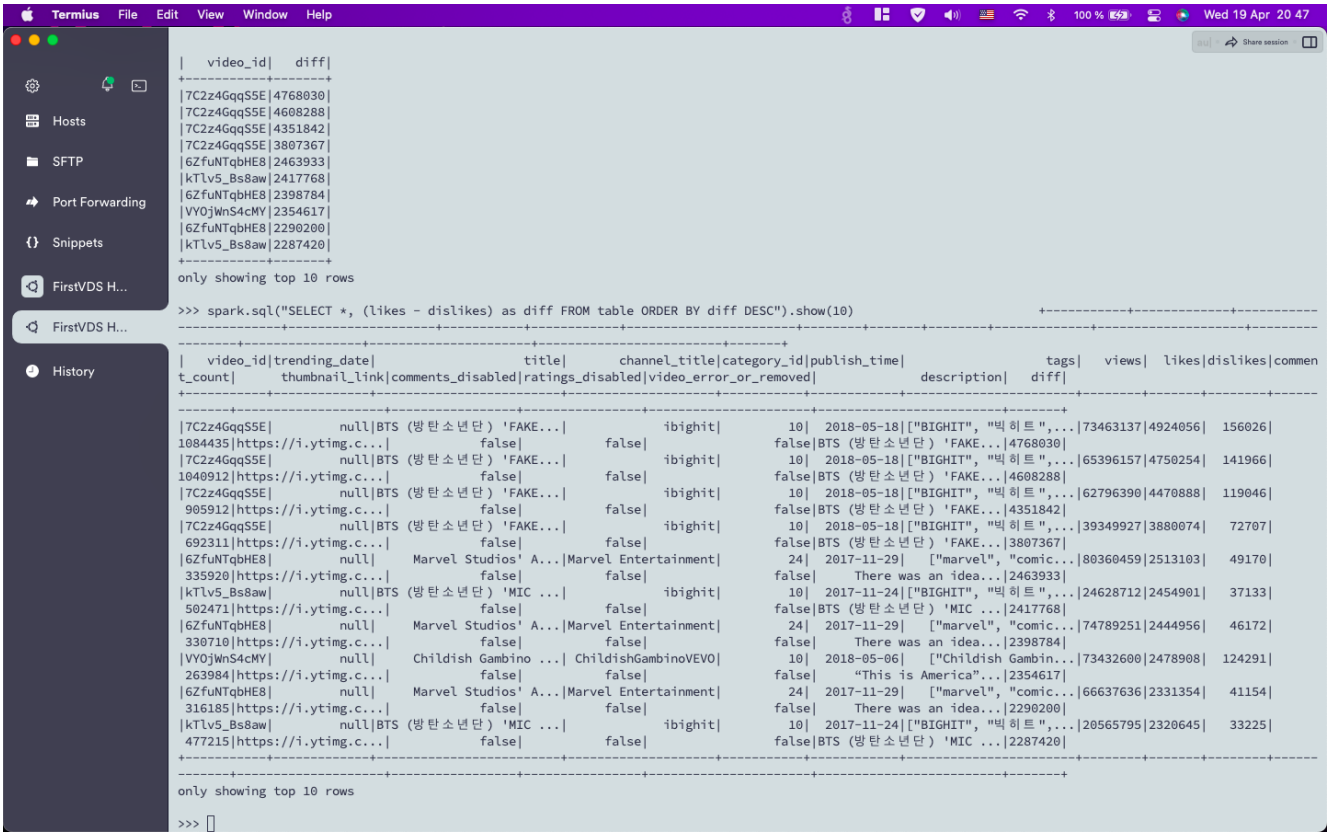


Your comments (optional)

