.._dataframe_final

FINISHED

Les dataframes en Spark: principes et application

Un dataframe, ce n'est jamais plus qu'une collection de **lignes** (du point de vu de Spark). Abstraitement, il s'agit de matrices by-dimensionnelles. Enfin, on peut analogiquement se les représenter comme les tables d'une BDD. Mais, plutôt que de s'intérogger sur ce que sont les dataframes, regardons comment: les créer, les manipuler, les sauvegarder sur HDFS.

NB: une partie de ce qui se passe sous le capot (lecture de l'HDFS, connexion avec Hive, etc) n'est pas explicite dans ce notebook. Cela provient du fait que Zeppelin vient avec sa propre `spark session` (et, par là même **spark context spark conf**) – i.e. à aucun moment nous n'avons besoin de marquer

```
spark  = SparkSession.builder
                     .master("ip")
                     .enableHiveSupport()
                     .getOrCreate()
```

comme nous aurions à le faire dans une vrai application spark (nous nous contentons de marquer `%spark2.pyspark`)

Création de dataframes “à la main”

Il existe de multiples manières de créer des dataframes en spark

mais commençons simple en les créant à la main à l'aide de la méthode `.createDataFrame`

Took 0 sec. Last updated by anonymous at August 22 2019, 6:45:28 AM.

```
%spark2.pyspark
from pyspark.sql import Row

l = [('Deluge', True, True, True, "GPL",True, "Python, C++"),
('BitTorrent', True, True, True, "proprietary", True, "C++"),
('BitComet', True, False, False, "proprietary", False, "C++"),
('Xtorrent', False, True, False, "proprietary", None, None),
('Vuze', True, True, True, "GPL", True, "Java, SWT"),
('Transmission', True, True, True, "GPL/MIT", True, "C, Objective-C"),
('LimeWire', True, True, True, "GPL", False, "Java")]

row_collection = sc.parallelize(l).map(lambda x: Row(name=x[0], run_win=x[1], ru

df = sqlContext.createDataFrame(row_collection)
```

SPARK JOB (http://sandbox-hdp.hortonworks.com:4040/jobs/job?id=315) FINISHED

Took 0 sec. Last updated by anonymous at August 22 2019, 6:45:29 AM.

...dataframe_final

```
%spark2.pyspark
help(df.toDF)
```

FINISHED

```
Help on method toDF in module pyspark.sql.dataframe:

toDF(self, *cols) method of pyspark.sql.dataframe.DataFrame instance
    Returns a new class:`DataFrame` that with new specified column names

    :param cols: list of new column names (string)

    >>> df.toDF('f1', 'f2').collect()
    [Row(f1=2, f2=u'Alice'), Row(f1=5, f2=u'Bob')]

Help on method toDF in module pyspark.sql.dataframe:

toDF(self, *cols) method of pyspark.sql.dataframe.DataFrame instance
    Returns a new class:`DataFrame` that with new specified column names

    :param cols: list of new column names (string)

    >>> df.toDF('f1', 'f2').collect()
```

Took 0 sec. Last updated by anonymous at August 22 2019, 6:45:29 AM.

FINISHED

Opérations de base sur les dataframes

ce qui va suivre rappelera un peu **pandas**

principales méthodes:

- `select` : selectionne une colonne => dataframe
- `filter` : selectionne les lignes satisfaisant la condition
- `withColumn` : rajoute une colonne au dataframe
- `where` : self explanatory (Attention toutefois, where prend des `col` en argument)

NB vous pouvez concevoir les Row de Spark comme des dictionnaires python – e.g. `r["champ"]` permet, pour la ligne `r`, de récupérer la valeur contenu dans la colonne `champ`

Took 0 sec. Last updated by anonymous at August 22 2019, 6:45:29 AM.

```
%spark2.pyspark
df.show()
df.printSchema()
```

SPARK JOBS FINISHED

```
+------------+--------------+-----------+------------+---------+-------+-------+
|ipv6_support|      language|    license|        name|run_linux|run_mac|run_win|
+------------+--------------+-----------+------------+---------+-------+-------+
```

```
|        true|   Python, C++|        GPL|      Deluge|     true|   true|   true|
|        true|           C++|proprietary|  BitTorrent|     true|   true|   true|
|       false|           C++|proprietary|    BitComet|    false|  false|   true|
|        null|          null|proprietary|    Xtorrent|    false|   true|  false|
|        true|     Java, SWT|        GPL|        Vuze|     true|   true|   true|
|        true|C, Objective-C|    GPL/MIT|Transmission|     true|   true|   true|
|       false|          Java|        GPL|    LimeWire|     true|   true|   true|
+------------+--------------+-----------+------------+---------+-------+-------+

root
 |-- ipv6_support: boolean (nullable = true)
 |-- language: string (nullable = true)
 |-- license: string (nullable = true)
 |-- name: string (nullable = true)
 |-- run_linux: boolean (nullable = true)
 |-- run_mac: boolean (nullable = true)
 |-- run_win: boolean (nullable = true)
```

Took 1 sec. Last updated by anonymous at August 22 2019, 6:45:30 AM.

FINISHED

Filtrage

Filtrer, c'est sélectionner des lignes satisfaisant n conditions. Il existe plusieurs manière de filtrer un dataframe

- à l'aide de la méthode `filter`
- à l'aide d'une requête sql adressée à une vue temporaire utilisant un `WHERE`

Selectionnons les clients bit torrent qui soient à la fois open source et supportent le protocole ipv6

Took 0 sec. Last updated by anonymous at August 22 2019, 6:45:30 AM.

```
%spark2.pyspark

# en pure pyspark
r = df.filter("ipv6_support").filter(df.license != "proprietary")
r.show()
```

SPARK JOBS FINISHED

```
+------------+--------------+-------+------------+---------+-------+-------+
|ipv6_support|      language|license|        name|run_linux|run_mac|run_win|
+------------+--------------+-------+------------+---------+-------+-------+
|        true|   Python, C++|    GPL|      Deluge|     true|   true|   true|
|        true|     Java, SWT|    GPL|        Vuze|     true|   true|   true|
|        true|C, Objective-C|GPL/MIT|Transmission|     true|   true|   true|
+------------+--------------+-------+------------+---------+-------+-------+

+------------+--------------+-------+------------+---------+-------+-------+
|ipv6_support|      language|license|        name|run_linux|run_mac|run_win|
+------------+--------------+-------+------------+---------+-------+-------+
|        true|   Python, C++|    GPL|      Deluge|     true|   true|   true|
```

```
|        true|     Java, SWT|    GPL|        Vuze|     true|   true|   true|
|        true|C, Objective-C|GPL/MIT|Transmission|     true|   true|   true|
+------------+--------------+-------+------------+---------+-------+-------+
```

