

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[-2:]))
# id_title_dict

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

```
%pychronic
rdd = sc.textFile("file://" + os.path.join(SOURCE_PATH, "DE_category_id.json"))

Took 0 sec. Last updated by anonymous at April 21 2023, 7:45:39 PM.
```

```
%python2
d = rdd.collect()
d = [str(s.decode("utf-8")).strip() for s in d]
s = "\n".join(d)
idx = s.find("items")

import re

s = s[idx:]

raw_pairs = re.findall(r'"id":[]]+>(?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

... {'42': 'Shorts', '24': 'Entertainment', '25': 'News & Politics', '26': 'Howto & Style', '27': 'Education', '20': 'Gaming', '21': 'Videoblogging', '22': 'People & Blogs', '23': 'Comedy', '44': 'Trailers', '28': 'Science & Technology', '43': 'Shows', '40': 'Sci-Fi/Fantasy', '41': 'Thriller', '1': 'Film & Animation', '2': 'Autos & Vehicles', '10': 'Music', '39': 'Horror', '38': 'Foreign', '15': 'Pets & Animals', '17': 'Sports', '19': 'Travel & Events', '18': 'Short Movies', '31': 'Anime/Animation', '30': 'Movies', '37': 'Family', '36': 'Drama', '35': 'Documentary', '34': 'Comedy', '33': 'Classics', '32': 'Action/Adventure'}
```

```
%pychronic
videos_path = "file://" + os.path.join(SOURCE_PATH, "DEvideos.csv")

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

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)
```

Brave File Edit View History Bookmarks People Tab Window Help

Ambari - Sandbox 188.120.225.39 Apache Zeppelin 0.6 Zeppelin - Spark Job Hadoop vs. Spark: Lab 4 - Apache Sp Assignment 4 - Ap [BD] Assignment04 Tue 18 Apr 19:53

Not Secure | http://188.120.225.39:9995/#/notebook/2HX5CJ5SJ

Zeppelin Notebook Job Search your Notes anonymous default

ass

```
.option("inferSchema", "true")\n.option("header", "true")\\n.load(videos_path)
```

videos_df.show()

False	False	False	False	False	False	False	False
1mmMPrcmJeAl	17.14.11 Diese Schlafroutine was u....	17 2017-11-12T09:00:... "Liebscher" "Brac... 109571 4357 303 357 https://i.ytimg.c...					
False	False	False	False	False	False	False	False
l3US1cViqulMi	17.14.11 ARAB Lifestyle ... PlanetKanax	23 2017-11-13T11:59:... "planetkanax" "pl... 999881 63971 298 429 https://i.ytimg.c...					
False	False	False	False	False	False	False	False
lOKYUthVhgMhcI	17.14.11 Die Welt in 30 Ja.... VOLKANI	24 2017-11-12T16:09:... "Welt" "Natur" "U... 378771 18391 327 170 https://i.ytimg.c...					
False	False	False	False	False	False	False	False
lspwqQDCW4TI	17.14.11 RUNDLAUF FUßBALL ... Brotatos!	24 2017-11-13T14:03:... "fußball challeng... 1113281 106921 115 548 https://i.ytimg.c...					
False	False	False	False	False	False	False	False
l-MAGdUa3H9Al	17.14.11 Schariger_Geist.... Ultraaktiv!	27 2017-11-12T13:00:... "standartskill" "... 936281 79241 118 308 https://i.ytimg.c...					
False	False	False	False	False	False	False	False
lrlV8xuBqUQ0l	17.14.11 Duell der Gigante... Inscopelifestyle	22 2017-11-12T16:34:... "inscopelifestyle" "... 113961 88911 260 1385 https://i.ytimg.c...					
False	False	False	False	False	False	False	False
lnNtrjfx2aGwl	17.14.11 KOCHEN mit TANZVE... TANZVERBOT!	26 2017-11-13T19:00:... "tanzverbot" "Koc... 1219281 164671 650 3492 https://i.ytimg.c...					
False	False	False	False	False	False	False	False

only showing top 20 rows

Took 25 sec. Last updated by anonymous at April 18 2023, 7:48:13 PM. (outdated)

spark2.pyspark

```
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! Erwachsene abzocken - LUKE! Die Woche und ich | SAT.1'), Row(title=u'The Tramp Presidency: Last Week Tonight with John Oliver (HBO)'), Row(title=u'Das Fermi-Paradoxon'), Row(title=u'18 SONGS mit Kelly MissesVlog (Sing-off')]

Took 1 sec. Last updated by anonymous at April 18 2023, 7:53:08 PM.

spark2.pyspark

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|Berühmt werden ka...
[2hu_evXPpMM| 17.14.11|Dagi Bee wird Hei...| HerrNewsTime|
ra...| 228574| 11349| 990| 1049|https://i.ytimg.c...|
False|Dagi Bee wird Hei...
[2ky565vSYSE| 17.14.11|WE WANT TO TALK A...| CaseyNeistat|
artin| 748374| 57532| 2966| 15954|https://i.ytimg.c...|
False|SHANTELL'S CHANNE...
[2Zp-Qm3WkA| 17.14.11|JP Performance - ...| JP Performance|
ps...| 465883| 19928| 216| 1240|https://i.ytimg.c...|
False|Mal schauen was u...
[mmMPrcmJea| 17.14.11|Diese Schlafposit...|Liebscher & Brach...
ac...| 109571| 4357| 303| 357|https://i.ytimg.c...|
False|Weitere Infos zu ...
[3U5icViquLM| 17.14.11|ARAB Lifestyle | ...| PlanetKanax|
pl...| 99988| 6397| 298| 429|https://i.ytimg.c...|
False|Video mit Bodyfor...
[OKVUthvgMhc| 17.14.11|Die Welt in 30 Ja...| VOLKAN|
"U...| 37877| 1839| 327| 170|https://i.ytimg.c...|
False|Über die Zukunft ...
[spwqQDCW4Tl| 17.14.11|RUNDLAUF FUßBALL ...| Brotatos|
ng...| 111328| 10692| 115| 548|https://i.ytimg.c...|
False|Die Rundlauf Fußb...
[-MagduA3H9A| 17.14.11|Schauriger_Geist....| Ultralativ|
"[...| 93628| 7924| 118| 308|https://i.ytimg.c...|
False|Geister gibt es ü...
[riV8xUBqQ0| 17.14.11|Duell der Gigante...| Inscopelifestyle|
le...| 113961| 8891| 260| 1385|https://i.ytimg.c...|
False|Ihr wolltet ein z...
[nNtrjFX2aGw| 17.14.11|KOCHEN mit TANZE...| TANZVERBOT|
oc...| 121928| 16467| 650| 3492|https://i.ytimg.c...|
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! Erwachsene 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)

```

```

ra...| 228574| 11349|     990|          1049|https://i.ytimg.c...|
False|Dagi Bee wird Hei...|
[2kySGSvSYSE| 17.14.11|WE WANT TO TALK A...| CaseyNeistat|
artin| 748374| 57532|    2966|        15954|https://i.ytimg.c...|
False|SHANTELL'S CHANNEL...|
[22p-Qm3wJkA| 17.14.11|JP Performance - ...| JP Performance|
ps...| 465883| 19928|     216|       1240|https://i.ytimg.c...|
False|Mal schauen was u...|
[mmMPrcrn3eA| 17.14.11|Diese Schlafpositi...|Liebscher & Brach...|
ac...| 109571| 4357|     303|       357|https://i.ytimg.c...|
False|Weitere Infos zu ...|
[3U51cViqulM| 17.14.11|ARAB Lifestyle | ...| PlanetKanax|
pl...| 99988| 6397|     298|       429|https://i.ytimg.c...|
False|Video mit Bodyfor...|
[OKYUthVhgMhc| 17.14.11|Die Welt in 30 Ja...| VOLKAN|
"U...| 37877| 1839|     327|       170|https://i.ytimg.c...|
False|Über die Zukunft ...|
[sppwQDCW4TI| 17.14.11|RUNDLAUF FURBALL ...| Brotatos|
ng...| 111328| 10692|     115|       548|https://i.ytimg.c...|
False|Die Rundlauf FuBb...|
[-MagDuA3H9A| 17.14.11|Schauriger Geist....| Ultraletiv|
"..."| 93628| 7924|     118|       308|https://i.ytimg.c...|
False|Geister gibt es ü...|
[rIV8xuBqUQ0| 17.14.11|Duell der Gigante...| Inscopelifestyle|
le...| 113961| 8891|     260|       1385|https://i.ytimg.c...|
False|Ihr wolltet ein z...|
[nNtrjfX2aGw| 17.14.11|KOCHEN mit TANZVE...| TANZVERBOT|
oc...| 121928| 16467|     650|       3492|https://i.ytimg.c...|
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! Etwachsende 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')]
>>> video_category_views = videos_df.select("views").rdd
>>> video_category_views.take(5)
[Row(views=u'252786'), Row(views=u'797196'), Row(views=u'2418783'), Row(views=u'380247'), Row(views=u'822213')]
>>>

