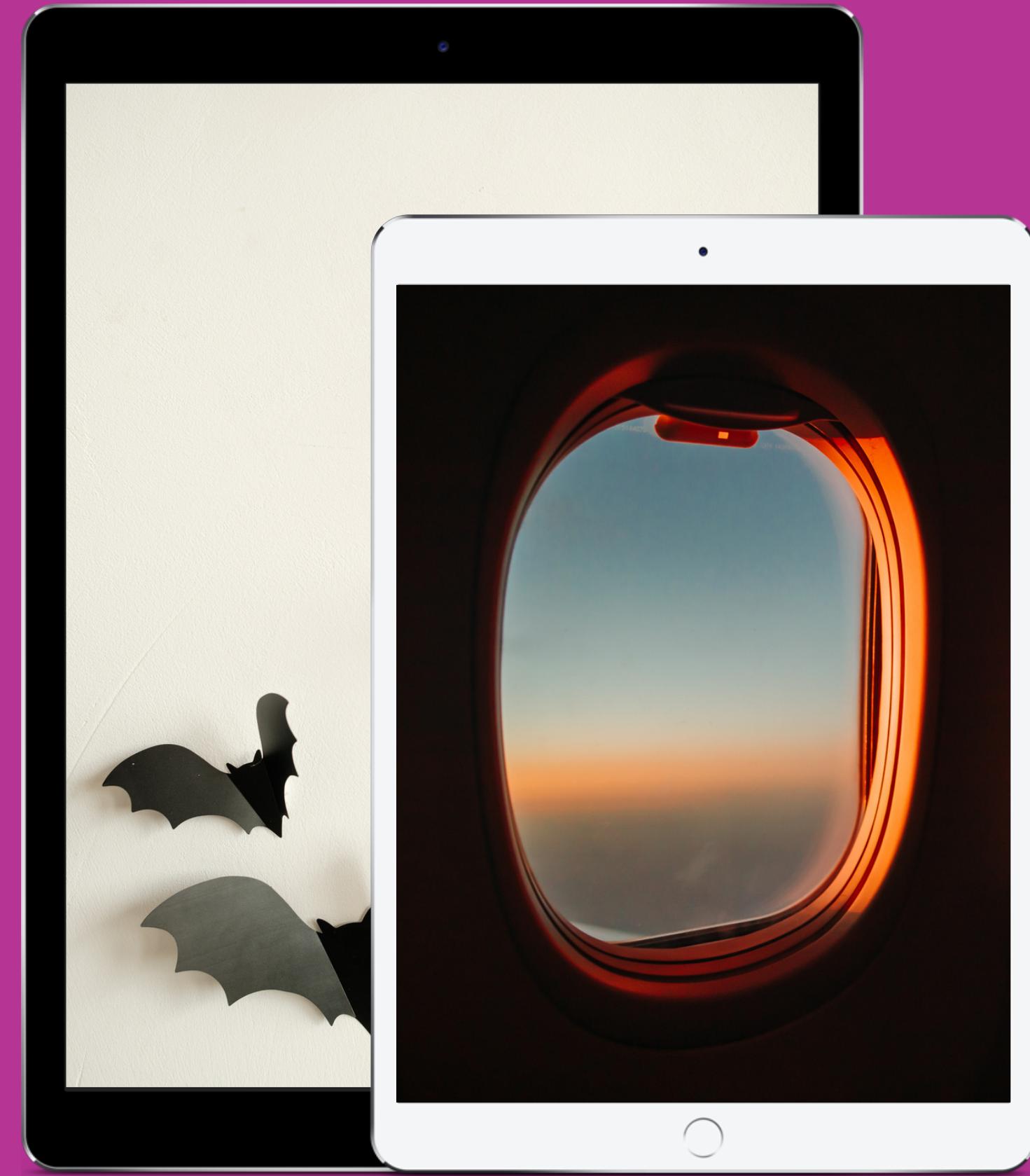


Jairo Sierra - 11811364

Renato Lizardo - 11811148

# Proyecto Final - Hadoop



Dataset

WallStreetBets  
Comments

# ¿Qué es WallStreetBets?

Es un SubReddit en donde los participantes discuten el comercio de acciones y opiniones. Se ha vuelto notable sus estrategias comerciales agresivas y su papel en la contracción corta de GameStop, el cual es un minorista estadounidense de productos de videojuegos, electrónica de consumo y juegos que causó pérdidas en posiciones cortas en empresas estadounidenses que superaron los 70 mil millones de dólares en unos pocos días a principios de 2021. El SubReddit es famoso por su jerga colorida y términos.