..._dataframe_final

Took 1 sec. Last updated by anonymous at August 22 2019, 6:45:31 AM.

```
%spark2.pyspark

# en sql
df.createOrReplaceTempView("torrent_clients")

r = df.filter(df.license != "proprietary").filter(df.language.contains("C"))
r.show()
```

SPARK JOBS FINISHED

```
+------------+--------------+-------+------------+---------+-------+-------+
|ipv6_support|      language|license|        name|run_linux|run_mac|run_win|
+------------+--------------+-------+------------+---------+-------+-------+
|        true|   Python, C++|    GPL|      Deluge|     true|   true|   true|
|        true|C, Objective-C|GPL/MIT|Transmission|     true|   true|   true|
+------------+--------------+-------+------------+---------+-------+-------+

+------------+--------------+-------+------------+---------+-------+-------+
|ipv6_support|      language|license|        name|run_linux|run_mac|run_win|
+------------+--------------+-------+------------+---------+-------+-------+
|        true|   Python, C++|    GPL|      Deluge|     true|   true|   true|
|        true|C, Objective-C|GPL/MIT|Transmission|     true|   true|   true|
+------------+--------------+-------+------------+---------+-------+-------+
```

Took 0 sec. Last updated by anonymous at August 22 2019, 6:45:31 AM.

…_dataframe_final

```
%spark2.pyspark

# TODO: à l'aide de la méthode de votre choix, trouver le(s) client(s) open sour
df.createOrReplaceTempView("torrent_clients")
r = spark.sql("SELECT * FROM torrent_clients WHERE language LIKE '%C%' AND licen
r.show()
```

SPARK JOBS FINISHED

```
+------------+--------------+-------+------------+---------+-------+-------+
|ipv6_support|      language|license|        name|run_linux|run_mac|run_win|
+------------+--------------+-------+------------+---------+-------+-------+
|        true|   Python, C++|    GPL|      Deluge|     true|   true|   true|
|        true|C, Objective-C|GPL/MIT|Transmission|     true|   true|   true|
+------------+--------------+-------+------------+---------+-------+-------+

+------------+--------------+-------+------------+---------+-------+-------+
|ipv6_support|      language|license|        name|run_linux|run_mac|run_win|
+------------+--------------+-------+------------+---------+-------+-------+
|        true|   Python, C++|    GPL|      Deluge|     true|   true|   true|
|        true|C, Objective-C|GPL/MIT|Transmission|     true|   true|   true|
+------------+--------------+-------+------------+---------+-------+-------+
```

Took 1 sec. Last updated by anonymous at August 22 2019, 6:45:32 AM.

Pour la suite du tuto, nous allons travailler sur un dataset plus intéressant FINISHED

Création de RDD par la lecture de fichiers en HDFS

Nous allons nous intéresser à la consommation de contenus vidéo et tenter de comparer nos habitudes de français avec celles de nos voisins britanniques
https://www.kaggle.com/datasnaek/youtube-new (https://www.kaggle.com/datasnaek/youtube-new)

chargez les données dans l'HDFS en `/user/raj_ops/youtube`

Took 0 sec. Last updated by anonymous at August 22 2019, 6:45:33 AM.

```
%spark2.pyspark
from pyspark.sql.types import *
from pyspark.sql.functions import *

# read n'infère pas les types. Il faut les donner
schema = StructType([
        StructField("video_id", StringType(), True),
        StructField("trending_date", StringType(), True),
        StructField("title", StringType(), True),
        StructField("channel_title", StringType(), True),
        StructField("category_id", IntegerType(), True),
        StructField("publish_time", TimestampType(), True),
        StructField("tags", StringType(), True),
        StructField("views", FloatType(), True),
```

FINISHED

..._dataframe_final

Took 0 sec. Last updated by anonymous at August 22 2019, 6:45:33 AM.

```
%spark2.pyspark
```

SPARK JOB (http://sandbox-hdp.hortonworks.com:4040/jobs/job?id=324) FINISHED

```
frdf.show()
```

```
+-----------+-------------+--------------------+--------------------+-----------+
--------------------+--------------------+---------+--------+--------+-----------
-+--------------------+-----------------+----------------+----------------------+
--------------------+
|   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|
+-----------+-------------+--------------------+--------------------+-----------+
--------------------+--------------------+---------+--------+--------+-----------
-+--------------------+-----------------+----------------+----------------------+
--------------------+
|Ro6eob0LrCY|     17.14.11|Malika LePen : Fe...| Le Raptor Dissident|         24|
2017-11-13 17:32:55|"Raptor""|""Dissi...| 212702.0| 29282.0|  1108.0|       3817.
0|https://i.ytimg.c...|            false|           false|                 false|
Dimanche.\n18h30....|
|Yo84eqYwP98|     17.14.11|LA PIRE PARTIE ft...|             Le Labo|         24|
2017-11-12 15:00:02|              [none]| 432721.0| 14053.0|   576.0|       1161.
0|https://i.ytimg.c...|            false|           false|                 false|
Le jeu de société...|
|ceqntSXE-10|     17.14.11|DESSINS ANIMÉS F...|     Daniil le Russe|         23|2
017-11-13 17:00:38|"cartoon""|""poké...| 482153.0| 76203.0|   477.0|       9580.0
|https://i.ytimg.c...|            false|           false|                 false|U
ne nouvelle dose...|
|WuTFI5qftCE|     17.14.11|PAPY GRENIER - ME...|   Joueur Du Grenier|         20|
2017-11-12 17:00:02|"Papy grenier""|"...| 925222.0| 85016.0|   550.0|       4303.
0|https://i.ytimg.c...|            false|           false|                 false|
Nouvel ,épisode d...|
|ee6OFs8TdEg|     17.14.11|QUI SAUTERA LE PL...|   Aurelien Fontenoy|         17|
2017-11-13 16:30:03|"vélo""|""vtt""|"...| 141695.0|  8091.0|    72.0|        481.
0|https://i.ytimg.c...|            false|           false|                 false|
Sauts à plus de 4...|
|teXaL6GdQRk|     17.14.11|STRANGER JOKES : ...|Le Jeu, C'est Sér...|         23|
2017-11-13 15:48:57|"Stranger Jokes""...| 141253.0| 14354.0|   202.0|        417.
0|https://i.ytimg.c...|            false|           false|                 false|
Contenu commandit...|
|nduL7G_gJoY|     17.14.11|De retour dans le...|         silent jill|         24|
```