```

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)
```

```

of Eminem: https://goo.gl/AquNpo\nSubscribe for more: https://goo.gl/DxCrDV\n\nFor more visit: \nhttp://eminem.c
om\nhttp://facebook.com/eminem\nhttp://twitter.com/eminem\nhttp://instagram.com/eminem\nhttp://eminem.tumblr.com
\nhttp://shadyrecords.com\nhttp://facebook.com/shadyrecords\nhttp://twitter.com/shadyrecords\nhttp://instagram.com
/shadyrecords\nhttp://vevo.ly/gA7xKt", Row(video_id='0dBIkQ4MzIM', trending_date='17.14.11', title='PLUSH - Bad Unboxi
ng Fan Mail', channel_title='iDubbbzTV', category_id='23', publish_time='2017-11-13T17:00:00.000Z', tags='plush
|bad unboxing|unboxing|fan mail|iDubbzTV|"idubbztv2|"things|"best|"packages|"plushies|"chontent chop"
, views='1014651', likes='127791', dislikes='1687', comment_count='13030', thumbnail_link='https://i.ytimg.com/vi/0dBIkQ4MzIM/default.jpg', comments_disabled='False', ratings_disabled='False', video_error_or_removed='False', d
escription='Still got a lot of packages. Probably will last for another year. On a side note, more 2nd channel vids
soon. editing with premiere from now on, gon\' be a tedious transition, but i think it\'s for the best. \n\n\nSUBSCRIBE \u25ba http://www.youtube.com/subscription_center?add_user=iDubbzTV\n\nMain Channel \u25ba https://www.yo
utube.com/user/iDubbzTV\nSecond Channel \u25ba https://www.youtube.com/channel/UC-tsNNJ3yIw8MtPHGPWFQA\nGaming Ch
annel \u25ba https://www.youtube.com/channel/UCvNfFXNY023-mbrTh10YRXA\n\nWebsite \u25ba http://www.idubbz.com\n\n
Instagram \u25ba https://instagram.com/idubbz/\nTwitter \u25ba https://twitter.com/Idubbz\nFacebook \u25ba http
://www.facebook.com/IDubbz\nTwitch \u25ba https://www.twitch.tv/idubbz\n", Row(video_id='1jCsM1vtv78', trending
_date='17.14.11', title='THE LOGANG MADE HISTORY. LOL. AGAIN.', channel_title='Logan Paul Vlogs', category_id='24
', publish_time='2017-11-12T20:19:24.000Z', tags='logan paul vlog|logan paul|"logan|"paul|"olympics|"logan p
aul youtube|"vlog||"daily||"comedy||"hollywood||"parrot||"maverick||"bird||"maverick clothes||"logan paul dubai||"du
bai||"logan paul uae||"uae||"logan paul meet and greet||"logan paul meet up||"logan paul dubai mall||"the dubai mall"
||"logan paul arab||"arab||"arab||"logan paul age||"the maverick||"maverick merch", views='4477587', likes='29283
7', dislikes='4123', comment_count='36391', thumbnail_link='https://i.ytimg.com/vi/JzCsM1vtv78/default.jpg', comme
nts_disabled='False', ratings_disabled='False', video_error_or_removed='False', description='Join the movement. Be
a Maverick \u25ba https://ShopLoganPaul.com\n\nON CAN STOP US. \n\nSUBSCRIBE FOR DAILY VLOGS! \u25ba http://bit.
ly/Subscribe2Logan\n\nMusic:\nExistence Problem - Marius:\nhttps://fanlink.to/MariusExistenceProblem2017\nAnimal - The Siege:\nhttps://soundcloud.com/theseigemusic/animal\nCtrl+Alt+Del - The Siege:\nhttps://soundcloud.com/t
heseigemusic/crtlaltdel\n\nWatch Yesterday:\n2019s Vlog \u25ba https://youtu.be/TIqqEJg4qC\n\nADD ME ON:\nINSTA
GRAM: https://www.instagram.com/LoganPaul/\nTWITTER: https://twitter.com/LoganPaul\n\nA 22 year old kid l
iving in Los Angeles. I make comedy vids, travel a lot, I have a pretty colorful parrot named Maverick and a savage d
og named Kong. This is my life.\nhttps://www.youtube.com/LoganPaulVLogs)]\n>>> videos_rdd.filter(lambda x: int(x['views']) > 1000000).take(1)
[Row(video_id='1ZAPwftrAFY', trending_date='17.14.11', title='The Trump Presidency: Last Week Tonight with John Ol
iver (HBO)', channel_title='LastWeekTonight', category_id='24', publish_time='2017-11-13T07:30:00.000Z', tags='l
ast week tonight trump presidency||"last week tonight donald trump||"john oliver trump||"donald trump", views='2418
783', likes='97190', dislikes='6146', comment_count='12703', thumbnail_link='https://i.ytimg.com/vi/1ZAPwftrAFY/d
efault.jpg', comments_disabled='False', ratings_disabled='False', video_error_or_removed='False', description='On
e year after the presidential election, John Oliver discusses what we've learned so far and enlists our catheter cow
boy to teach Donald Trump what he hasn't.\n\nConnect with Last Week Tonight online.\n\nSubscribe to the Last Wee
k Tonight YouTube channel for more almost news as it almost happens: www.youtube.com/user/LastWeekTonight\n\nFind L
ast Week Tonight on Facebook like your mom would: http://Facebook.com/LastWeekTonight\n\nFollow us on Twitter for n
ews about jokes and jokes about news: http://Twitter.com/LastWeekTonight\n\nVisit our official site for all that ot
her stuff at once: http://www.hbo.com/lastweektonight")]
>>> videos_rdd.filter(lambda x: int(x['views']) > 1000000).map(lambda x: x["title"]).take(5)
[u'The Trump Presidency: Last Week Tonight with John Oliver (HBO)', u'Ed Sheeran - Perfect (Official Music Video)', u
'Eminem - Walk On Water (Audio) ft. Beyoncé', u'PLUSH - Bad Unboxing Fan Mail', u'THE LOGANG MADE HISTORY. LOL. AG
AIN.']
>>>

```

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)