# PreProcesamiento

**PreProcesamiento**

Cargar archivo

Analizar

**Ordenamiento**

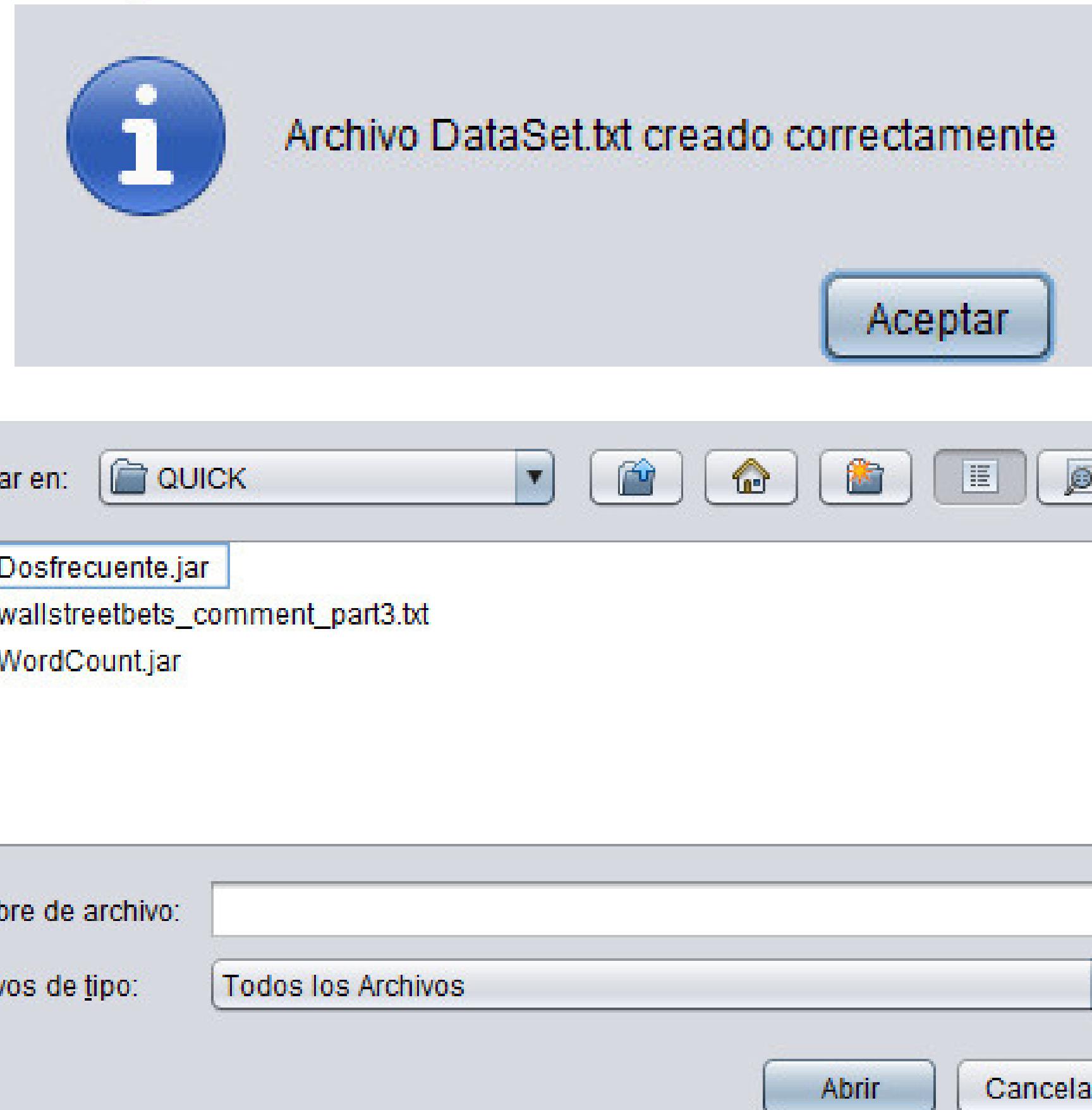
Un Frecuente

Dos Frecuente

Nombre de archivo:

