TWITTER STREAMING CON SPARK

PROYECTO FINAL BASES DE DATOS NO RELACIONALES

INTRODUCCIÓN

Tendencias: Global · Cambiar

#FelicesPosadas
Promocionado por Gasolinera Pemex

Lula
314 mil Tweets

#15TemmuzUnutma
3.961 Tweets

#TürkünHürriyetAteşi
42 mil Tweets

#KurandakiCennetHayatı
1.890 Tweets

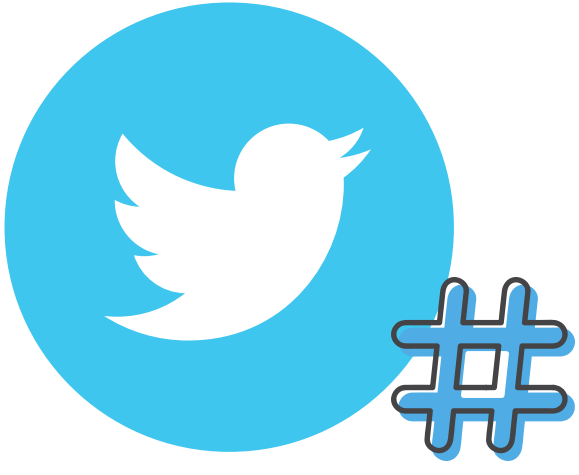
#ARSTOT
6.758 Tweets

#KadınGücü
55,1 mil Tweets

Hellboy
22,6 mil Tweets

Bale
51,9 mil Tweets

Marco Aurélio Mello
85,9 mil Tweets



OBJETIVOS

1

Analizar información en tiempo real proveniente de Twitter utilizando las herramientas de Streaming de Spark.

2

Extraer las palabras más usadas, localidades e idioma de un trending topic, otras hashtags relacionados y graficar su frecuencia

PROCESO

ANACONDA.

What is Anaconda?

Easily install 1,400+ data science packages

Manage packages, dependencies and environments with conda

Windows | macOS | Linux

Anaconda 5.3.1 For Windows Installe

Python 3.7 version *

Download

64-Bit Graphical Installer (633 MB)

32-Bit Graphical Installer (510 MB)

Python

64-Bit Graph

32-Bit Graph

We suggest the following mirror site for your download:

https://www-us.apache.org/dist/spark/spark-2.3.2/spark-2.3.2-bin-hadoop2.7.tgz

Other mirror sites are suggested below.

It is essential that you verify the integrity of the downloaded file using the PGP signature (`.asc` file) or a hash (`.md5` or

Please only use the backup mirrors to download KEYS, PGP and MD5 sigs/hashes or if no other mirrors are working.

HTTP

https://www-eu.apache.org/dist/spark/spark-2.3.2/spark-2.3.2-bin-hadoop2.7.tgz

https://www-us.apache.org/dist/spark/spark-2.3.2/spark-2.3.2-bin-hadoop2.7.tgz

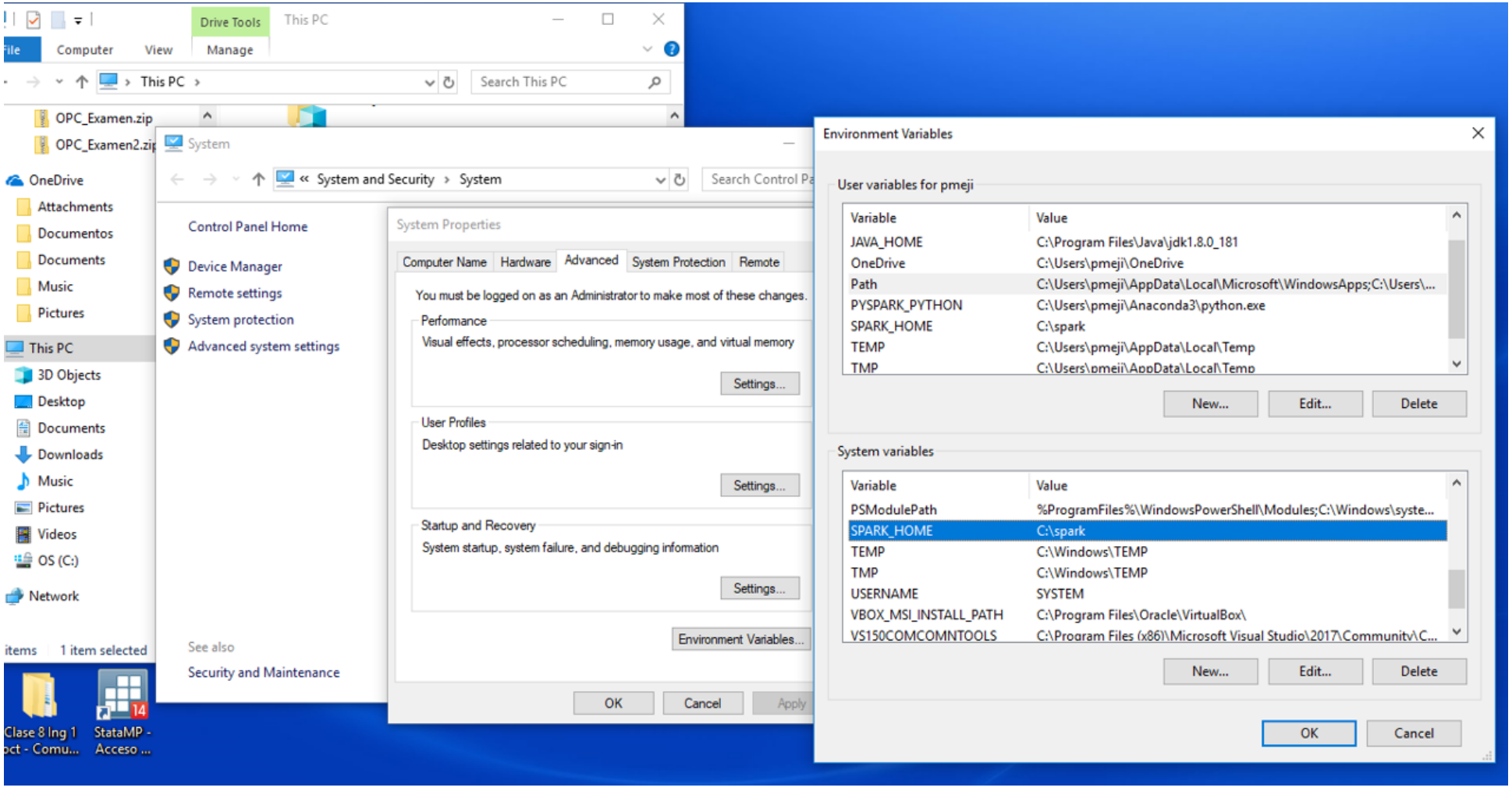
Windows binaries for Hadoop versions (built fron