```

Records\nhttp://velo.ly/gA7xKt"), Row(video_id=u'0dBiKQ4Mz1M', trending_date=u'17.14.11', title=u'PLUSH - Bad Unboxi
ng Fan Mail', channel_title=u'idubbbzTV', category_id=u'23', publish_time=u'2017-11-13T17:00:00.000Z', tags=u'"plush"
|"bad unboxing"|"unboxing"|"fan mail"|"idubbbzTV"|"idubbbzTV2"|"things"|"best"|"packages"|"plushies"|"chontent chop"
, views=u'1014651', likes=u'127791', dislikes=u'1687', comment_count=u'13030', thumbnail_link=u'https://i.ytimg.com/v
?idBiKQ4Mz1M/default.jpg', comments_disabled=u'False', ratings_disabled=u'False', video_error_or_removed=u'False', d
escription=u'Still got a lot of packages. Probably will last for another year. On a side note, more 2nd channel vids
soon. editing with premiere from now on, gon' be a tedious transition, but i think it's for the best. \n\n_\\|\nS
UBSCRIBE \u25ba http://www.youtube.com/subscription_center?add_user=idubbbzTV\nSecond Channel \u25ba https://www.youtube.com/channel/UC-tsNNJ3yIw8MtPH6PWFQA\\nGaming Ch
annel \u25ba https://www.youtube.com/channel/UCvhFFXNY023-mbrTh10YRXA\\n\\nWebsite \u25ba http://www.idubbbz.com\\n\\n
\\nInstagram \u25ba https://instagram.com/idubbbz/\\nTwitter \u25ba https://twitter.com/idubbbz\\nFacebook \u25ba http
://www.facebook.com/IDubbbz\\nTwitch \u25ba http://www.twitch.tv/idubbbz\\n", Row(video_id=u'jZCsM1vtv78', trending
_date=u'17.14.11', title=u'THE LOGANG MADE HISTORY', channel_title=u'Logan Paul Vlogs', category_id=u'24
', publish_time=u'2017-11-12T20:19:24.000Z', tags=u'"logan paul vlog"|"logan paul"|"logan"|"paul"|"olympics"|"logan p
aul youtube"|"vlog"|"daily"|"comedy"|"hollywood"|"parrot"|"maverick"|"bird"|"maverick clothes"|"logan paul dubai"|"du
bai"|"logan paul uae"|"uae"|"logan paul meet and greet"|"logan paul meet up"|"logan paul dubai mall"|"the dubai mall"
|"logan paul arab"|"arabic"|"arab"|"logan paul age"|"the maverick"|"maverick merch", views=u'4477587', likes=u'29283
7', dislikes=u'4123', comment_count=u'36391', thumbnail_link=u'https://i.ytimg.com/vi/jZCsM1vtv78/default.jpg', comme
nts_disabled=u'False', ratings_disabled=u'False', video_error_or_removed=u'False', description=u'Join the movement. B
e a Maverick \u25ba https://ShopLoganPaul.com\\nNO ONE CAN STOP US! \\nSUBSCRIBE FOR DAILY VLOGS! \u25ba http://bit.
ly/Subscribe2Logan\\n\\nMusic:\\nExistence Problem - Marius\\nhttps://fanlink.to/MariusExistenceProblem2017\\n\\nAnim
al - The Siege\\nhttps://soundcloud.com/theseigemusic/animal\\n\\nCtrl+Alt+Del - The Siege\\nhttps://soundcloud.com/t
heseigemusic/crtlaltdel\\n\\nWatch Yesterday\\n2019s Vlog \u25ba https://youtu.be/TIqqEg4q7c\\n\\nADE ME ON:\\nINSTA
GRAM: https://www.instagram.com/LoganPaul\\n\\nTWITTER: https://twitter.com/LoganPaul\\n\\n\\n22 year old kid l
iving in Los Angeles. I make comedy vids, travel a lot, I have a pretty colorful parrot named Maverick and a savage d
og named Kong. This is my life.\\nhttps://www.youtube.com/LoganPaulVlogs')
>>> videos_rdd.filter(lambda x: int(x["views"]) > 1000000).take(1)
[Row(video_id=u'12APwfrrtAFY', trending_date=u'17.14.11', title=u'The Trump Presidency: Last Week Tonight with John Ol
iver (HBO)', channel_title=u'LastWeekTonight', category_id=u'24', publish_time=u'2017-11-13T07:38:00.000Z', tags=u'"l
ast week tonight trump presidency"|"last week tonight donald trump"|"john oliver trump"|"donald trump"', views=u'2418
783', likes=u'97190', dislikes=u'6146', comment_count=u'12703', thumbnail_link=u'https://i.ytimg.com/vi/12APwfrrtAFY/d
efault.jpg', comments_disabled=u'False', ratings_disabled=u'False', video_error_or_removed=u'False', description=u'On
e year after the presidential election, John Oliver discusses what we've learned so far and enlists our catheter cowb
oy to teach Donald Trump what he hasn't.\\n\\nConnect with Last Week Tonight online...\\n\\nSubscribe to the Last Wee
k Tonight YouTube channel for more almost news as it almost happens: www.youtube.com/user/LastWeekTonight\\n\\nFind L
ast Week Tonight on Facebook like your mom would: http://Facebook.com/LastWeekTonight\\n\\nFollow us on Twitter for n
ews about jokes and jokes about news: http://Twitter.com/LastWeekTonight\\n\\nVisit our official site for all that ot
her stuff at once: http://www.hbo.com/lastweektonight')
>>> videos_rdd.filter(lambda x: int(x["views"]) > 1000000).map(lambda x: x["title"]).take(5)
[u'The Trump Presidency: Last Week Tonight with John Oliver (HBO)', u'Ed Sheeran - Perfect (Official Music Video)', u
'Eminem - Walk On Water (Audio) ft. Beyonc\x9e', u'PLUSH - Bad Unboxing Fan Mail', u'THE LOGANG MADE HISTORY. LOL. AG
AIN.']
>>> videos_rdd.filter(lambda x: int(x["views"]) > 1000000).map(lambda x: x["category_id"]).take(5)
[u'24', u'10', u'10', u'23', u'24']
>>> videos_rdd.filter(lambda x: int(x["views"]) > 1000000).map(lambda x: id_title.value[x["category_id"]]).take(5)
[u'Entertainment', u'Music', u'Music', u'Comedy', u'Entertainment']
>>>

```

Your comments (optional)

Task 7

Your query as a text

```

video_channel_titles = videos_df.select("channel_title").rdd
video_channel_titles.take(5)

```

```

i/0dBIkQ4Mz1M/default.jpg", comments_disabled=u'False', ratings_disabled=u'False', video_error_or_removed=u'False', description=u'Still got a lot of packages. Probably will last for another year. On a side note, more 2nd channel vids soon. editing with premiere from now on, gon\'t be a tedious transition, but i think it's for the best. \n\n_\\n\nUBSCRIBE \u25ba http://www.youtube.com/subscription_center?add_user=idubbbztv\\n\\nMain Channel \u25ba https://www.youtube.com/user/idubbbztv\\nSecond Channel \u25ba https://www.youtube.com/channel/UC-tsNNJ3yIw8MtPH6PWFQ\\nGaming Channel \u25ba https://www.youtube.com/channel/UChFFXNY0z3-mbrTh10YRXA\\n\\nWebsite \u25ba http://www.idubbbz.com\\n\\nInstagram \u25ba https://instagram.com/idubbbz/\\nTwitter \u25ba https://twitter.com/idubbbz\\nFacebook \u25ba http://www.facebook.com/IDubbbz\\nTwitch \u25ba http://www.twitch.tv/idubbbz\\n_\\nRow(video_id=u'jzCsMlvtvn78', trending_date=u'17.14.11', title=u'THE LOGANG MADE HISTORY. LOL. AGAIN.', channel_title=u'Logan Paul Vlogs', category_id=u'24', publish_time=u'2017-11-12T20:19:24.000Z', tags=u'"logan paul vlog"'||"logan paul"'||"logan"'||"paul"'||"olympics"'||"logan paul youtube"'||"vlog"'||"daily"'||"comedy"'||"hollywood"'||"parrot"'||"bird"'||"maverick clothes"'||"logan paul dubai"'||"dubai mall"'||"logan paul arab"'||"arabic"'||"arab"'||"logan paul age"'||"the maverick"'||"maverick merch"', views=u'4477587', likes=u'292837', dislikes=u'4123', comment_count=u'36391', thumbnail_link=u'https://i.ytimg.com/vi/JzCsMlvtvn78/default.jpg', comments_disabled=u'False', ratings_disabled=u'False', video_error_or_removed=u'False', description=u'Join the movement. Be a Maverick \u25ba https://ShopLoganPaul.com\\n\\nON CAN STOP US. \\nSUBSCRIBE FOR DAILY VLOGS! \u25ba http://bit.ly/Subscribe2Logan\\n\\nMusic:\\nExistence Problem - Marius\\nhttps://fanlink.to/MariusExistenceProblem2017\\n\\nAnim al - The Siege\\nhttps://soundcloud.com/theseigemusic/animal\\n\\nCtrl+Alt+Del - The Siege\\nhttps://soundcloud.com/hesigemusic/crlaltdel\\n\\nWatch Yesterday\\u2019 Vlog \u25ba https://youtu.be/TIqqE3g4q7c\\n\\nADD ME ON:\\nINSTAGRAM: https://www.instagram.com/LoganPaul\\n\\nTWITTER: https://twitter.com/LoganPaul\\n\\nI\\u2019m a 22 year old kid living in Los Angeles. I make comedy vids, travel a lot, I have a pretty colorful parrot named Maverick and a savage dog named Kong. This is my life.\\nhttps://www.youtube.com/LoganPaulVLogs\\n\\gt;>> videos_rdd.filter(lambda x: int(x["views"]) > 1000000).take(1)
[Row(video_id=u'12APwfraTFY', trending_date=u'17.14.11', title=u'The Trump Presidency: Last Week Tonight with John Oliver (HBO)', channel_title=u'LastWeekTonight', category_id=u'24', publish_time=u'2017-11-13T07:30:00.000Z', tags=u'"last week tonight trump presidency"'||"last week tonight donald trump"'||"john oliver trump"'||"donald trump\"", views=u'2418783', likes=u'97190', dislikes=u'6146', comment_count=u'12703', thumbnail_link=u'https://i.ytimg.com/vi/12APwfraTFY/default.jpg', comments_disabled=u'False', ratings_disabled=u'False', video_error_or_removed=u'False', description=u'One year after the presidential election, John Oliver discusses what we've learned so far and enlists our catheter cowboy to teach Donald Trump what he hasn't.\\n\\nConnect with Last Week Tonight online:\\n\\nSubscribe to the Last Week Tonight YouTube channel for more almost news as it almost happens: www.youtube.com/user/LastWeekTonight\\n\\nFind Last Week Tonight on Facebook like your mom would: http://Facebook.com/LastWeekTonight\\n\\nFollow us on Twitter for news about jokes and jokes about news: http://Twitter.com/LastWeekTonight\\n\\nVisit our official site for all that other stuff at once: http://www.hbo.com/lastweektonight\\n\\gt;>> videos_rdd.filter(lambda x: int(x["views"]) > 1000000).map(lambda x: x["title"]).take(5)
[u'The Trump Presidency: Last Week Tonight with John Oliver (HBO)', u'Ed Sheeran - Perfect (Official Music Video)', u'Eminem - Walk On Water (Audio) ft. Beyonc\x9e9', u'PLUSH - Bad Unboxing Fan Mail', u'THE LOGANG MADE HISTORY. LOL. AGAIN.']
\\gt;>> videos_rdd.filter(lambda x: int(x["views"]) > 1000000).map(lambda x: x["category_id"]).take(5)
[u'24', u'10', u'10', u'23', u'24']
\\gt;>> videos_rdd.filter(lambda x: int(x["views"]) > 1000000).map(lambda x: id_title.value[x["category_id"]]).take(5)
[u'Entertainment', u'Music', u'Music', u'Comedy', u'Entertainment']
\\gt;>> video_channel_titles = videos_df.select("channel_title").rdd
\\gt;>> video_channel_titles.take(5)
[Row(channel_title=u'inscope21'), Row(channel_title=u'LUKE! Die Woche und ich'), Row(channel_title=u'LastWeekTonight'), Row(channel_title=u'100SekundenPhysik'), Row(channel_title=u'rezo')]
\\gt;>>

```

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()

```

Screenshot of Termius application window showing a terminal session. The terminal output is filled with repeated occurrences of '(None, 5855),'. The session includes configuration panels for 'Terminal themes' (Termius Dark selected), 'Terminal font' (Source Code Pro), and 'Text Size' (14). A 'Share session' button is at the top right.