...dataframe_final

```
2017-11-12 19:00:08|"fantome""|""espr...| 187654.0|  9286.0|  1381.0|       2419.
0|https://i.ytimg.c...|           false|           false|               false|
Bonsoir à tous, \...|
|QmpWE_SODZA|     17.14.11|T'es qui toi ? Sq...|Salut les terriens !|         24|
2017-11-12 17:30:01|"salut les terrie...|  91051.0|  1674.0|  1903.0|        701.
0|https://i.ytimg.c...|           false|           false|               false|
Dans Salut Les Te...|
|GBVxEpQr8R8|     17.14.11|ON VOUS DÉVOILE N...|     Mcfly & Carlito|         24|
2017-11-12 08:59:25|"mcfly""|""carlit...|2340941.0|200598.0|  6018.0|       7575.
0|https://i.ytimg.c...|           false|           false|               false|
Nouvelle vidéo to...|
|0RFhWyM6qbA|     17.14.11|Benzema balance s...|            HALIRIPA|         22|
2017-11-12 20:16:45|"Karim benzema fo...| 635236.0|  5945.0|   722.0|       1483.
0|https://i.ytimg.c...|           false|           false|               false|
Abonnez-vous et p...|
|LhjhGOBIoHM|     17.14.11|Jérémy Ferrari - ...| On n'est pas couché|         24|
2017-11-12 00:53:02|"onpc""|""on n'es...| 294065.0|     0.0|     0.0|          0.
0|https://i.ytimg.c...|            true|            true|               false|
Jérémy Ferrari - ...|
|lnFaRuWOLN0|     17.14.11|Emilie, 10 ans, C...|            Lolywood|         23|
2017-11-12 15:00:03|"Chope squad""|""...|1504950.0|108635.0|  1562.0|       2913.
0|https://i.ytimg.c...|           false|           false|               false|
Tu sais que ça va...|
|LGbUBietbJc|     17.14.11|J'ABANDONNE UNE P...|               RAZMO|         24|
2017-11-12 18:00:01|              [none]| 335121.0| 23410.0|  2034.0|       3004.
0|https://i.ytimg.c...|           false|           false|               false|
GO 500K !!! JE CO...|
|JfanclGyrZY|     17.14.11|L'incroyable Joya...|           Nota Bene|         27|
2017-11-13 11:00:23|"Nota Bene""|""fr...|  53248.0|  5164.0|    43.0|        234.
0|https://i.ytimg.c...|           false|           false|               false|
INCROYABLE ! On e...|
|PpECwr15oQQ|     17.14.11|  Cocovoit - M comme|            Cocovoit|         23|
2017-11-13 17:16:37|"cocovoit""|""et ...|  55313.0|  4153.0|    97.0|        396.
0|https://i.ytimg.c...|           false|           false|               false|
Pour tout savoir ...|
|ZTTpRHC5ZH4|     17.14.11|Une nouvelle lune...|            Studio 4|         23|
2017-11-12 14:00:02|"nouvelles écritu...| 172608.0|  8163.0|    95.0|        237.
0|https://i.ytimg.c...|           false|           false|               false|
Vous avez lu la B...|
|RpEJCznFreQ|     17.14.11|FAIRE UNE GLACE R...|    Alexandre Calvez|         26|
2017-11-12 18:00:02|"Alexandre Calvez...| 373013.0| 23645.0|   411.0|        813.
0|https://i.ytimg.c...|           false|           false|               false|
👉 Ma page facebo...|
|tsMw-VMUtNU|     17.14.11|Kid Barely Avoids...|            Pirateay|         24|
2017-11-11 18:38:02|"near accident""|...|  79611.0|    56.0|     4.0|         25.
0|https://i.ytimg.c...|           false|           false|               false|
A young kid barel...|
|hdOzPPa_bFY|     17.14.11|JE SUIS PASSÉ À L...|            Panormal|         20|
2017-11-13 02:49:46|              [none]| 255349.0| 20069.0|   245.0|        950.
0|https://i.ytimg.c...|           false|           false|               false|
LISEZ LA DESCRIPT...|
|7779JdxVAg0|     17.14.11|10 FILMS que LES ...|          Lama Faché|         24|
2017-11-12 15:59:38|"dessin animé""|"...| 653398.0| 27773.0|  2778.0|       3976.
0|https://i.ytimg.c...|           false|           false|               false|
Si vous pensiez q...|
```

```
+-----------+-------------+--------------------+--------------------+-----------+
------------------+--------------------+---------+--------+--------+-----------
-+--------------------+-----------------+----------------+----------------------+
--------------------+
only showing top 20 rows
```