<Spark SQL part>

Task 1

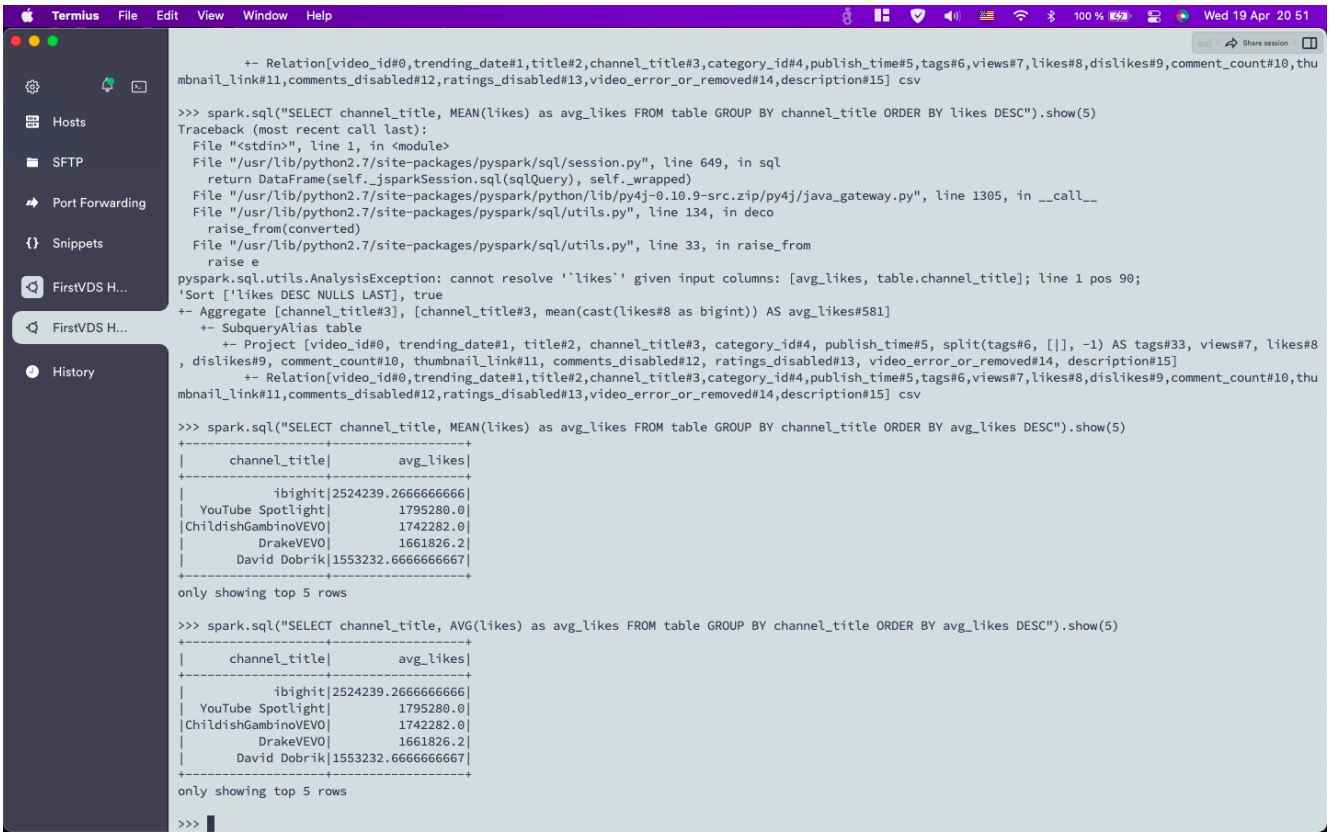
Your query as a text
spark.sql("SELECT *, (likes - dislikes) as diff FROM table ORDER BY diff DESC").show(10)



Your comments (optional)
df2.createOrReplaceTempView("table")