Screenshot of Termius application window showing a terminal session. The terminal output is identical to the one above, filled with repeated occurrences of '(None, 5855),'. The session includes configuration panels for 'Terminal themes' (Termius Dark selected), 'Terminal font' (Source Code Pro), and 'Text Size' (14). A 'Share session' button is at the top right.

<p>Your comments (optional)</p> <p>I'm not sure if this title is just a problem with the dataset. But I've checked there are empty video titles</p> <pre>from pyspark.sql.functions import col,isnan,when,count >>> df2 = df.select([count(when(col(c).contains('None') \ ... col(c).contains('NULL')) \</pre>

```
...     (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 ...
|PaWTa6Lie0|      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 lehrt 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"|"A...| 1
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 CHANNEL...
|[22p-Qm3wJkA|      null|JP Performance - ...|       JP Performance|   2| 2017-11-13|"V8"|"MAX"|"Tops...| 4
65883| 19928|      216|           1240|https://i.ytimg.c...|        false|        false|false|Ma
l schauen was u...
|[mmMPprcmJeA|      null|Diese Schlaufposit...|       Liebscher & Brach...|  17| 2017-11-12|"Liebscher"|"Brac...| 1
69571| 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...| 1
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...| 1
37877| 1839|      327|           170|https://i.ytimg.c...|        false|        false|false|Üb
er die Zukunft ...
|[spwgQDCW4TI|      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"|"...| 1
93628| 7924|      118|           308|https://i.ytimg.c...|        false|        false|false|Ge
ister gibt es ü...
|[riV8xuBqQ0|      null|Duell der Gigante...|       InscopeLifestyle| 22| 2017-11-12|"inscopelifestyle"|"...| 1
13961| 8891|      260|           1385|https://i.ytimg.c...|        false|        false|false|Ih
r wollten 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
>>> [REDACTED]

```

Terminal themes
 Termius Dark
 Termius Light
 Basic
 Homebrew
 Grass
 Man Page
 Novel
 Ocean

Terminal font
 Source Code Pro
 Text Size - 14 +
 Cancel Save

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]
```

The screenshot shows the Termius application window. On the left, there's a sidebar with icons for Hosts, SFTP, Port Forwarding, Snippets, and History. The main area displays a terminal session with the following Python code:

```

+-----+
+-----+
>>> likes = df.filter(f.col("publish_time") < f.lit("2010-01-01")).select("likes")
>>> dislikes = df.filter(f.col("publish_time") < f.lit("2010-01-01")).select("dislikes")
>>> 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
DataFrame[sum(likes): bigint]
>>> dislikes
DataFrame[sum(dislikes): bigint]
>>> likes.collect()
[Row(sum(likes)=None)]
>>> dislikes.collect()
[Row(sum(dislikes)=None)]
>>> likes - dislikes
Traceback (most recent call last):
  File "<stdin>", line 1, in <module>
TypeError: unsupported operand type(s) for -: 'DataFrame' and 'DataFrame'
>>> likes.value - dislikes.value
Traceback (most recent call last):
  File "<stdin>", line 1, in <module>
    File "/usr/lib/python2.7/site-packages/pyspark/sql/dataframe.py", line 1401, in __getattribute__
      "'%s' object has no attribute '%s'" % (self.__class__.__name__, name)
AttributeError: 'DataFrame' object has no attribute 'value'
>>> dislikes = df.filter(f.col("publish_time") > f.lit("2010-01-01")).select(f.sum("dislikes"))
>>> dislikes
DataFrame[sum(dislikes): bigint]
>>> dislikes.collect()
[Row(sum(dislikes)=57059031)]
>>> dislikes[0]
Column<sum(dislikes)>
>>> dislikes.value
Traceback (most recent call last):
  File "<stdin>", line 1, in <module>
    File "/usr/lib/python2.7/site-packages/pyspark/sql/dataframe.py", line 1401, in __getattribute__
      "'%s' object has no attribute '%s'" % (self.__class__.__name__, name)
AttributeError: 'DataFrame' object has no attribute 'value'
>>> dislikes.collect()[0][0]
57059031
>>> dislikes = df.filter(f.col("publish_time") < f.lit("2010-01-01")).select(f.sum("dislikes"))
>>> dislikes.collect()[0][0]
>>> likes.collect()[0][0] - dislikes.collect()[0][0]
Traceback (most recent call last):
  File "<stdin>", line 1, in <module>
TypeError: unsupported operand type(s) for -: 'NoneType' and 'NoneType'
>>> 

```

To the right of the terminal, there's a sidebar titled "Terminal themes" with a list of themes: Termius Dark (selected), Termius Light, Basic, Homebrew, Grass, Man Page, Novel, and Ocean. Below that is "Terminal font" set to "Source Code Pro" at size 14, with "Text Size" controls and "Cancel" and "Save" buttons.

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)
```

The screenshot shows the Termius application window. On the left, there's a sidebar with icons for Hosts, SFTP, Port Forwarding, Snippets, and History. The main terminal window displays several lines of Python code and their output. The code uses the pandas library to group data by channel title and calculate average likes. It includes operations like aggregation with aliasing, sorting, and filtering. The output shows results for channels like Achgut.Pogo, Nikad nije kasno, Emperor Caligula, NickiTV, and Bir Milletin Diri... The right side of the window has a sidebar for 'Terminal themes' (with options like Termius Dark, Termius Light, Basic, Homebrew, Grass, Man Page, Novel, and Ocean) and 'Terminal font' (set to Source Code Pro at size 14).

```
|     channel_title|      avg(likes)|  
+-----+-----+  
| Achgut.Pogo|993.68181818181|  
| Nikad nije kasno|          0.0|  
| Emperor Caligula|       310.0|  
|           NickiTV|      545.25|  
|Bir Milletin Diri...|      381.0|  
+-----+  
only showing top 5 rows  
  
>>> df.groupby("channel_title").agg(f.avg("likes").alias("avg_likes")).show(5)  
+-----+-----+  
|     channel_title|      avg_likes|  
+-----+-----+  
| Achgut.Pogo|993.68181818181|  
| Nikad nije kasno|          0.0|  
| Emperor Caligula|       310.0|  
|           NickiTV|      545.25|  
|Bir Milletin Diri...|      381.0|  
+-----+  
only showing top 5 rows  
  
>>> df.groupby("channel_title").agg(f.avg("likes").alias("avg_likes")).sort("avg_likes").show(5)  
+-----+-----+  
|     channel_title|      avg_likes|  
+-----+-----+  
| dass er ihn gest...|      null|  
| doch der deckt s...|      null|  
| als Feddersen be...|      null|  
| um ihr das Reite...|      null|  
| ihr Sohn Moritz ...|      null|  
+-----+  
only showing top 5 rows  
  
>>> df.groupby("channel_title").agg(f.avg("likes").alias("avg_likes")).sort(f.col("avg_likes").desc()).show(5)  
+-----+-----+  
|     channel_title|      avg_likes|  
+-----+-----+  
| ibight|2524239.266666666|  
| YouTube Spotlight|    1795280.0|  
| ChildishGambinoVEVO| 1742282.0|  
| DrakeVEVO|    1661826.2|  
| David Dobrik|1553232.666666667|  
+-----+  
only showing top 5 rows  
  
>>> |
```