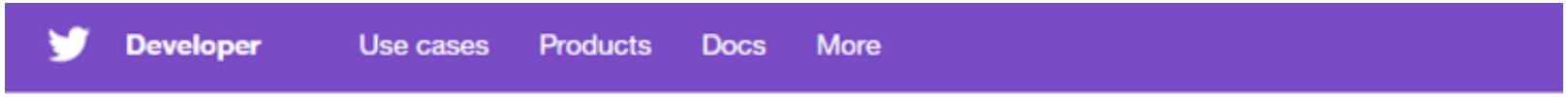
24 commits | 1 branch

Branch: master ▾ | New pull request

steveloughran improving script for adding the artifacts

hadoop-2.6.0/bin	Add Hadoop-2.6.
hadoop-2.6.3/bin	add gpg2 signatu
hadoop-2.6.4	add 2.6.4 and 2.7
hadoop-2.7.1	add 2.6.4 and 2.7
hadoop-2.8.0-RC3/bin	sign Hadoop arti
hadoop-2.8.1	sign Hadoop arti
hadoop-2.8.3/bin	Windows binarie
hadoop-3.0.0/bin	Hadoop 3.0.0 wir
.gitattributes	add gitattributes
.gitignore	add 2.6.4 and 2.7
KEYS	add my new key


Drive Tools
This PC
File
Computer
View
Manage
This PC
Search This PC
OPC_Examen.zip
OPC_Examen2.zip
OneDrive
Attachments
Documentos
Documents
Music
Pictures
This PC
3D Objects
Desktop
Documents
Downloads
Music
Pictures
Videos
OS (C:)
Network
items
1 item selected
Clase 8 Ing 1 ıct - Comu...
StataMP - Acceso ...
System
System and Security
System
Search Control Pa
Control Panel Home
Device Manager
Remote settings
System protection
Advanced system settings
See also
Security and Maintenance
System Properties
Computer Name
Hardware
Advanced
System Protection
Remote
You must be logged on as an Administrator to make most of these changes.
Performance
Visual effects, processor scheduling, memory usage, and virtual memory
Settings...
User Profiles
Desktop settings related to your sign-in
Settings...
Startup and Recovery
System startup, system failure, and debugging information
Settings...
Environment Variables...
OK
Cancel
Apply
Environment Variables
User variables for pmeji
Variable
Value
JAVA_HOME
C:\Program Files\Java\jdk1.8.0_181
OneDrive
C:\Users\pmeji\OneDrive
Path
C:\Users\pmeji\AppData\Local\Microsoft\WindowsApps;C:\Users\...
PYSPARK_PYTHON
C:\Users\pmeji\Anaconda3\python.exe
SPARK_HOME
C:\spark
TEMP
C:\Users\pmeji\AppData\Local\Temp
TMP
C:\Users\pmeji\AppData\Local\Temp
New...
Edit...
Delete
System variables
Variable
Value
PSModulePath
%ProgramFiles%\WindowsPowerShell\Modules;C:\Windows\syste...
SPARK_HOME
C:\spark
TEMP
C:\Windows\TEMP
TMP
C:\Windows\TEMP
USERNAME
SYSTEM
VBOX_MSI_INSTALL_PATH
C:\Program Files\Oracle\VirtualBox\
VS150COMCOMNTOOLS
C:\Program Files (x86)\Microsoft Visual Studio\2017\Community\C...
New...
Edit...
Delete
OK
Cancel



Get started with Twitter APIs and tools.