..._dataframe_final

```
+-----------+-------------+--------------------+--------------------+-----------+
------------------+--------------------+---------+--------+--------+-----------
-+--------------------+-----------------+----------------+----------------------+
--------------------+
|   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|
+-----------+-------------+--------------------+--------------------+-----------+
------------------+--------------------+---------+--------+--------+-----------
-+--------------------+-----------------+----------------+----------------------+
--------------------+
|Ro6eob0LrCY|     17.14.11|Malika LePen : Fe...| Le Raptor Dissident|         24|
2017-11-13 17:32:55|"Raptor""|""Dissi...| 212702.0| 29282.0|  1108.0|       3817.
0|https://i.ytimg.c...|            false|           false|                false|
Dimanche.\n18h30....|
|Yo84eqYwP98|     17.14.11|LA PIRE PARTIE ft...|             Le Labo|         24|
2017-11-12 15:00:02|              [none]| 432721.0| 14053.0|   576.0|       1161.
0|https://i.ytimg.c...|            false|           false|                false|
Le jeu de société...|
|ceqntSXE-10|     17.14.11|DESSINS ANIMÉS F...|     Daniil le Russe|         23|2
017-11-13 17:00:38|"cartoon""|""poké...| 482153.0| 76203.0|   477.0|       9580.0
|https://i.ytimg.c...|            false|           false|                false|U
ne nouvelle dose...|
|WuTFI5qftCE|     17.14.11|PAPY GRENIER - ME...|   Joueur Du Grenier|         20|
2017-11-12 17:00:02|"Papy grenier""|"...| 925222.0| 85016.0|   550.0|       4303.
0|https://i.ytimg.c...|            false|           false|                false|
Nouvel ,épisode d...|
|ee6OFs8TdEg|     17.14.11|QUI SAUTERA LE PL...|   Aurelien Fontenoy|         17|
2017-11-13 16:30:03|"vélo""|""vtt""|"...| 141695.0|  8091.0|    72.0|       481.
0|https://i.ytimg.c...|            false|           false|                false|
Sauts à plus de 4...|
|teXaL6GdQRk|     17.14.11|STRANGER JOKES : ...|Le Jeu, C'est Sér...|         23|
2017-11-13 15:48:57|"Stranger Jokes""...| 141253.0| 14354.0|   202.0|       417.
0|https://i.ytimg.c...|            false|           false|                false|
Contenu commandit...|
|nduL7G_gJoY|     17.14.11|De retour dans le...|         silent jill|         24|
2017-11-12 19:00:08|"fantome""|""espr...| 187654.0|  9286.0|  1381.0|       2419.
0|https://i.ytimg.c...|            false|           false|                false|
Bonsoir à tous, \...|
|QmpWE_SODZA|     17.14.11|T'es qui toi ? Sq...|Salut les terriens !|         24|
2017-11-13 17:30:01|"salut les terrie...|  91051.0|  1674.0|  1903.0|        701.
0|https://i.ytimg.c...|            false|           false|                false|
Dans Salut Les Te...|
|GBVxEpQr8R8|     17.14.11|ON VOUS DÉVOILE N...|     Mcfly & Carlito|         24|
2017-11-12 08:59:25|"mcfly""|""carlit...|2340941.0|200598.0|  6018.0|       7575.
0|https://i.ytimg.c...|            false|           false|                false|
Nouvelle vidéo to...|
|0RFhWyM6qbA|     17.14.11|Benzema balance s...|            HALIRIPA|         22|
```

```
2017-11-12 20:16:45|"Karim benzema fo...| 635236.0|  5945.0|   722.0|       1483.
0|https://i.ytimg.c...|            false|           false|                 false|
Abonnez-vous et p...|
|LhjhGOBIoHM|     17.14.11|Jérémy Ferrari - ...| On n'est pas couché|         24|
2017-11-12 00:23:02|"Enpc | On n'es...| 294065.0|     0.0|     0.0|          0.
0|https://i.ytimg.c...|             true|            true|                 false|
Jérémy Ferrari - ...|
|lnFaRuWOLN0|     17.14.11|Emilie, 10 ans, C...|            Lolywood|         23|
2017-11-12 15:00:03|"Chope squad""|""...|1504950.0|108635.0|  1562.0|       2913.
0|https://i.ytimg.c...|            false|           false|                 false|
Tu sais que ça va...|
|LGbUBietbJc|     17.14.11|J'ABANDONNE UNE P...|               RAZMO|         24|
2017-11-12 18:00:01|              [none]| 335121.0| 23410.0|  2034.0|       3004.
0|https://i.ytimg.c...|            false|           false|                 false|
GO 500K !!! JE CO...|
|JfanclGyrZY|     17.14.11|L'incroyable Joya...|           Nota Bene|         27|
2017-11-13 11:00:23|"Nota Bene""|""fr...|  53248.0|  5164.0|    43.0|        234.
0|https://i.ytimg.c...|            false|           false|                 false|
```


Took 0 sec. Last updated by anonymous at August 22 2019, 6:45:33 AM.

```
%spark2.pyspark

frdf.printSchema()
```

FINISHED

```
root
 |-- video_id: string (nullable = true)
 |-- trending_date: string (nullable = true)
 |-- title: string (nullable = true)
 |-- channel_title: string (nullable = true)
 |-- category_id: integer (nullable = true)
 |-- publish_time: timestamp (nullable = true)
 |-- tags: string (nullable = true)
 |-- views: float (nullable = true)
 |-- likes: float (nullable = true)
 |-- dislikes: float (nullable = true)
 |-- comment_count: float (nullable = true)
 |-- thumbnail_link: string (nullable = true)
 |-- comments_disabled: boolean (nullable = true)
 |-- ratings_disabled: boolean (nullable = true)
 |-- video_error_or_removed: boolean (nullable = true)
 |-- description: string (nullable = true)

root
 |-- video_id: string (nullable = true)
 |-- trending_date: string (nullable = true)
 |-- title: string (nullable = true)
 |-- channel_title: string (nullable = true)
 |-- category_id: integer (nullable = true)
 |-- publish_time: timestamp (nullable = true)
 |-- tags: string (nullable = true)
 |-- views: float (nullable = true)
 |-- likes: float (nullable = true)
 |-- dislikes: float (nullable = true)
 |-- comment_count: float (nullable = true)
```

```
 |-- thumbnail_link: string (nullable = true)
 |-- comments_disabled: boolean (nullable = true)
 |-- ratings_disabled: boolean (nullable = true)
 |-- video_error_or_removed: boolean (nullable = true)
 |-- description: string (nullable = true)
```

...dataframe_final

Took 0 sec. Last updated by anonymous at August 22 2019, 6:45:33 AM.

```
%spark2.pyspark

#Assurons nous de la taille de nos data frames

assert frdf.count() == 46138
assert gbdf.count() == 43295
```

SPARK JOBS FINISHED

Took 1 sec. Last updated by anonymous at August 22 2019, 6:45:35 AM.

FINISHED

Dataframe merge: on ne peut comparer que le disemblable

Took 0 sec. Last updated by anonymous at August 22 2019, 6:45:35 AM.

```
%spark2.pyspark

df = frdf.withColumn("country", lit("fr")).union(gbdf.withColumn("country", lit(
df.count()
#df.show()
```

SPARK JOB (http://sandbox-hdp.hortonworks.com:4040/jobs/job?id=327) FINISHED

```
89433
89433
```

Took 1 sec. Last updated by anonymous at August 22 2019, 6:45:36 AM.

FINISHED

Quelques statistiques descriptives et visualisation

Si tout s'est bien passé jusque là, passons à l'élément central: l'analyse. Pour le graphing, nous allons nous servir de Seaborn (https://seaborn.pydata.org/) (une surcouche matplotlib aussi belle que concise). Son installation va nécessiter quelques manip du fait que Hortonworks tourne avec un version bientôt obsolète de python

Veuillez d'abord l'installer à l'aide de la commande suivante. En `root` dans shell in a box:

```
yum -y install python-pip
yum -y install gcc

pip install --upgrade pip
pip install --upgrade setuptools

pip install pyarrow==0.8.0

pip install matplotlib==1.5.3
pip install seaborn

# doit donner
Successfully .................. seaborn-0.9.0
```

Visualisation

nécessite de convertir le dataframe en un df pandas à l'aide de la méthode `toPandas()`. Ne choisir que les colonnes pertinentes

Statisitique

Les faire directement en Spark pour tirer parti de la parallelisation. La grande majorité des fonctions statistiques pertinentes sont documentés ici https://spark.apache.org/docs/latest/mllib-statistics.html (https://spark.apache.org/docs/latest/mllib-statistics.html)

Took 0 sec. Last updated by anonymous at August 22 2019, 6:45:37 AM.

```
%spark2.pyspark
import matplotlib
matplotlib.use("Agg") # IMPORTANT: ne surtout pas toucher à cela
import matplotlib.pyplot as plt
import seaborn as sns
import pandas as pd

spark.conf.set("spark.sql.execution.arrow.enabled", "true")
sns.set(style="whitegrid")
```