Task 2

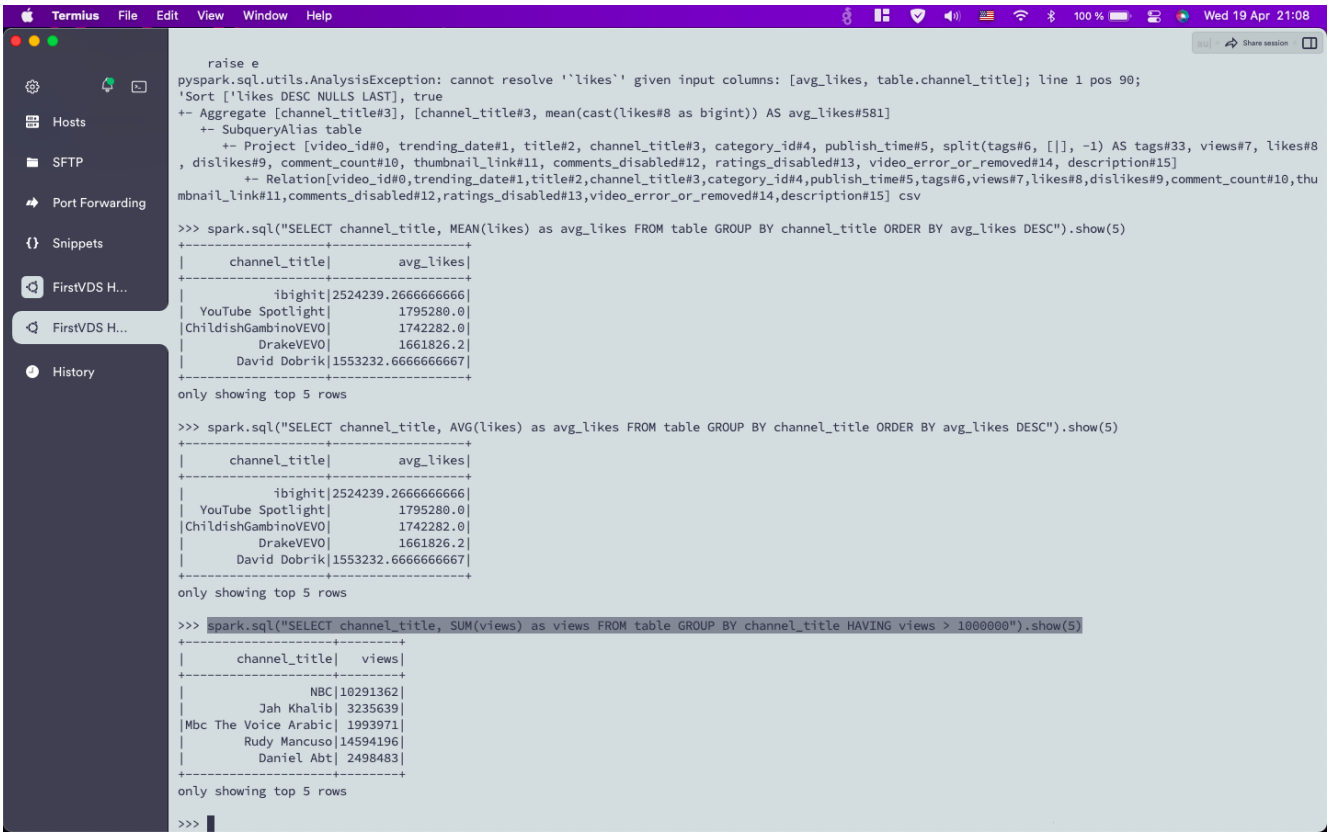
Your query as a text
spark.sql("SELECT channel_title, AVG(likes) as avg_likes FROM table GROUP BY channel_title ORDER BY avg_likes DESC").show(5)



Your comments (optional)

Task 3

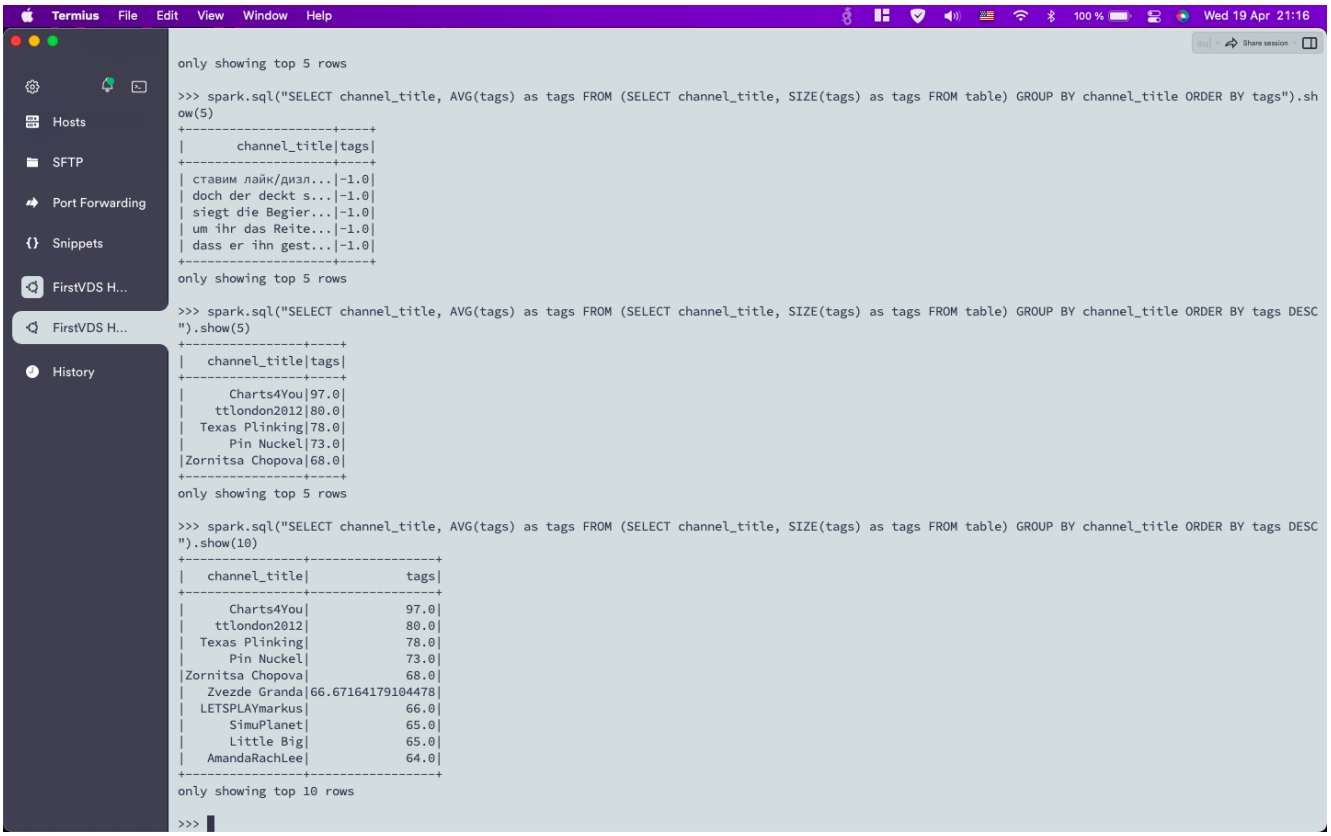
Your query as a text
spark.sql("SELECT channel_title, SUM(views) as views FROM table GROUP BY channel_title HAVING views > 1000000").show(5)