Apply for access.

All new developers must apply for a developer account to access T
use our standard APIs and our new premium APIs.

Apply for a developer account >

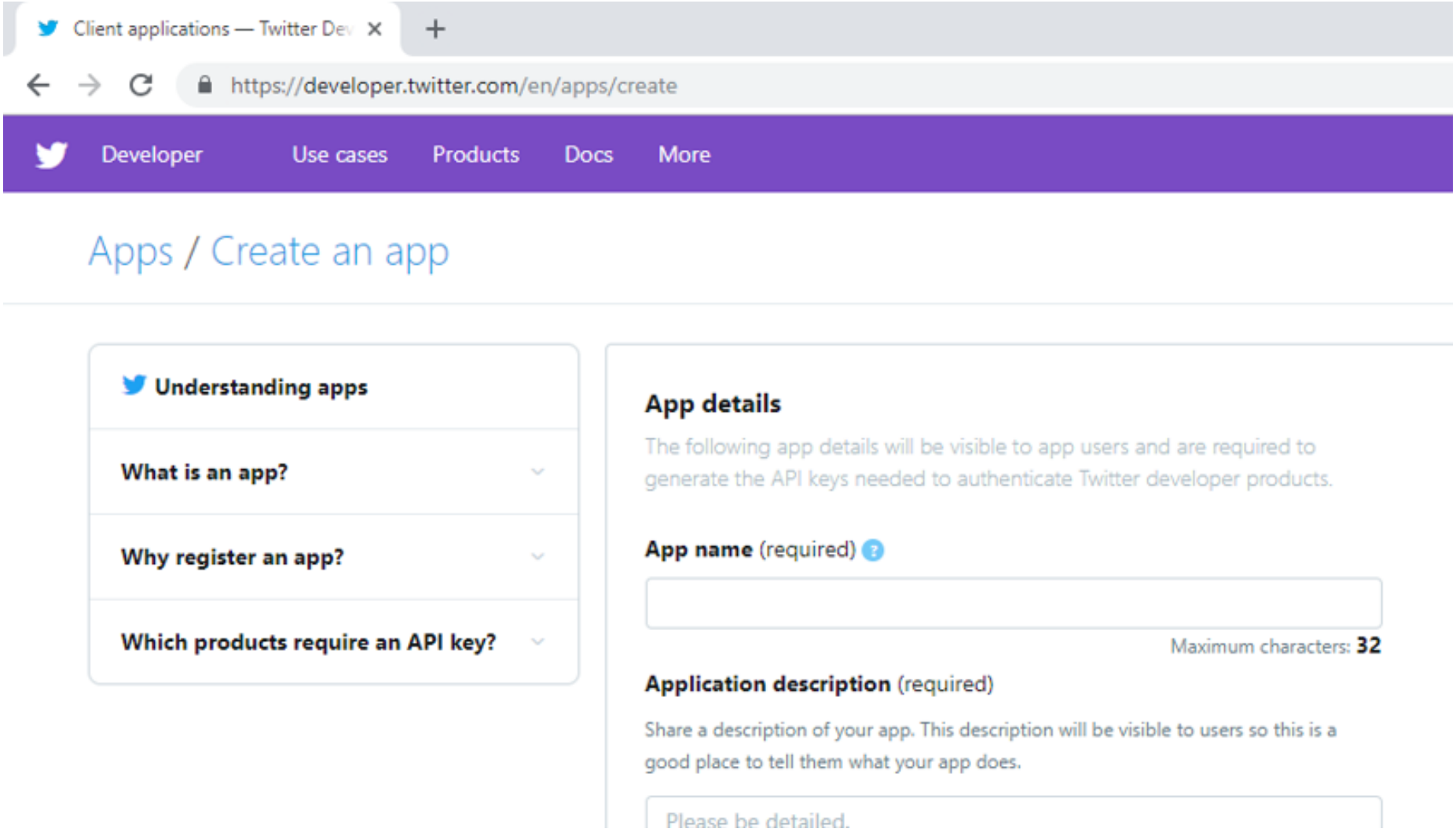
Restricted use cases >

Twitter developer account application [ref:_00DA0K0A8._5004A1YlOxb:ref]