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()
```

The screenshot shows a terminal window titled "Termius" with a dark theme. The left sidebar contains icons for Hosts, SFTP, Port Forwarding, Snippets, and two entries for "FirstVDS H...". The main pane displays a large list of YouTube channel titles and their total views. The list is scrollable and includes channels like 'REACT', 'KMNGANG', 'FilmSelect Trailer', 'SelenaGomezVEVO', 'NickiMinajAtVEVO', 'David Dobrik', 'BostonDynamics', 'ZaynVEVO', 'Fazilet Han\u0131im ve \u0131\u0131lar\u0131', 'Clash Royale', 'Dharma Productions', 'Screen Junkies', 'Casey Neistat', 'Nicki Minaj', 'Boston Dynamics', 'VikatanTV', 'Lucas the Spider', 'Daily Dose Of Internet', 'CinemaSins', 'JustintieberLakeVEVO', 'How Ridiculous', 'LuisFonsiVEVO', 'Eurovision Song Contest', 'a tv', 'Eurovision', 'NFL', 'FailArmy', 'The Film Theorists', 'Post MaloneVEVO', 'Julien Bam', 'SpaceX', 'Gelin', 'Maroon5VEVO', 'Bruno Mars', 'The Tonight Show Starring Jimmy Fallon', 'AsapSCIENCE', 'T-Series', 'Cardi B', 'Warner Bros. Pictures', 'DrakeVEVO', 'The Daily Show with Trevor Noah', 'Late Night with Seth Meyers', 'The Ellen Show', 'EminemVEVO', 'FoxStarHindi', 'LastWeekTonight', 'Late Show with Stephen Colbert', 'Jimmy Kimmel Live!', 'Logan Paul Vlogs', 'Kanye West', 'Kylie Jenner', 'Sen Anlat Karadeniz', 'MLG Highlights', 'WWE', 'Disney\u2022 Pixar', 'Nicky Jam', '20th Century Fox', '5-Minute Crafts', 'Ed Sheeran', 'MalumaVEVO', 'TaylorSwiftVEVO', 'ChildishGambinoVEVO', 'Universal Pictures', 'Sony Pictures Entertainment', 'PewDiePie', 'YouTube Spotlight', 'Dude Perfect', 'Marvel Entertainment', and 'Marvel Entertainment'. The bottom right corner of the terminal window has a "Terminal themes" section with options for "Termius Dark", "Termius Light", "Basic", "Homebrew", "Grass", "Man Page", "Novel", and "Ocean", along with a "Terminal font" section for "Source Code Pro" and a "Text Size" slider set to 14.

```

983), Row(channel_title=u'REACT', total_views=61172161), Row(channel_title=u'KMNGANG', total_views=61858571), Row(channel_title=u'FilmSelect Trailer', total_views=61994849), Row(channel_title=u'SelenaGomezVEVO', total_views=63273240), Row(channel_title=u'Marshmello', total_views=63382798), Row(channel_title=u'David Dobrik', total_views=63627292), Row(channel_title=u'NickiMinajAtVEVO', total_views=63810829), Row(channel_title=u'Dharma Productions', total_views=63966565), Row(channel_title=u'ZaynVEVO', total_views=64082613), Row(channel_title=u'Fazilet Han\u0131im ve \u0131\u0131lar\u0131', total_views=65954460), Row(channel_title=u'Dharmo Productions', total_views=66903489), Row(channel_title=u'Clash Royale', total_views=66984444), Row(channel_title=u'Screen Junkies', total_views=67345864), Row(channel_title=u'Casey Neistat', total_views=69439407), Row(channel_title=u'\u0410\u043b\u043d\u043b\u043e\u0431\u043b\u043e\u0432\u0430\u043d\u0430\u043b\u0435\u043d\u0439 \u0410\u043d\u0430\u043b\u0435\u043d\u0439', total_views=69529242), Row(channel_title=u'h3h3productions', total_views=7126747), Row(channel_title=u'VikatanTV', total_views=71334442), Row(channel_title=u'Lucas the Spider', total_views=71859515), Row(channel_title=u'Daily Dose Of Internet', total_views=71863309), Row(channel_title=u'\u0410\u043b\u043d\u043b\u043e\u0431\u043b\u043e\u0432\u0430\u043d\u0430\u043b\u0435\u043d\u0439', total_views=72375060), Row(channel_title=u'JP Performance', total_views=73996717), Row(channel_title=u'CinemaSins', total_views=74120746), Row(channel_title=u'justintieberLakeVEVO', total_views=74249232), Row(channel_title=u'How Ridiculous', total_views=75339388), Row(channel_title=u'LuisFonsiVEVO', total_views=75970330), Row(channel_title=u'Eurovision Song Contest', total_views=76246801), Row(channel_title=u'a tv', total_views=78195358), Row(channel_title=u'NFL', total_views=80619255), Row(channel_title=u'FailArmy', total_views=80836524), Row(channel_title=u'The Film Theorists', total_views=82989821), Row(channel_title=u'\u0430\u0432\u043d\u0430\u043b\u0435\u043d\u0439', total_views=83999811), Row(channel_title=u'Post MaloneVEVO', total_views=85599472), Row(channel_title=u'netd m\xfcyczki', total_views=86887238), Row(channel_title=u'jbalvinVEVO', total_views=88350630), Row(channel_title=u'Julien Bam', total_views=92517837), Row(channel_title=u'SpaceX', total_views=92937080), Row(channel_title=u'Troom Troom', total_views=93170594), Row(channel_title=u'\u0130stanbulu Gelin', total_views=93824821), Row(channel_title=u'FBE', total_views=99162586), Row(channel_title=u'MigosVEVO', total_views=100372196), Row(channel_title=u'Maroon5VEVO', total_views=101507314), Row(channel_title=u'Bruno Mars', total_views=106275320), Row(channel_title=u'The Tonight Show Starring Jimmy Fallon', total_views=107310809), Row(channel_title=u'AsapSCIENCE', total_views=107520729), Row(channel_title=u'T-Series', total_views=108701208), Row(channel_title=u'Cardi B', total_views=110322480), Row(channel_title=u'WORLDSTARHIPHOP', total_views=11212522), Row(channel_title=u'Warner Bros. Pictures', total_views=112727135), Row(channel_title=u'The Daily Show with Trevor Noah', total_views=112780388), Row(channel_title=u'DrakeVEVO', total_views=112993132), Row(channel_title=u'Philip DeFranco', total_views=114390135), Row(channel_title=u'Ariana GrandeVEVO', total_views=114816843), Row(channel_title=u'The Late Late Show with James Corden', total_views=115163227), Row(channel_title=u'Jimmy Kimmel Live!', total_views=11691672), Row(channel_title=u'Late Night with Seth Meyers', total_views=116918411), Row(channel_title=u'Logan Paul Vlogs', total_views=126182692), Row(channel_title=u'TheEllenShow', total_views=126306992), Row(channel_title=u'EminemVEVO', total_views=128904513), Row(channel_title=u'KSI', total_views=132550437), Row(channel_title=u'BuzzFeedVideo', total_views=133479295), Row(channel_title=u'FoxStarHindi', total_views=13533554), Row(channel_title=u'LastWeekTonight', total_views=147089108), Row(channel_title=u'The Late Show with Stephen Colbert', total_views=148944918), Row(channel_title=u'MLG Highlights', total_views=152333313), Row(channel_title=u'WWE', total_views=162801036), Row(channel_title=u'Disney\u2022 Pixar', total_views=177491595), Row(channel_title=u'Sen Anlat Karadeniz', total_views=185173924), Row(channel_title=u'Nicky Jam', total_views=190345736), Row(channel_title=u'Kylie Jenner', total_views=191009543), Row(channel_title=u'20th Century Fox', total_views=193517965), Row(channel_title=u'5-Minute Crafts', total_views=198762998), Row(channel_title=u'S\xfc\xfdz Dizi', total_views=204077348), Row(channel_title=u'MalumaVEVO', total_views=204787293), Row(channel_title=u'Ed Sheeran', total_views=205159549), Row(channel_title=u'TaylorSwiftVEVO', total_views=205965944), Row(channel_title=u'\u0131c7ukur', total_views=206207313), Row(channel_title=u'Universal Pictures', total_views=216443732), Row(channel_title=u'ChildishGambinoVEVO', total_views=227197666), Row(channel_title=u'Sony Pictures Entertainment', total_views=233030169), Row(channel_title=u'PewDiePie', total_views=30755426), Row(channel_title=u'YouTube Spotlight', total_views=368298641), Row(channel_title=u'ibright', total_views=368806027), Row(channel_title=u'Dude Perfect', total_views=408515774), Row(channel_title=u'Marvel Entertainment', total_views=585900476)]>>> [
```

Your comments (optional)

Task 5

Your query as a text