Your comments (optional)

Task 4

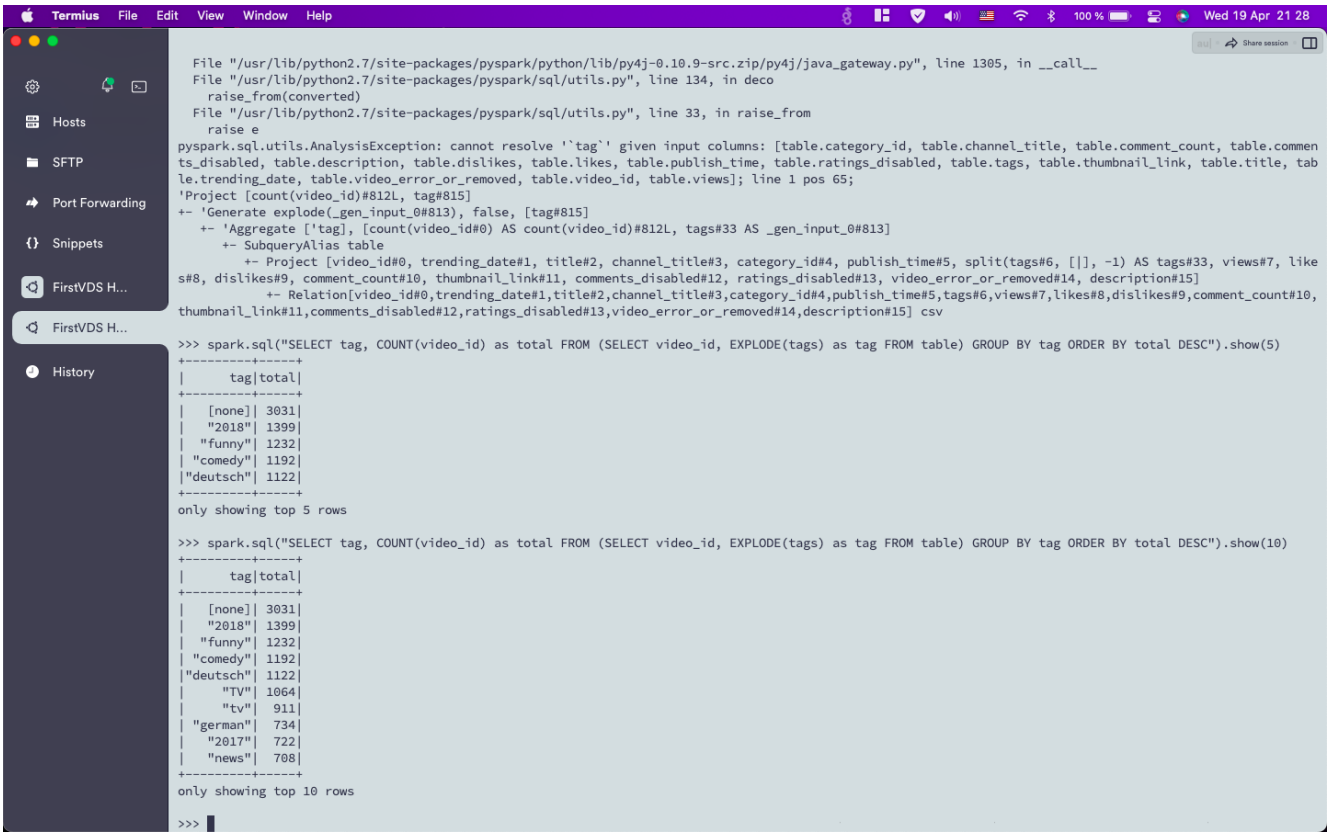
Your query as a text
spark.sql("SELECT channel_title, AVG(tags) as tags FROM (SELECT channel_title, SIZE(tags) as tags FROM table) GROUP BY channel_title ORDER BY tags DESC").show(10)