FINISHED

Took 0 sec. Last updated by anonymous at August 22 2019, 6:45:37 AM.

```
%spark2.pyspark
# extraire les colonnes des dataframes et les convertir en df pandas
dfp = df.select("views", "likes", "dislikes", "country").toPandas()
```

SPARK JOB (http://sandbox-hdp.hortonworks.com:4040/jobs/job?id=328) FINISHED

Took 1 sec. Last updated by anonymous at August 22 2019, 6:45:39 AM.

```
%spark2.pyspark
df.describe()
```

SPARK JOB (http://sandbox-hdp.hortonworks.com:4040/jobs/job?id=329) FINISHED

```
DataFrame[summary: string, video_id: string, trending_date: string, title: string
```

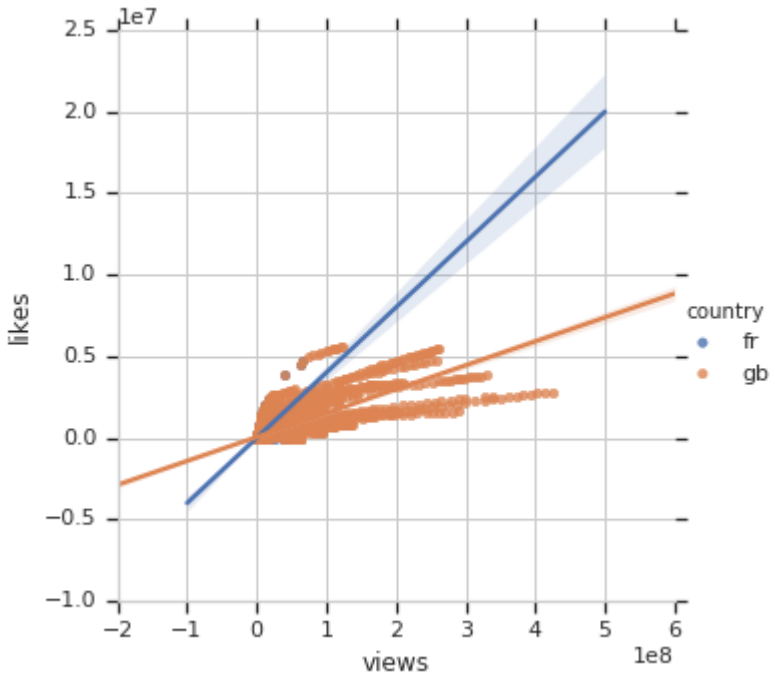
```
, channel_title: string, category_id: string, tags: string, views: string, likes:
string, dislikes: string, comment_count: string, thumbnail_link: string, descript
ion: string, country: string]
DataFrame[summary: string, video_id: string, trending_date: string, title: string
, channel_title: string, category_id: string, tags: string, views: string, likes:
string, dislikes: string, comment_count: string, thumbnail_link: string, descript
ion: string, country: string]
```

...._dataframe_final

Took 3 sec. Last updated by anonymous at August 22 2019, 6:45:43 AM.

```
%spark2.pyspark                                                    FINISHED
# a priori, il existe une importante différence entre les deux pays mais i) peut
sns.lmplot(x="views", y="likes", hue="country", data=dfp)
```

```
<seaborn.axisgrid.FacetGrid object at 0x7fdec54b6810>
<seaborn.axisgrid.FacetGrid object at 0x7fdec54b6810>
```



Took 36 sec. Last updated by anonymous at August 22 2019, 6:46:20 AM.

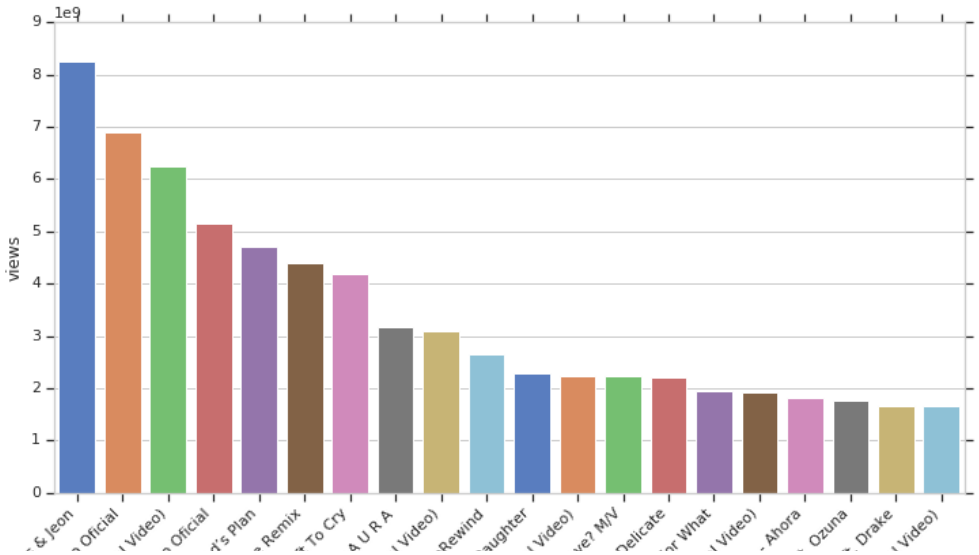
```
%spark2.pyspark SPARK JOB (http://sandbox-hdp.hortonworks.com:4040/jobs/job?id=330)  FINISHED
# top 20 des vidéos les plus vues (vous noterez qu'il y a un problème: le compre
data = df.dropna(subset=["title", "views"]).select("video_id","title", "views").
#help(df.limit)
fig, ax = plt.subplots(figsize = (12,6))
sns.barplot(x="title", y="views", data=data, palette="muted", ax=ax)

ax.set_xticklabels(labels=data["title"].values, rotation=45, ha='right')
```

```
[<matplotlib.text.Text object at 0x7fdec62d6690>, <matplotlib.text.Text object at
0x7fdec6102710>, <matplotlib.text.Text object at 0x7fdec6493c50>, <matplotlib.tex
t.Text object at 0x7fdec64a0dd0>, <matplotlib.text.Text object at 0x7fdec64a0290>
, <matplotlib.text.Text object at 0x7fdec6371dd0>, <matplotlib.text.Text object a
```

```
t 0x7fdec623c7d0>, <matplotlib.text.Text object at 0x7fdec4db3550>, <matplotlib.t
ext.Text object at 0x7fdec4db3dd0>, <matplotlib.text.Text object at 0x7fdecc1d9ad
0>, <matplotlib.text.Text object at 0x7fdecc47d1d0>, <matplotlib.text.Text object
at 0x7fdecc47d4d0>, <matplotlib.text.Text object at 0x7fdecc2e6cd0>, <matplotlib.
text.Text object at 0x7fdecc2e6350>, <matplotlib.text.Text object at 0x7fdecf03b3
50>, <matplotlib.text.Text object at 0x7fdecf03b490>, <matplotlib.text.Text objec
t at 0x7fdeceb357d0>, <matplotlib.text.Text object at 0x7fdeceb35e10>, <matplotli
b.text.Text object at 0x7fdec62aff50>, <matplotlib.text.Text object at 0x7fdec62a
f090>]
[<matplotlib.text.Text object at 0x7fdec62d6690>, <matplotlib.text.Text object at
0x7fdec6102710>, <matplotlib.text.Text object at 0x7fdec6493c50>, <matplotlib.tex
t.Text object at 0x7fdec64a0dd0>, <matplotlib.text.Text object at 0x7fdec64a0290>
, <matplotlib.text.Text object at 0x7fdec6371dd0>, <matplotlib.text.Text object a
t 0x7fdec623c7d0>, <matplotlib.text.Text object at 0x7fdec4db3550>, <matplotlib.t
ext.Text object at 0x7fdec4db3dd0>, <matplotlib.text.Text object at 0x7fdecc1d9ad
0>, <matplotlib.text.Text object at 0x7fdecc47d1d0>, <matplotlib.text.Text object
at 0x7fdecc47d4d0>, <matplotlib.text.Text object at 0x7fdecc2e6cd0>, <matplotlib.
```