```
df.stat.corr("views", "likes")
```

```

Row(channel_title=u'Marshmello', total_views=6382798), Row(channel_title=u'David Dobrik', total_views=63627292), Ro
w(channel_title=u'NickiMinajAtEVO', total_views=63810829), Row(channel_title=u'BostonDynamics', total_views=63966565
), Row(channel_title=u'ZaynVEVO', total_views=64082613), Row(channel_title=u'Fazilet Han\u0131m ve Klu\u0131zlar\u0131', total_
views=65954460), Row(channel_title=u'Dharma Productions', total_views=66903489), Row(channel_title=u'Clash Ro
yale', total_views=66984444), Row(channel_title=u'Screen Junkies', total_views=67345864), Row(channel_title=u'CaseyNe
istat', total_views=69439407), Row(channel_title=u'\u0410\u043b\u043d\u043b\u043f\u0435\u043d\u0430\u0435\u043d\u0430\u0435', total_
views=69529242), Row(channel_title=u'h3h3productions', total_views=7126747), Row(channel_title=u'Daily Dose of Internet', total_
views=71863309), Row(channel_title=u'\u0410\u043b\u043d\u043b\u043f\u0435\u043d\u0430\u0435', total_views=71859515), Row(chann
el_title=u'\u043e\u043b\u043d\u043b\u043f\u0435\u043d\u0430\u0435', total_views=72375060), Row(channel_title=u'JP Performance', tot
al_views=73996717), Row(channel_title=u'CinemaSins', total_views=74120746), Row(channel_title=u'justintimberlakeEVO'
), total_views=74249232), Row(channel_title=u'How Ridiculous', total_views=75339388), Row(channel_title=u'LuisFonsiVEV
O', total_views=75970330), Row(channel_title=u'Eurovision Song Contest', total_views=76246801), Row(channel_title=u'a
tv', total_views=78195358), Row(channel_title=u'NFL', total_views=80619255), Row(channel_title=u'FailArmy', total_v
iews=80836524), Row(channel_title=u'The Film Theorists', total_views=82989821), Row(channel_title=u'\u0432\u0432\u0432\u0432
\u0432\u0432', total_views=83070300), Row(channel_title=u'Lil pump', total_views=83999811), Row(channel_title=u'PostMa
loneEVO', total_views=85599472), Row(channel_title=u'netd m\xfczci', total_views=86887238), Row(channel_title=u'jbal
vinEVO', total_views=88350630), Row(channel_title=u'Julien Bam', total_views=92517837), Row(channel_title=u'SpaceX'
), total_views=92937080), Row(channel_title=u'Trooo Troom', total_views=93170594), Row(channel_title=u'\u0130stanbullu
Gelin', total_views=93824821), Row(channel_title=u'FBE', total_views=99162586), Row(channel_title=u'MigosVEVO', total_
views=100372196), Row(channel_title=u'Maroon5VEVO', total_views=101507314), Row(channel_title=u'Brune Mars', total_v
iews=106275320), Row(channel_title=u'The Tonight Show Starring Jimmy Fallon', total_views=107310809), Row(channel_tit
le=u'AsapSCIENCE', total_views=107520729), Row(channel_title=u'T-Series', total_views=108701208), Row(channel_title=u
'Cardi B', total_views=110322480), Row(channel_title=u'WORLDSTARHIPHOP', total_views=112122522), Row(channel_title=u
'Warner Bros. Pictures', total_views=112727135), Row(channel_title=u'The Daily Show with Trevor Noah', total_views=112
780388), Row(channel_title=u'DrakeEVO', total_views=112993132), Row(channel_title=u'Philip DeFranco', total_views=11
4390135), Row(channel_title=u'ArianaGrandeVeo', total_views=114816843), Row(channel_title=u'The Late Late Show with
James Corden', total_views=115163227), Row(channel_title=u'Jimmy Kimmel Live!', total_views=116791672), Row(channel_tit
le=u'Late Night with Seth Meyers', total_views=116918411), Row(channel_title=u'Logan Paul Vlogs', total_views=126182
692), Row(channel_title=u'TheEllenShow', total_views=126306992), Row(channel_title=u'EminemVEVO', total_views=1289045
13), Row(channel_title=u'KSI', total_views=132550437), Row(channel_title=u'BuzzFeedVideo', total_views=13479295), Ro
w(channel_title=u'FoxStarHindi', total_views=13533554), Row(channel_title=u'LastWeekTonight', total_views=147089108)
, Row(channel_title=u'DrakeVEVO', total_views=148944918), Row(channel_title=u'MLG Highlights', total_views=15233313), Row
(channel_title=u'WWE', total_views=162801036), Row(channel_title=u'Disney\u0202 Pixar', total_views=177491595), Row(chann
el_title=u'Sean Anlat Karadeniz', total_views=185173924), Row(channel_title=u'NickyJa mtV', total_views=190345736), Row(chann
el_title=u'Kylie Jenner', total_views=191009543), Row(channel_title=u'20th Century Fox', total_views=193517965), Row(chann
el_title=u'5\xfcf62 Dizi', total_views=19550437), Row(channel_title=u'5-Minute Crafts', total_views=198762998), Row(chann
el_title=u'Ed Sheran', total_views=205159549), Row(channel_title=u'TaylorSwiftEVO', total_views=204787293), Row(channel_title=u
'\xc3\x7ukur', total_views=206207313), Row(channel_title=u'Universal Pictures', total_views=227197666), Row(channel_title=u
'ChildishGambinoVEVO', total_views=233030169), Row(channel_title=u'PewDiePie', total_views=307554426), Row(channel_title=u'YouTube Spotlight', total_
views=368860627), Row(channel_title=u'Dude Perfect', total_views=408515774), Row(channel_title=u'Marvel Entertainment', total_
views=585900476)
>>> df.stat().corr("views", "likes")
0.8253504735516626
>>> 
```

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)

```

The screenshot shows a macOS desktop with the Termius application open. The window title is 'Termius'. The main area displays a terminal session with the following content:

```

-- Project [video_id#950, trending_date#951, title#952, channel_title#953, category_id#954, publish_time#955, tags#1082, views#957, likes#958, dislikes#959, comment_count#960, thumbnail_link#961, comments_disabled#962, ratings_disabled#963, video_error_or_removed#964, description#965]
    -- Project [video_id#950, trending_date#951, title#952, channel_title#953, category_id#954, publish_time#955, tags#956, views#957, likes#958, dislikes#959, comment_count#960, thumbnail_link#961, comments_disabled#962, ratings_disabled#963, video_error_or_removed#964, description#965] csv

>>> 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")).show(10)
+-----+
| channel_title|avg_tags_size|
+-----+
| doch der deckt s...|      -1.0|
| ставим лайк/дизл...|      -1.0|
| Schlaganfall|      -1.0|
| ihr Sohn Moritz ...|      -1.0|
| scheint alle Müh...|      -1.0|
| als diese nach e...|      -1.0|
| and culture.\n|      -1.0|
| dass er ihn gest...|      -1.0|
| um ihr das Reite...|      -1.0|
| siegt die Begier...|      -1.0|
+-----+
only showing top 10 rows

>>> 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)
+-----+
| channel_title| avg_tags_size|
+-----+
| Charts4You|      97.0|
| ttLondon2012|      80.0|
| Texas Plinking|      78.0|
| Pin Nuckel|      73.0|
| Zornitsa Chopova|      68.0|
| Zvezde Granda| 66.67164179104478|
| LETSPLAYmarkus|      66.0|
| SimuPlanet|      65.0|
| Little Big|      65.0|
| AmandaRachLee|      64.0|
+-----+
only showing top 10 rows
>>>

```

On the right side of the terminal window, there are configuration panels for 'Terminal themes' (Termius Dark, Termius Light, Basic, Homebrew, Grass, Man Page, Novel, Ocean) and 'Terminal font' (Source Code Pro). Below these are buttons for 'Text Size' (14), 'Cancel', and 'Save'.

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)
```

The screenshot shows a terminal session within the Termius app. The terminal window displays Scala code running on a DataFrame named df2. The code performs several operations: grouping by 'video_id', exploding the 'tags' column into multiple rows, aliasing 'tags' as 'tag', grouping by 'tag', and counting the number of 'video_id's for each tag. It then shows the top 5 rows of the resulting DataFrame. The terminal also shows two additional queries: one for the frequency of each tag and another for the top 10 most frequent tags, sorted in descending order.

```
Traceback (most recent call last):
  File "<stdin>", line 1, in <module>
AttributeError: 'GroupedData' object has no attribute 'select'
>>> df2.select(f.col("video_id"), f.explode("tags").alias("tag")).groupby("tag").agg(f.count("video_id")).show(5)
+-----+-----+
| tag|count(video_id)|
+-----+-----+
| "Sat.1"| 100|
| "meteorit"| 1|
| "Sarkasmus"| 61|
| "Drama"| 83|
| "nusret şenay"| 125|
+-----+
only showing top 5 rows

>>> df2.select(f.col("video_id"), f.explode("tags").alias("tag")).groupby("tag").agg(f.count("video_id")).alias("freq")
).show(5)
+-----+-----+
| tag|freq|
+-----+-----+
| "Sat.1"| 100|
| "meteorit"| 1|
| "Sarkasmus"| 61|
| "Drama"| 83|
| "nusret şenay"| 125|
+-----+
only showing top 5 rows