Your comments (optional)

Task 5

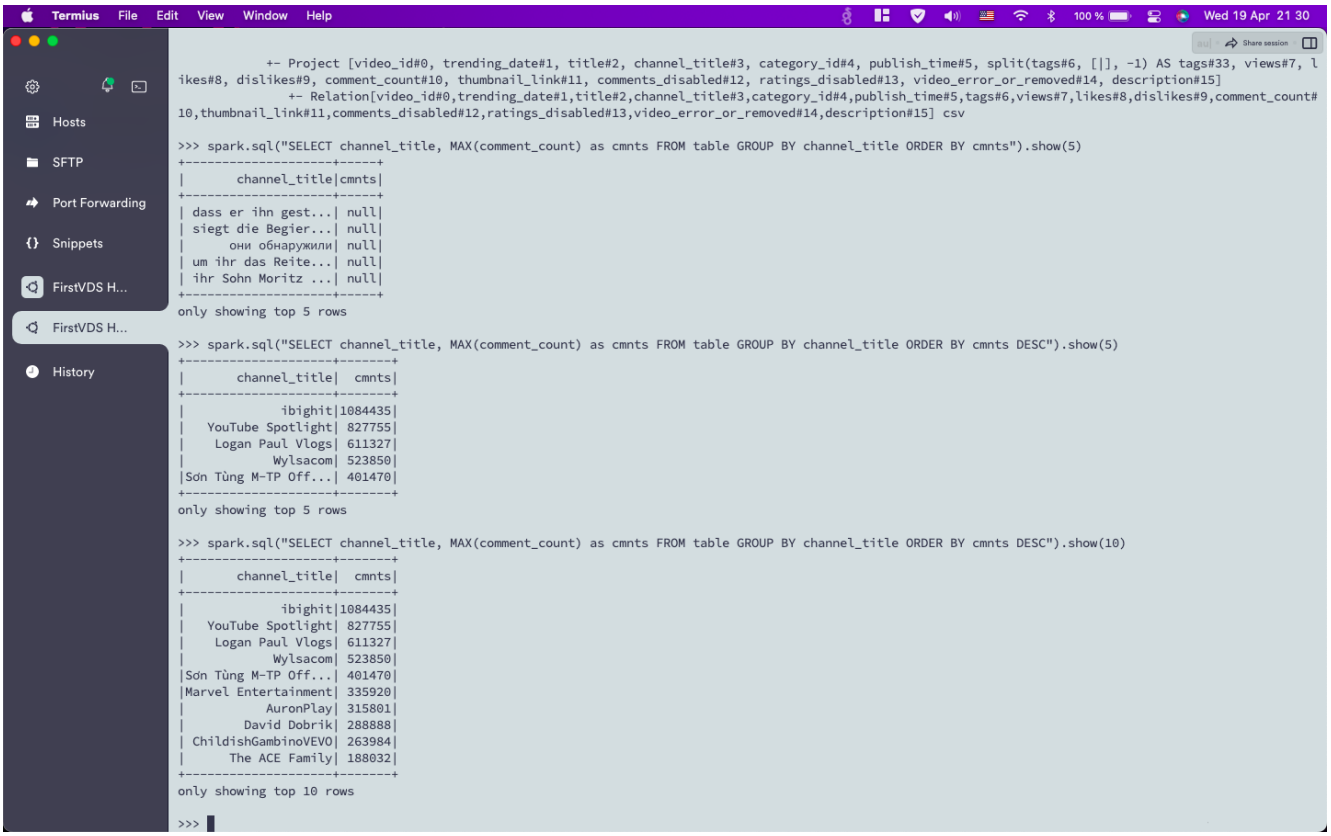
Your query as a text
spark.sql("SELECT tag, COUNT(video_id) as total FROM (SELECT video_id, EXPLODE(tags) as tag FROM table) GROUP BY tag ORDER BY total DESC").show(10)



Your comments (optional)

Task 6

Your query as a text
spark.sql("SELECT channel_title, MAX(comment_count) as cmnts FROM table GROUP BY channel_title ORDER BY cmnts DESC").show(10)



Your comments (optional)