..._dataframe_final



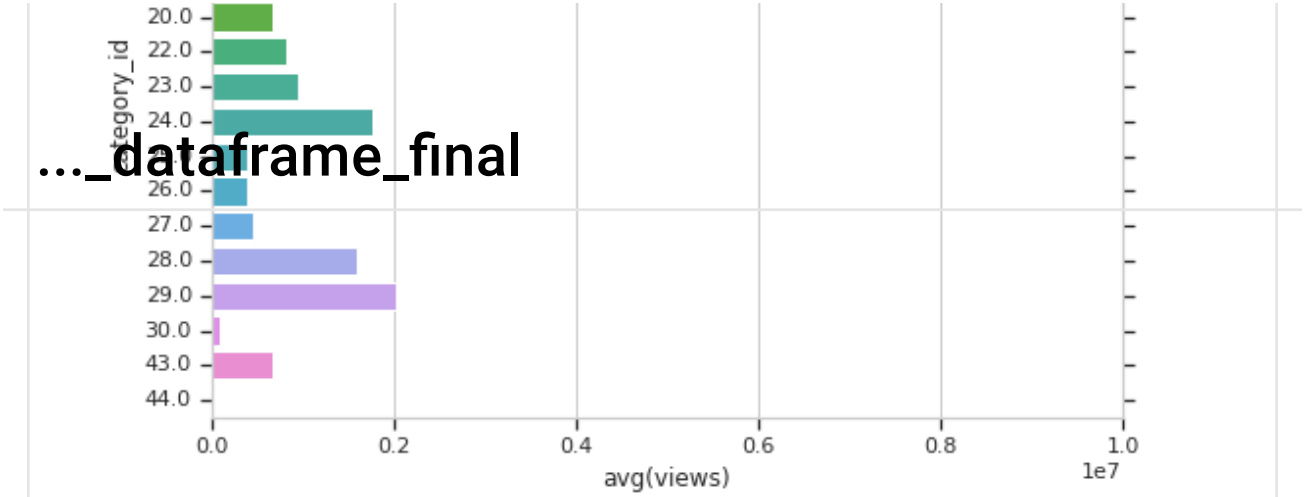
Took 10 sec. Last updated by anonymous at August 22 2019, 6:46:30 AM.

```
%spark2.pyspark
# aggrégé par categories / moyenne

sns.barplot(x = "avg(views)", y="category_id" , data=df.groupBy("category_id").a
```

SPARK JOB (http://sandbox-hdp.hortonworks.com:4040/jobs/job?id=331) FINISHED

```
<matplotlib.axes._subplots.AxesSubplot object at 0x7fdec4a90c10>
<matplotlib.axes._subplots.AxesSubplot object at 0x7fdec4a90c10>
```



Took 5 sec. Last updated by anonymous at August 22 2019, 6:46:35 AM.

FINISHED

Une ANOVA?

Sachant que nous disposons de deux variables catégorielles at que nous avons montré que l'indépendance statisitque de la distribution des vidéos entre les différentes catégories au sein des deux groupes, une two way ANOVA (country * category) semblerait une bonne approche. Les variables pertinentes

- views
- comment count
- likes
- dislikes

NB: l'ANOVA stipule que la variable explicative (dans notre cas, les likes) doit être continue. Tel n'est pas le cas des likes. De plus, il convient d'abord de s'assurer que la distribution suit une loi normale. Nous allons:

- Grapher les distribution
- réaliser un test de Kolmogorov-Smirnov

Le test de Shapiro-Wilk ne semble pas exister en spark. Nous devrions pouvoir lui substituer le test de Kolmogorov-Smirnov (https://fr.wikipedia.org/wiki/Test_de_Kolmogorov-Smirnov) documentation (https://spark.apache.org/docs/2.3.0/api/java/org/apache/spark/mllib/stat/test/KolmogorovSmirnovTest.html)

Approche visuelle

Took 0 sec. Last updated by anonymous at August 22 2019, 6:46:35 AM.

```
%spark2.pyspark
df_range = df.select("views", "comment_count", "likes", "dislikes")#.withColumn(
df_range.describe().show()
```

SPARK JOB (http://sandbox-hdp.hortonworks.com:4040/jobs/job?id=332) FINISHED

```
+-------+------------------+-----------------+-----------------+-------------
----+
|summary|             views|    comment_count|            likes|        disl
ikes|
+-------+------------------+-----------------+-----------------+-------------
----+
|  count|             79640|            79640|            79640|            7
9640|
|   mean| 3103592.3616147665|7332.636300853842| 74624.6488448016|4136.60110497
2376|
| stddev|1.3622226102699319E7|37241.603765286716|259175.35148122633|36697.6566676
1799|
|    min|             223.0|              0.0|              0.0|
0.0|
|    max|       4.24538912E8|         1626501.0|         5613827.0|        19449
71.0|
+-------+------------------+-----------------+-----------------+-------------
```

Took 2 sec. Last updated by anonymous at August 22 2019, 6:46:38 AM.

```
%spark2.pyspark
df_range = df.select("views", "comment_count", "likes", "dislikes")#.withColumn(

df_range.describe().show()
```