>>> 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)
+-----+-----+
| tag|freq|
+-----+-----+
| [none]| 3031|
| "2018"| 1399|
| "funny"| 1232|
| "comedy"| 1192|
| "deutsch"| 1122|
| "TV"| 1064|
| "tv"| 911|
| "german"| 734|
| "2017"| 722|
| "news"| 708|
+-----+
only showing top 10 rows
>>> 
```

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)
```

```

| Logan Paul Vlogs | 611327 |
| Wylascom | 523850 |
| Sơn Tùng M-TP Off... | 401470 |
| Marvel Entertainment | 335920 |
| AuronPlay | 315801 |
| David Dobrik | 288888 |
| ChildishGambinoEVO | 263984 |
| The ACE Family | 188032 |
+-----+
only showing top 10 rows

>>> df2.filter(f.col("comment_count") == 1084435).show(5)
+-----+-----+-----+-----+-----+-----+-----+-----+
| 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|
+-----+-----+-----+-----+-----+-----+-----+-----+
| 7C2z4GqqS5E | null|BTS (방탄소년단) 'FAKE...| ibighit| 10| 2018-05-18|[BIGHIT, "弼...|7
3463137|4924056| 156026| 1084435|https://i.ytimg.c...| false| false| false
| BTS (방탄소년단) 'FAKE... |
+-----+-----+-----+-----+-----+-----+-----+-----+
>>> df2.groupby("channel_title").agg(f.max("comment_count").alias("max_comment_count")).sort(f.col("max_comment_count").desc()).show(10)
+-----+-----+
| channel_title|max_comment_count|
+-----+-----+
| ibighit| 1084435 |
| YouTube Spotlight| 827755 |
| Logan Paul Vlogs| 611327 |
| Wylascom| 523850 |
| Sơn Tùng M-TP Off...| 401470 |
| Marvel Entertainment| 335920 |
| AuronPlay| 315801 |
| David Dobrik| 288888 |
| ChildishGambinoEVO| 263984 |
| The ACE Family| 188032 |
+-----+
only showing top 10 rows

>>>

```

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)

```

The screenshot shows the Termius application window on a Mac OS X desktop. The main area is a terminal window displaying Scala code. The code defines a function `extract_link_udf` that splits a string into a list of URLs if they start with "https:", "http:", or "@".

```
...     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)
+-----+-----+
|<lambda>(description)|
+-----+-----+
|[https://www.inscope21.com/|
zu, https://www.youtube.com/user/unsympathischtv
Shpendi:, https://www.youtube.com/channel/UCm32Mjdv30hNrDT-RXExViA
Peter:, https://www.youtube.com/channel/UCvZxLwvrXrRf1mENhqJKkw
[]|
[]|
|[http://Facebook.com/LastWeekTonight
Follow, http://Twitter.com/LastWeekTonight
Visit, http://www.hbo.com/lastweektonight]
|
|[http://bit.ly/1fa7Tw3
@Snapchat:, http://on.fb.me/YJFlNT
@Twitter:, http://bit.ly/Zqgnv4
@Abonnieren:, http://bit.ly/10jgdi2
100SekundenPhysik:]
|[https://youtu.be/1ZMZTC67F6k
MEIN, http://www.rezo-shop.de
@, http://instagram.com/rezomusik
Twitter: https://twitter.com/rezomusik
Facebook:, https://www.facebook.com/rezomusik/
Snapchat:]|
+-----+
only showing top 5 rows
>>> |
```

To the right of the terminal is a sidebar titled "Terminal themes" which lists several color schemes: Termius Dark (selected), Termius Light, Basic, Homebrew, Grass, Man Page, Novel, and Ocean. Below the themes is a "Terminal font" section with "Source Code Pro" selected and a "Text Size" slider set to 14. At the bottom of the sidebar are "Cancel" and "Save" buttons.

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)
```

á, http://short.t-designz.de/vUyf3

á á , http://short.t-designz.de/V45MA, http://short.t-designz.de/3Zgrk
á, http://short.t-designz.de/Iyse
á, http://short.t-designz.de/SgJOF
á, http://short.t-designz.de/g7rnQ
á, http://short.t-designz.de/vUyf3
á, http://short.t-designz.de/szLOX
á, http://amzn.to/2vRhzCw
á, http://short.t-designz.de/rAyj0
á, http://short.t-designz.de/Cinfq
á, http://short.t-designz.de/IhgsM
á, http://amzn.to/2BuGUL
á, http://short.t-designz.de/gTnJP
á, http://short.t-designz.de/vBXh1
á, http://short.t-designz.de/S8NHZ
á, http://short.t-designz.de/YiQDM
á, http://short.t-designz.de/VBGtu
á, http://amzn.to/2k10Bna
áObjektiv:, http://amzn.to/2i9LEF2
áMikrofon:, http://amzn.to/2i9o7PY

áí,, https://lets-bastel.de/weiteres/newsletter
Á
á", http://short.t-designz.de/ls3aQ
Á
á , http://short.t-designz.de/zaltF
Á
áá , http://short.t-designz.de/4cJKL
á-á , http://short.t-designz.de/gVbM1
Á
áöÁ , http://short.t-designz.de/Vtdc6
áÁ , http://short.t-designz.de/jIq2W
á?Á 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)
+-----+
| title | list |
+-----+
malloreddus	[]
! Trump ! - Make ...	[https://www.yout...]
!! THIS VIDEO IS ...	[https://markipli...]
!EXKLUSIV! Der Fa...	[https://volksleh...]
#1 Bandsäge sel...	[http://amzn.to/2...]
+-----+
only showing top 5 rows
>>> |

-

| malloreddus | [] |
+-----+
| ! Trump ! - Make the Amerikan - Fake again ! Agenda der Regierungen - Bush Clinton Obama 4.0|[https://www.youtube.com/channel/UCug6UysHiBASbPX6ASdmedw/videos, @, https://vk.com/uwe.tagesschlau
| , https://tagesschlau-online.jimdo.com, https://www.youtube.com/channel/UCug6UysHiBASbPX6ASdmedw
|-]

| !!! THIS VIDEO IS NOTHING BUT PAIN !! | Getting Over It - Part 7 | [https://markiplier.co
m/]

| !EXKLUSIV! Der Fall Ursache - Der Volkslehrer im Gespräch mit Adrian | [https://volkslehrer.i
nfoä¶, https://www.youtube.com/watch?v=imhE7z8oLd8]

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)

The screenshot shows a terminal window titled 'Termius' with a purple header bar. The menu bar includes File, Edit, View, Window, Help, and a date/time indicator 'Wed 19 Apr 2047'. The terminal window displays a command-line interface with several tabs: 'Hosts', 'SFTP', 'Port Forwarding', 'Snippets', 'FirstVDS H...', 'FirstVDS H...', and 'History'. The 'History' tab is active, showing the execution of a Spark SQL query:

```
|>>> spark.sql("SELECT *, (likes - dislikes) as diff FROM table ORDER BY diff DESC").show(10)
```

The output shows the top 10 rows of the DataFrame, including columns like video_id, diff, trending_date, title, channel_title, category_id, publish_time, tags, views, likes, dislikes, and comments.

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)

```
+-- Relation[video_id#0,trending_date#1,title#2,channel_title#3,category_id#4,publish_time#5,tags#6,views#7,likes#8,dislikes#9,comment_count#10,thumbna...  
>>> spark.sql("SELECT channel_title, MEAN(likes) as avg_likes FROM table GROUP BY channel_title ORDER BY likes DESC").show(5)  
Traceback (most recent call last):  
  File "<stdin>", line 1, in <module>  
    File "/usr/lib/python2.7/site-packages/pyspark/sql/session.py", line 649, in sql  
      return DataFrame(self._jsparkSession.sql(sqlQuery), self._wrapped)  
    File "/usr/lib/python2.7/site-packages/pyspark/python/lib/py4j-0.10.9-src.zip/py4j/java_gateway.py", line 1305, in __call__  
    File "/usr/lib/python2.7/site-packages/pyspark/sql/utils.py", line 134, in deco  
      raise_from(converted)  
    File "/usr/lib/python2.7/site-packages/pyspark/sql/utils.py", line 33, in raise_from  
      raise e  
pyspark.sql.utils.AnalysisException: cannot resolve 'likes' given input columns: [avg_likes, table.channel_title]; line 1 pos 90;  
'Sort ['likes DESC NULLS LAST], true  
'-- Aggregate [channel_title#3], [channel_title#3, mean(cast(likes#8 as bigint)) AS avg_likes#581]  
  +- SubqueryAlias table  
    +- Project [video_id#0, trending_date#1, title#2, channel_title#3, category_id#4, publish_time#5, split(tags#6, [], -1) AS tags#33, views#7, likes#8, dislikes#9, comment_count#10, thumbnail_link#11, comments_disabled#12, ratings_disabled#13, video_error_or_removed#14, description#15]  
    +- Relation[video_id#0,trending_date#1,title#2,channel_title#3,category_id#4,publish_time#5,tags#6,views#7,likes#8,dislikes#9,comment_count#10,thumbna...  
>>> spark.sql("SELECT channel_title, MEAN(likes) as avg_likes FROM table GROUP BY channel_title ORDER BY avg_likes DESC").show(5)  
+-----+-----+  
| channel_title | avg_likes |  
+-----+-----+  
| ibighit|2524239.2666666666|  
| YouTube Spotlight| 1795280.0|  
| ChildishGambinoVEVO| 1742282.0|  
| DrakeVEVO| 1661826.2|  
| David Dobrik|1553232.6666666667|  
+-----+-----+  
only showing top 5 rows  
  
>>> spark.sql("SELECT channel_title, AVG(likes) as avg_likes FROM table GROUP BY channel_title ORDER BY avg_likes DESC").show(5)  
+-----+-----+  
| channel_title | avg_likes |  
+-----+-----+  
| ibighit|2524239.2666666666|  
| YouTube Spotlight| 1795280.0|  
| ChildishGambinoVEVO| 1742282.0|  
| DrakeVEVO| 1661826.2|  
| David Dobrik|1553232.6666666667|  
+-----+-----+  
only showing top 5 rows  
  
>>> 
```