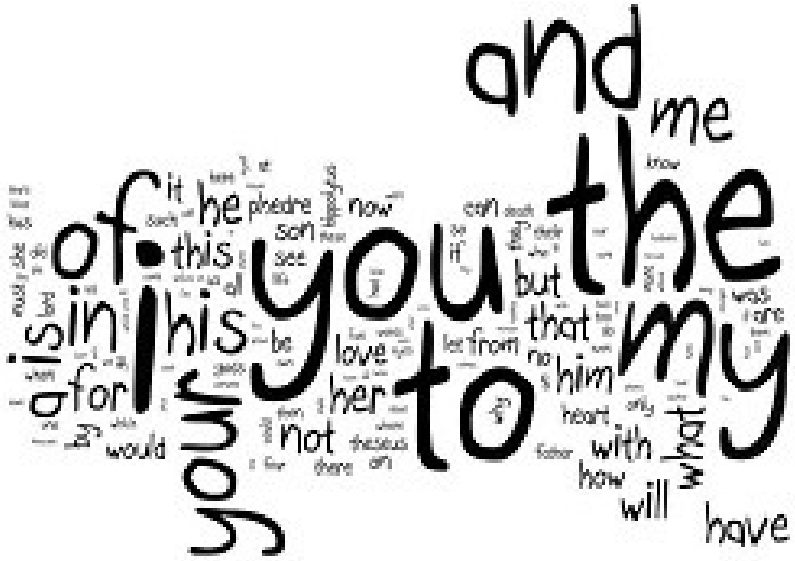
Recibidos ×

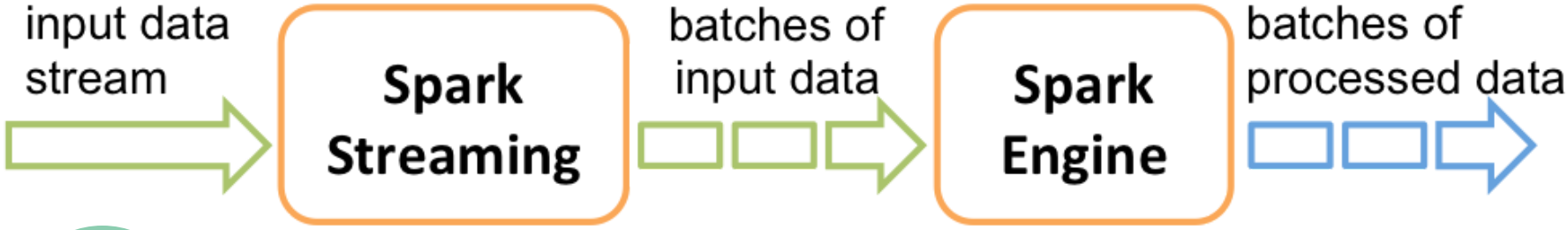
developer-accounts@twitter.com <developer-accounts@twitter.com>
para mí

inglés > español Traducir mensaje

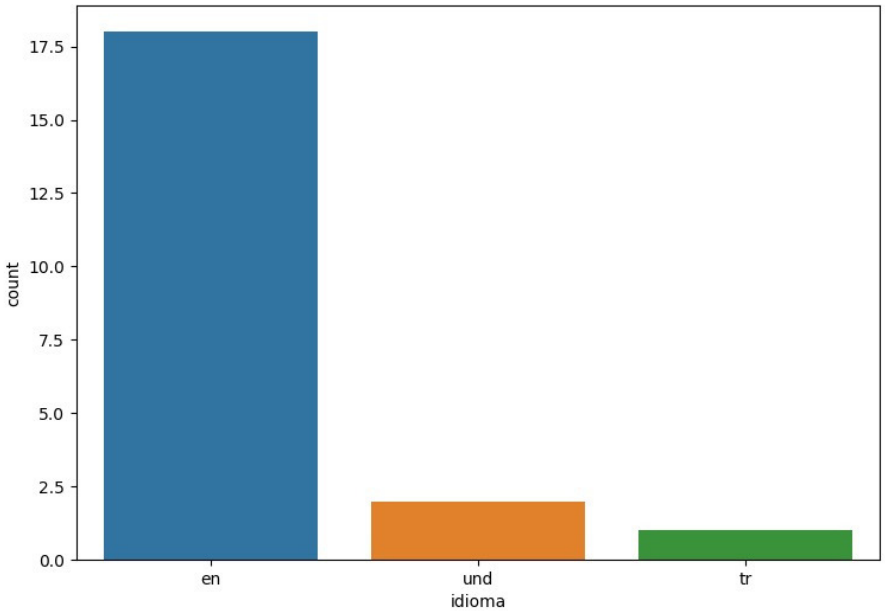

Client applications — Twitter De
https://developer.twitter.com/en/apps/create
Developer
Use cases
Products
Docs
More
Apps / Create an app
Understanding apps
What is an app?
Why register an app?
Which products require an API key?
App details
The following app details will be visible to app users and are required to generate the API keys needed to authenticate Twitter developer products.
App name (required)
Maximum characters: 32
Application description (required)
Share a description of your app. This description will be visible to users so this is a good place to tell them what your app does.
Please be detailed.