SPARK JOB (http://sandbox-hdp.hortonworks.com:4040/jobs/job?id=333) FINISHED

```
+-------+------------------+-----------------+-----------------+-------------
---+
|summary|             views|    comment_count|            likes|        dislik
es|
+-------+------------------+-----------------+-----------------+-------------
---+
|  count|             79640|            79640|            79640|           796
40|
|   mean| 3103592.3616147665|7332.636300853842| 74624.6488448016|4136.601104972
376|
| stddev|1.3622226102699319E7|37241.60376528672|259175.35148122633|36697.65666761
799|
|    min|             223.0|              0.0|              0.0|
0.0|
|    max|       4.24538912E8|        1626501.0|         5613827.0|       194497
1.0|
+-------+------------------+-----------------+-----------------+-------------
---+

+-------+------------------+-----------------+-----------------+-------------
---+
|summary|             views|    comment_count|            likes|        dislik
es|
+-------+------------------+-----------------+-----------------+-------------
---+
|  count|             79640|            79640|            79640|           796
40|
```

```
|   mean|  3103592.3616147665|7332.636300853842|  74624.6488448016|4136.601104972
376|
| stddev|1.3622226102699319E7|37241.60376528672|259175.35148122633|36697.65666761
799|
|    min|                 0.0|              0.0|               0.0|
0.0|
|    max|        4.24538912E8|        1626501.0|         5613827.0|         194497
1.0|
+-------+--------------------+-----------------+------------------+-------------
---+
```

...._dataframe_final

Took 2 sec. Last updated by anonymous at August 22 2019, 6:46:40 AM.

```
%spark2.pyspark

# TROUVEZ UN FONCTION QUI PERMET D AGGREGER VIEWS EN 20 VALEURS
def get_interval(colonne_name, nb):
    return int(df.select(colonne_name).rdd.max()[colonne_name]/20)

views_range = df.select("views").dropna(subset=["views"]).rdd.map(lambda x: x["v
comment_count_range = df.select("comment_count").dropna(subset=["comment_count"]
likes_range = df.select("likes").dropna(subset=["likes"]).rdd.map(lambda x: x["l
dislikes_range = df.select("dislikes").dropna(subset=["dislikes"]).rdd.map(lambd

data_views = pd.DataFrame.from_dict(views_range.countByValue(),orient='index')
data_comments = pd.DataFrame.from_dict(comment_count_range.countByValue(),orient
data_likes = pd.DataFrame.from_dict(likes_range.countByValue(),orient='index')
data_dislikes = pd.DataFrame.from_dict(dislikes_range.countByValue(),orient='ind
```

SPARK JOBS FINISHED

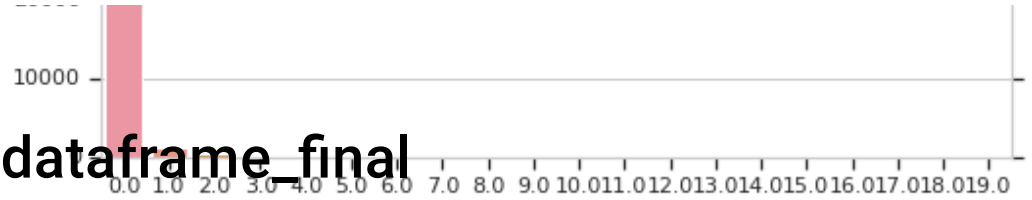
Took 4 sec. Last updated by anonymous at August 22 2019, 6:47:15 AM.

```
%spark2.pyspark

sns.barplot(x=data_views.index , y=data_views[0], data=data_views)
```

FINISHED

```
<matplotlib.axes._subplots.AxesSubplot object at 0x7fdec44b4bd0>
<matplotlib.axes._subplots.AxesSubplot object at 0x7fdec44b4bd0>
```



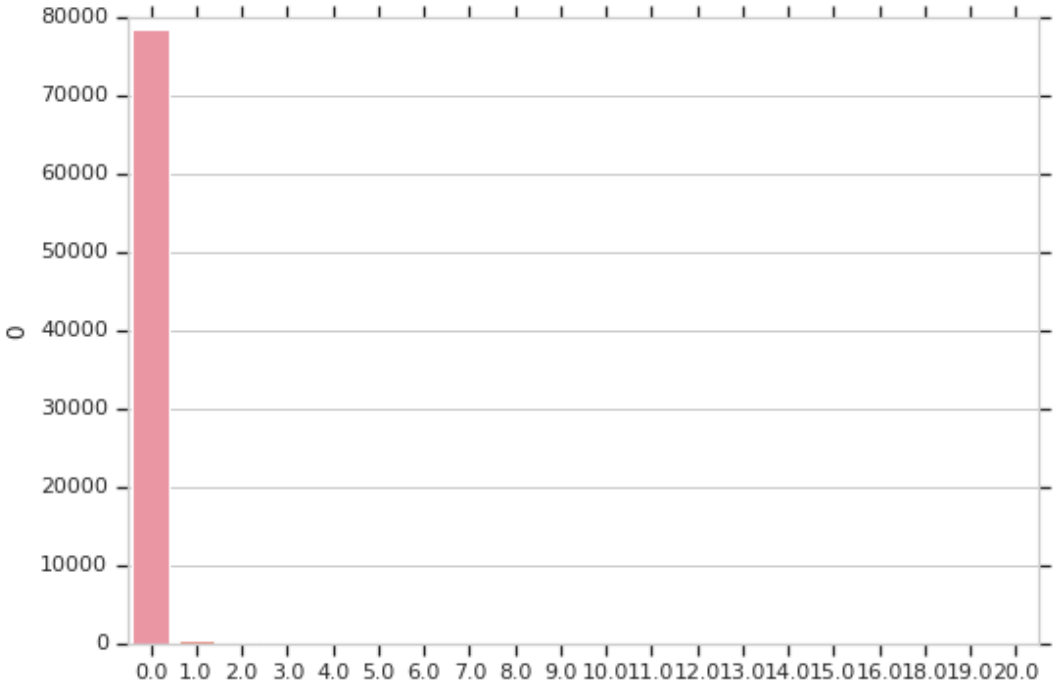
..._dataframe_final

Took 4 sec. Last updated by anonymous at August 22 2019, 6:47:22 AM.

```
%spark2.pyspark
sns.barplot(x=data_comments.index , y=data_comments[0], data=data_comments)
```

FINISHED

```
<matplotlib.axes._subplots.AxesSubplot object at 0x7fdec6ba26d0>
<matplotlib.axes._subplots.AxesSubplot object at 0x7fdec6ba26d0>
```



Took 4 sec. Last updated by anonymous at August 22 2019, 6:47:29 AM.