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)
```

```

raise e
pyspark.sql.utils.AnalysisException: cannot resolve ``likes'' given input columns: [avg_likes, table.channel_title]; line 1 pos 90;
"Sort ['likes DESC NULLS LAST], true
-- Aggregate [channel_title#3], [channel_title#3, mean(cast(likes#8 as bigint)) AS avg_likes#581]
  +- SubqueryAlias table
    +- Project [video_id#0, trending_date#1, title#2, channel_title#3, category_id#4, publish_time#5, split(tags#6, [], -1) AS tags#33, views#7, likes#8
, dislikes#9, comment_count#10, thumbnail_link#11, comments_disabled#12, ratings_disabled#13, video_error_or_removed#14, description#15]
      +- Relation[video_id#0,trending_date#1,title#2,channel_title#3,category_id#4,publish_time#5,tags#6,views#7,likes#8,dislikes#9,comment_count#10,thu
mbnail_link#11,comments_disabled#12,ratings_disabled#13,video_error_or_removed#14,description#15] csv

>>> spark.sql("SELECT channel_title, MEAN(likes) as avg_likes FROM table GROUP BY channel_title ORDER BY avg_likes DESC").show(5)
+-----+-----+
| channel_title| avg_likes|
+-----+-----+
| ibighit|2524239.2666666666|
| YouTube Spotlight| 1795280.0|
| ChildishGambinoVEVO| 1742282.0|
| DrakeVEVO| 1661826.2|
| David Dobrik|1553232.666666667|
+-----+-----+
only showing top 5 rows

>>> spark.sql("SELECT channel_title, AVG(likes) as avg_likes FROM table GROUP BY channel_title ORDER BY avg_likes DESC").show(5)
+-----+-----+
| channel_title| avg_likes|
+-----+-----+
| ibighit|2524239.2666666666|
| YouTube Spotlight| 1795280.0|
| ChildishGambinoVEVO| 1742282.0|
| DrakeVEVO| 1661826.2|
| David Dobrik|1553232.666666667|
+-----+-----+
only showing top 5 rows

>>> spark.sql("SELECT channel_title, SUM(views) as views FROM table GROUP BY channel_title HAVING views > 1000000").show(5)
+-----+-----+
| channel_title| views|
+-----+-----+
| NBC|10291362|
| Jah Khalib| 3235639|
| Mbc The Voice Arabic| 1993971|
| Rudy Mancuso|14594196|
| Daniel Abt| 2498483|
+-----+-----+
only showing top 5 rows
>>> 

```

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)
```

The screenshot shows a Termius session window with the following content:

```
only showing top 5 rows
>>> 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").show(5)
+-----+-----+
| channel_title|tags|
+-----+-----+
| ставим лайк/дизл...|-1.0|
| doch der deckt s...|-1.0|
| siegt die Begier...|-1.0|
| um ihr das Reite...|-1.0|
| dass er ihn gest...|-1.0|
+-----+
only showing top 5 rows

>>> 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(5)
+-----+-----+
| channel_title|tags|
+-----+-----+
| Charts4You|97.0|
| ttlondon2012|80.0|
| Texas Plinking|78.0|
| Pin Nuckel|73.0|
| Zornitsa Chopova|68.0|
+-----+
only showing top 5 rows

>>> 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)
+-----+-----+
| channel_title|      tags|
+-----+-----+
| Charts4You|      97.0|
| ttlondon2012|     80.0|
| Texas Plinking|    78.0|
| Pin Nuckel|    73.0|
| Zornitsa Chopova| 68.0|
| Zvezde Granda| 66.67164179104478|
| LETSPLAYmarkus| 66.0|
| SimuPlanet| 65.0|
| Little Big| 65.0|
| AmandaRachLee| 64.0|
+-----+
only showing top 10 rows
>>> 
```

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)
```

```

File "/usr/lib/python2.7/site-packages/pyspark/python/lib/py4j-0.10.9-src.zip/py4j/java_gateway.py", line 1305, in __call__
  File "/usr/lib/python2.7/site-packages/pyspark/sql/utils.py", line 134, in deco
    raise_from(converted)
  File "/usr/lib/python2.7/site-packages/pyspark/sql/utils.py", line 33, in raise_from
    raise e
pyspark.sql.utils.AnalysisException: cannot resolve ``tag`` given input columns: [table.category_id, table.channel_title, table.comment_count, table.comments_disabled, table.description, table.dislikes, table.likes, table.publish_time, table.ratings_disabled, table.tags, table.thumbnail_link, table.title, table.trending_date, table.video_error_or_removed, table.video_id, table.views]; line 1 pos 65;
'Project [count(video_id)#812L, tag#815]
+- 'Generate explode(_gen_input_0#813), false, [tag#815]
   +- 'Aggregate ['tag], [count(video_id#0) AS count(video_id#0) AS count(video_id#812L, tag#815) AS _gen_input_0#813]
      +- SubqueryAlias table
         +- Project [video_id#0, trending_date#1, title#2, channel_title#3, category_id#4, publish_time#5, split(tags#6, [], -1) AS tags#33, views#7, like#8, dislikes#9, comment_count#10, thumbnail_link#11, comments_disabled#12, ratings_disabled#13, video_error_or_removed#14, description#15]
            +- Relation[video_id#0,trending_date#1,title#2,channel_title#3,category_id#4,publish_time#5,tags#6,views#7,likes#8,dislikes#9,comment_count#10,thumbnail_link#11,comments_disabled#12,ratings_disabled#13,video_error_or_removed#14,description#15] csv
only showing top 5 rows

>>> 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(5)
+-----+---+
| tag|total|
+-----+---+
| [none] | 3031|
| "2018" | 1399|
| "funny" | 1232|
| "comedy" | 1192|
| "deutsch" | 1122|
+-----+---+
only showing top 5 rows

>>> 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)
+-----+---+
| tag|total|
+-----+---+
| [none] | 3031|
| "2018" | 1399|
| "funny" | 1232|
| "comedy" | 1192|
| "deutsch" | 1122|
| "TV" | 1064|
| "tv" | 911|
| "german" | 734|
| "2017" | 722|
| "news" | 708|
+-----+---+
only showing top 10 rows

>>> 
```

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)
```

The screenshot shows a terminal window within the Termius application. The terminal window has a dark background and displays the following Python code and its output:

```
+-- Project [video_id#0, trending_date#1, title#2, channel_title#3, category_id#4, publish_time#5, split(tags#6, [], -1) AS tags#33, views#7, likes#8, dislikes#9, comment_count#10, thumbnail_link#11, comments_disabled#12, ratings_disabled#13, video_error_or_removed#14, description#15]
+-- Relation[video_id#0,trending_date#1,title#2,channel_title#3,category_id#4,publish_time#5,tags#6,views#7,likes#8,dislikes#9,comment_count#10,thumbnail_link#11,comments_disabled#12,ratings_disabled#13,video_error_or_removed#14,description#15] csv

>>> spark.sql("SELECT channel_title, MAX(comment_count) as cmnts FROM table GROUP BY channel_title ORDER BY cmnts").show(5)
+-----+
|   channel_title|cmnts|
+-----+
| dass er ihn gest...| null|
| siegt die Begier...| null|
| они обнаружили null|
| um ihr das Reite...| null|
| ihr Sohn Moritz ...| null|
+-----+
only showing top 5 rows

>>> spark.sql("SELECT channel_title, MAX(comment_count) as cmnts FROM table GROUP BY channel_title ORDER BY cmnts DESC").show(5)
+-----+
|   channel_title| cmnts|
+-----+
| ibighit|1084435|
| YouTube Spotlight| 827755|
| Logan Paul Vlogs| 611327|
| Wylascom| 523850|
| Sơn Tùng M-TP Off...| 401470|
+-----+
only showing top 5 rows

>>> spark.sql("SELECT channel_title, MAX(comment_count) as cmnts FROM table GROUP BY channel_title ORDER BY cmnts DESC").show(10)
+-----+
|   channel_title| cmnts|
+-----+
| ibighit|1084435|
| YouTube Spotlight| 827755|
| Logan Paul Vlogs| 611327|
| Wylascom| 523850|
| Sơn Tùng M-TP Off...| 401470|
| Marvel Entertainment| 335920|
| AuronPlay| 315801|
| David Dobrik| 288888|
| ChildishGambinoEVO| 263984|
| The ACE Family| 188032|
+-----+
only showing top 10 rows

>>> 
```

Your comments (optional)