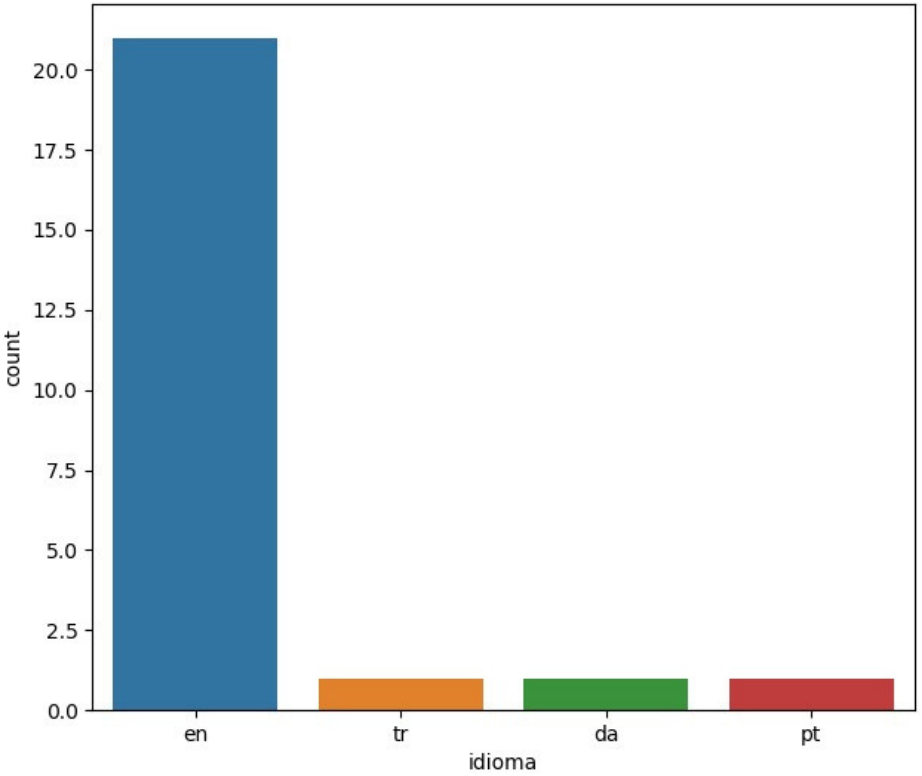
PROCESO REAL

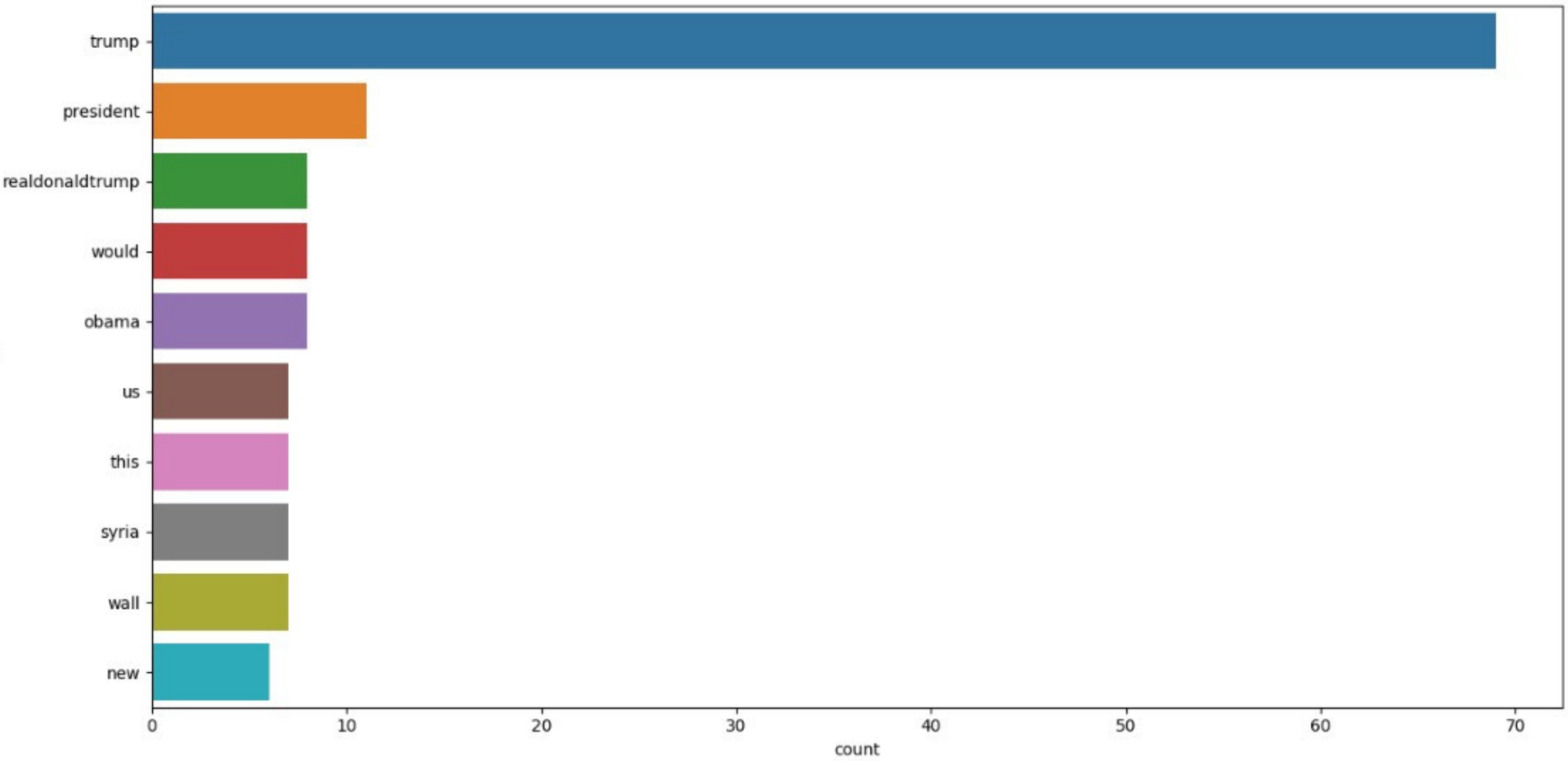


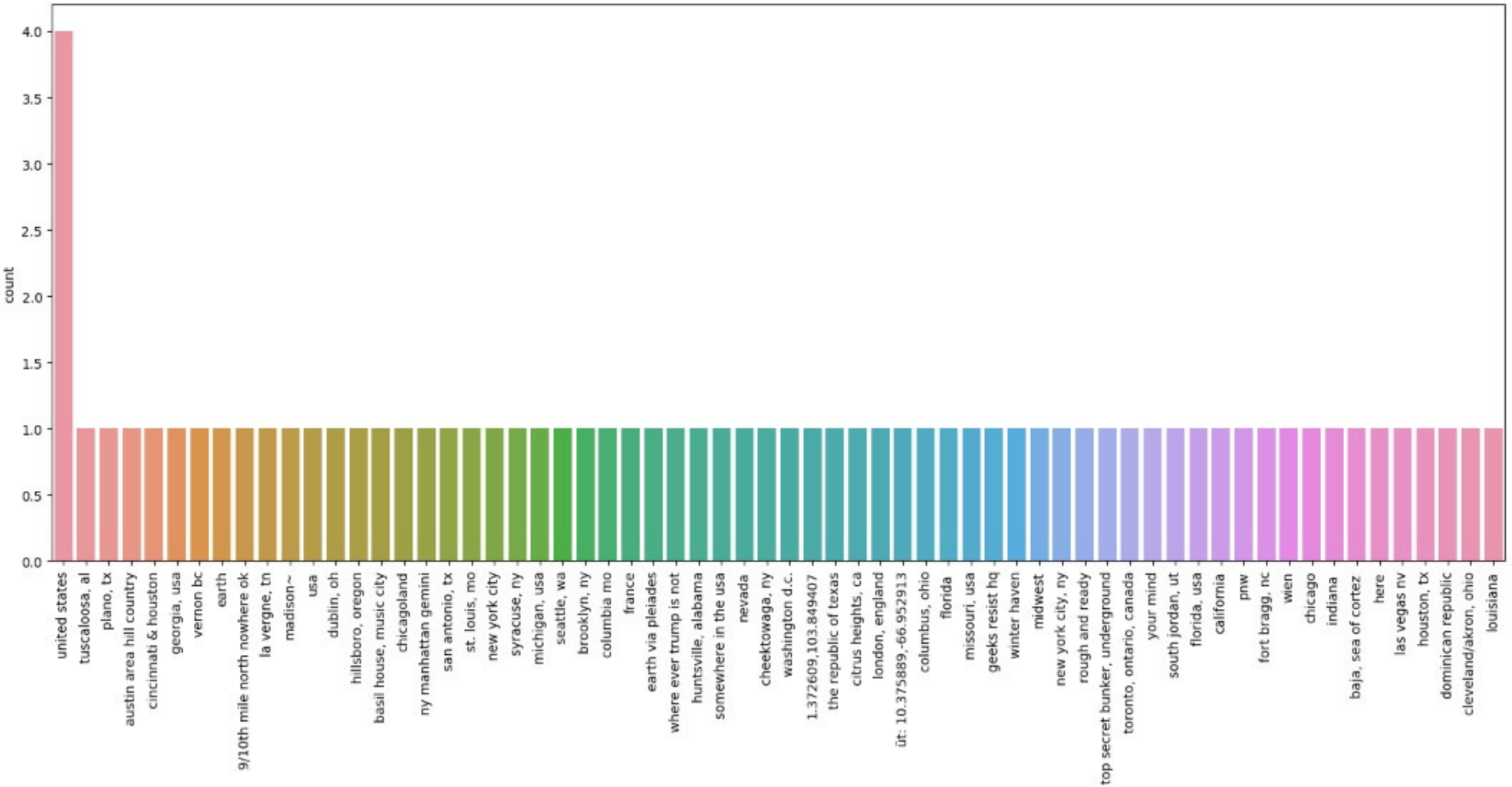


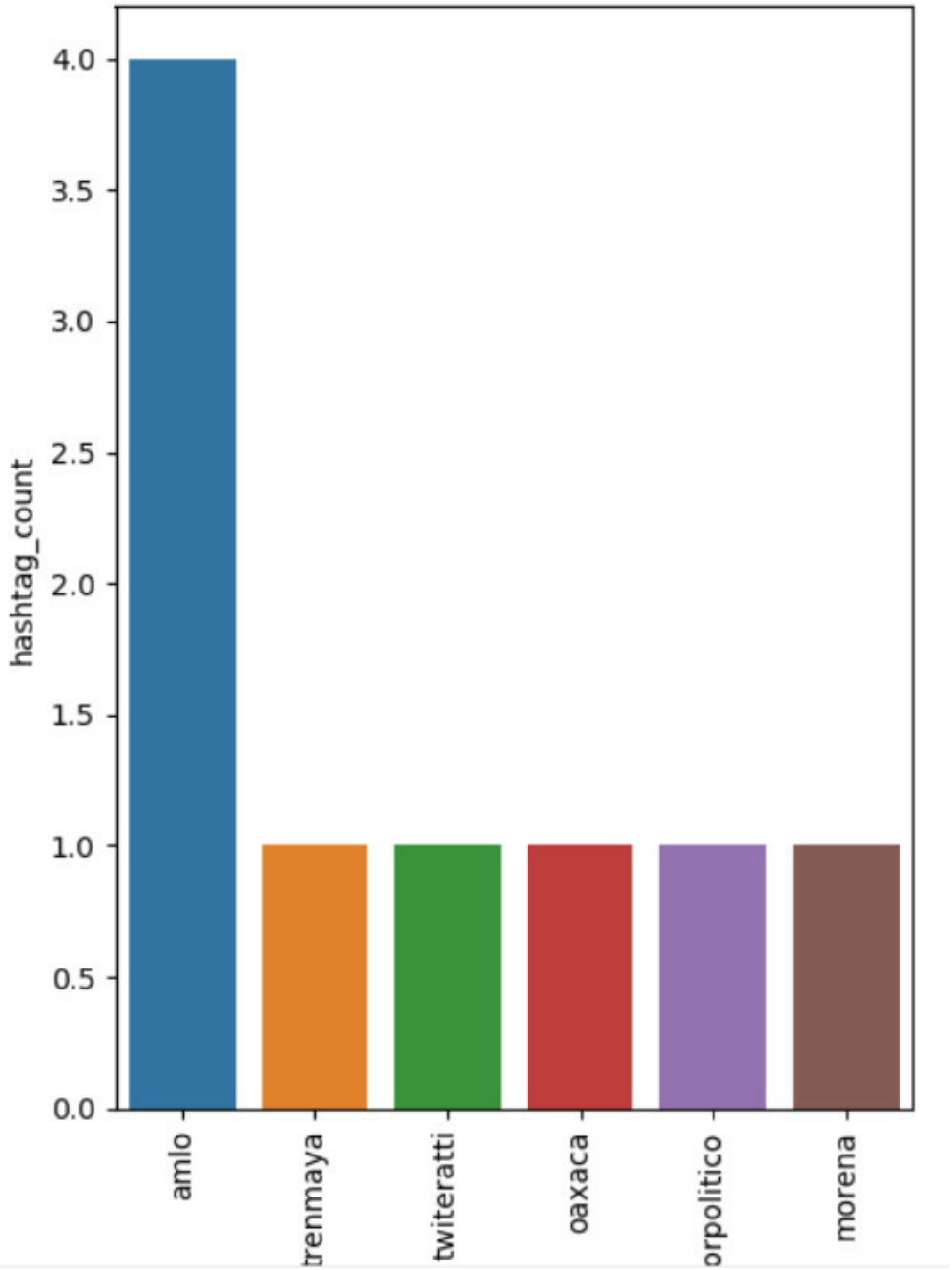
EJEMPLOS

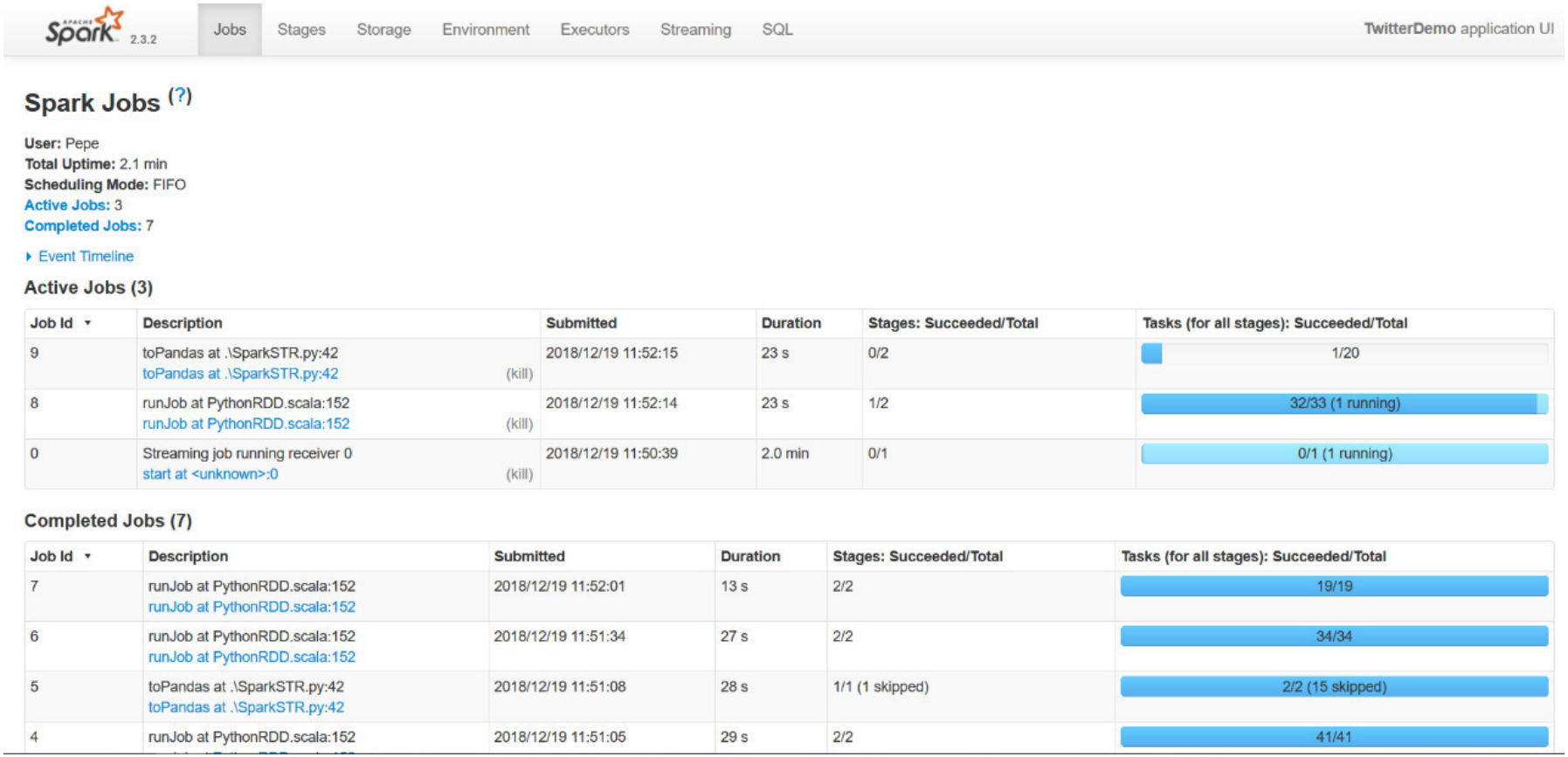

17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
count
en
und
tr
idioma


20.0
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
count
en
tr
da
pt
idioma


trump
president
realdonaldtrump
would
obama
us
this
syria
wall
new
0
10
20
30
40
50
60
70
count


count
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
united states
tuscaloosa, al
plano, tx
austin area hill country
cincinnati & houston
georgia, usa
vernon bc
earth
9/10th mile north nowhere ok
la vergne, tn