FINISHED

Test de Kolmogorov

Took 0 sec. Last updated by anonymous at August 22 2019, 6:35:21 AM.

```
%spark2.pyspark
from pyspark.mllib.stat import Statistics
# je veux tester si "views" suit une loi normale. Que dois-je ajouter?


result = Statistics.kolmogorovSmirnovTest(df.select("views").rdd.map(lambda r: r
print(result)
```

SPARK JOBS FINISHED

```
Kolmogorov-Smirnov test summary:
degrees of freedom = 0
statistic = 0.8[illegible]
pValue = 1.236532209958341E-9
Very strong presumption against null hypothesis: Sample follows theoretical distr
ibution.
```

Took 3 sec. Last updated by anonymous at August 22 2019, 6:47:59 AM.

FINISHED

Je sais, “*Sample follows theoretical distribution*” prète à confusion. Mais rejet de H0, dans ce cas, rejet de l'hypothèse que les données suivent une loi normale

https://stackoverflow.com/questions/37335408/kolmogorov-smirnov-test-in-spark-python-not-working (https://stackoverflow.com/questions/37335408/kolmogorov-smirnov-test-in-spark-python-not-working)

Nous ne pouvons pas, *a priori* nous servir d'une ANOVA

Took 0 sec. Last updated by anonymous at August 22 2019, 6:38:16 AM.

FINISHED

Corrélation des variables numériques

Took 0 sec. Last updated by anonymous at August 22 2019, 6:38:04 AM.

SPARK JOBS FINISHED

```
%spark2.pyspark
# Utilisation de l'ancienne librairie (mllib)
#Existe-t-il globalement une corrélation entre le nombre de vue et les likes?

df_filtered = df.dropna(subset=["views", "likes"]).where((col("views") != 0.0) &

# ouh que c'est vilain!!! Pourquoi devons nous faire ce map?
views = df_filtered.select("views").rdd.map(lambda r: r["views"])
likes = df_filtered.select("likes").rdd.map(lambda r: r["likes"])

corr = Statistics.corr(views, likes, method="pearson")
print corr
```

```
0.79992349445
```

Took 3 sec. Last updated by anonymous at August 22 2019, 6:48:12 AM.

SPARK JOBS FINISHED

```
%spark2.pyspark
from pyspark.ml.stat import Correlation
from pyspark.ml.feature import VectorAssembler
# CONTEXT/ migration de mllib => ml . recalculer la corrélation de Pearson entre

df_filtered = df.dropna(subset=["views", "likes", "comment_count", "dislikes"])

features = ["views", "comment_count", "likes", "dislikes"]
```

..._dataframe_final

```
<matplotlib.axes._subplots.AxesSubplot object at 0x7fdecc20af50>
```

Took 5 sec. Last updated by anonymous at August 22 2019, 6:48:22 AM.

FINISHED

Test d'indépendance

Les catégories expliquent-elles la corrélation observée entre `views` et `comment_count`

Took 0 sec. Last updated by anonymous at August 22 2019, 6:44:04 AM.

SPARK JOBS FINISHED

```
%spark2.pyspark
from pyspark.ml.stat import ChiSquareTest

df_filtered = df.dropna(subset=["views", "likes", "comment_count", "dislikes"]).

x = df_filtered.groupBy("category_id").agg(mean('views').alias('views'), mean('c
vecAssembler = VectorAssembler(inputCols=["views", "comment_count"], outputCol="
x = vecAssembler.transform(x)

r = ChiSquareTest.test(x, "features", "category_id").head()
print "pValues: " + str(r.pValues)
print "degreesOfFreedom: " + str(r.degreesOfFreedom)
```

```
pValues: [0.2350832814454744,0.2350832814454744]
degreesOfFreedom: [256, 256]
statistics: [272.0000000000008,272.0000000000008]
```

Took 6 sec. Last updated by anonymous at August 22 2019, 6:48:43 AM.

...dataframe_final

FINISHED

Tentons une softmax regression

https://spark.apache.org/docs/2.2.0/api/python/pyspark.ml.html#pyspark.ml.classification.LogisticRegression (https://spark.apache.org/docs/2.2.0/api/python/pyspark.ml.html#pyspark.ml.classification.LogisticRegression)

Attention, contrairement à ce que laisserait penser penser le nom `LogisticRegression` , le fait d'ajouter `, family="multinomial"` en fait une softmax

Took 0 sec. Last updated by anonymous at August 22 2019, 6:41:18 AM.

SPARK JOBS FINISHED

```
%spark2.pyspark
from pyspark.ml.classification import LogisticRegression
from pyspark.ml.feature import StringIndexer

indexer = StringIndexer(inputCol="category_id", outputCol="label")
x = indexer.fit(df_filtered).transform(df_filtered)


vecAssembler = VectorAssembler(inputCols=["comment_count","likes"], outputCol="f
x = vecAssembler.transform(x).select("label", "features")
#x.show()
lr = LogisticRegression(maxIter=50, regParam=0.3, elasticNetParam=0.8, family="m

lrModel = lr.fit(x)

#dir(lrModel)
#lrModel.standardization()
print("Coefficients: " )
print(lrModel.coefficientMatrix.toArray())
print("Intercept: " + str(lrModel.interceptVector))
# pas probant
```

```
Coefficients:
[[ 1.64535827e-05  2.36786823e-06]
 [-3.62170625e-05  1.34561121e-05]
 [ 2.24810162e-05 -4.01657832e-06]
 [-7.52205378e-06  6.06715405e-06]
 [ 2.05072004e-05 -3.98254201e-06]
 [ 4.97460320e-05 -3.07441563e-05]
 [ 5.29469144e-06  3.74231383e-06]
 [ 2.47347196e-05 -7.01969179e-06]
```

```
 [ 8.83035972e-07  2.40435819e-06]
 [ 1.74916407e-05  1.67117756e-06]
 [-6.29475777e-06  8.42316927e-07]
 [ 2.36396260e-05 -3.64640346e-05]
 [-1.63e-06  2.04e-07]
 [-2.83865275e-05 -1.51334559e-05]
 [ 1.94941799e-05  2.29948490e-06]
 [-3.02653658e-05 -1.18207765e-05]
 [-3.39527292e-06 -1.73018106e-06]]
Intercept: [2.3772381764528827,1.848670703286689,1.7554800661225438,1.24041551253
8595,1.429535635670561,1.4145482182536695,0.9888091814484424,1.111803573195483,0.
6768531298117472,-0.27667448487983665,-0.2966676128486425,-0.32007604403100764,-0
.6704267867472261,-1.7037688035881366,-2.3135928801009085,-2.348226130575937,-4.9
13921454008919]
```

Took 22 sec. Last updated by anonymous at August 22 2019, 6:49:10 AM.

```
%spark2.pyspark
```

READY