madison~
usa
dublin, oh
hillsboro, oregon
basil house, music city
chicagoland
ny manhattan gemini
san antonio, tx
st. louis, mo
new york city
syracuse, ny
michigan, usa
seattle, wa
brooklyn, ny
columbia mo
france
earth via pleiades
where ever trump is not
huntsville, alabama
somewhere in the usa
nevada
cheektowaga, ny
washington d.c.
1.372609,103.849407
the republic of texas
citrus heights, ca
london, england
üt: 10.375889,-66.952913
columbus, ohio
florida
missouri, usa
geeks resist hq
winter haven
midwest
new york city, ny
rough and ready
top secret bunker, underground
toronto, ontario, canada
your mind
south jordan, ut
florida, usa
california
pnw
fort bragg, nc
wien
chicago
indiana
baja, sea of cortez
here
las vegas nv
houston, tx
dominican republic
cleveland/akron, ohio
louisiana


hashtag_count
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
amlo
trenmaya
twiteratti
oaxaca
orpolitico
morena


Spark 2.3.2
Jobs
Stages
Storage
Environment
Executors
Streaming
SQL
TwitterDemo application UI
Spark Jobs (?)
User: Pepe
Total Uptime: 2.1 min
Scheduling Mode: FIFO
Active Jobs: 3
Completed Jobs: 7
Event Timeline
Active Jobs (3)
Job Id
Description
Submitted
Duration
Stages: Succeeded/Total
Tasks (for all stages): Succeeded/Total
9
toPandas at .\SparkSTR.py:42
toPandas at .\SparkSTR.py:42 (kill)
2018/12/19 11:52:15
23 s
0/2
1/20
8
runJob at PythonRDD.scala:152
runJob at PythonRDD.scala:152 (kill)
2018/12/19 11:52:14
23 s
1/2
32/33 (1 running)
0
Streaming job running receiver 0
start at <unknown>:0 (kill)
2018/12/19 11:50:39
2.0 min
0/1
0/1 (1 running)
Completed Jobs (7)
Job Id
Description
Submitted
Duration
Stages: Succeeded/Total
Tasks (for all stages): Succeeded/Total
7
runJob at PythonRDD.scala:152
runJob at PythonRDD.scala:152
2018/12/19 11:52:01
13 s
2/2
19/19
6
runJob at PythonRDD.scala:152
runJob at PythonRDD.scala:152
2018/12/19 11:51:34
27 s
2/2
34/34
5
toPandas at .\SparkSTR.py:42
toPandas at .\SparkSTR.py:42
2018/12/19 11:51:08
28 s
1/1 (1 skipped)
2/2 (15 skipped)
4
runJob at PythonRDD.scala:152
2018/12/19 11:51:05
29 s
2/2
41/41

RESULTADOS

CONCLUSIONES

DEMO

LA BANDA GANGRENA