Archivos de tipo: Todos los Archivos

Aceptar



# PreProcesamiento - Flex

```
EliminarEsto = as |zx| at | be |dm |zy |mo|dh |Anyone |mu|wt |eqd |nqgj |gn| ycf1 |qyn |qsm| ds | dq | jxzg|lmao|lgko|vq|yw  
|ag|bv|ya|em|oy|dj|sk|dz|dl|Nah |rjrcs |gt |re |zrn |ozw |dr|yo|subredditmessagecompositorwallstreetbets|anyone |pm | nf  
| zpy | wz | ur | pump|Su |obef | dfomqwa | zukv| uze | ukx | bm | bitch | because | pu | geci | been | before | dfolzyu  
| qtbk | HEHEHEH | being | below | between | both | ejburritos | but | by | can | can't | cannot | could | couldn't | did  
| didn't | do | does | doesn't | doing | don't | down | during | deleted | each | few | for | from | further| fuck | Fucking  
|faggot| stupid | fucking | had | hadn't | has | hasn't | have | haven't | having | he | he'd | he'll | he's | her | here  
| here's | hers | herself | him | himself | his | how | how's | i | i'd | i'll | i'm | i've | if | in | into | is | isn't  
| it | it's | its| vq | itself | let's | me | more | most | mustn't | my | myself| anyway | once | only | or | other | ought  
| our | ours | ourselves | out | over | own | same | shan't | nqgj | she | she'd | she'll | she's | should | shouldn't | so  
| some | such | than | that | that's | the | their | theirs | them | themselves | then | there|TrueLibertyorDeath | there's  
| these | they | they'd | they'll | they're | they've| bv |ag |yw | xa | this | those | through | to | too | under | until  
| up | very | was | wasn't | we | we'd | we'll | we're | we've | were | weren't | what | what's | when | when's | where  
| where's | which | while | who | who's | whom | why | why's | with | won't | would | wouldn't | you | you'd | you'll  
| you're | you've | your | yours | yourself | yourselves | quot | arent | cant | couldnt | didnt | doesnt | dont | hadnt  
| hasnt | havent | hed | hes | heres | hows | im | ive | isnt | its | shes | shouldnt | shit | thats | theres | theyd  
| theyll | theyre | theyve | wasnt | a+ | about | above | after | again | against | all | th | am | an | and | any | are  
| aren't | wed | weve | were | wouldnt | youd | youll | youre | youve | twice | just | like | likely | later | try | even  
| will | simply | almost | get | ok | must | else | without | put | unless | much | ah | yeah | maybe | need | also | used  
| may | yet | obviously | obvious | many | us | use | etc | oh | since | yah | come | knows | know | knew | yes | true  
| something | really | anything | sense | b+ | c+ | d+ | e+ | f+ | g+ | h+ | i+ | j+ | k+ | l+ | m+ | n+ | o+ | p+ | q+  
| r+ | s+ | t+ | u+ | v+ | w+ | x+ | y+ | z+ | lol | within | it'll | ha+ | HA+ | ja+| je+ | wallstreetbets | True  
| False|ea |di |wsb | dx |cut | werent | whats | whos | whys | wont|penis | no | nor | not | of | off | on
```

```
mala = {malo}{palabra}  
malo = dfo | qmt | qtb | dty | dtx | ecw | aaz | dh | AMAT | zz | zy | zx | zv | zw | AAA | zuu | zuf | zt | zs | yw | AAHAHA  
| EEE | shit | qsm | woo | WOO | vjq | vjp | ebn | zucc |zrn  
palabra = [a-zA-Z]+  
numero = [0-9]+  
espacio = [^\x00-\x7F]+ | [\r\s\v\f]+
```

# Iniciar Hadoop

cmd Administrador: Símbolo del sistema

```
0000001, C:\hadoop\data\namenode\current\fimage_0000000000000000, C:\hadoop\data\namenode  
00000000.md5, C:\hadoop\data\namenode\current\seen_txid, C:\hadoop\data\namenode\current\VER  
21/09/17 15:25:19 INFO common.Storage: Storage directory C:\hadoop\data\namenode has been suc  
21/09/17 15:25:19 INFO namenode.FSImageFormatProtobuf: Saving image file C:\hadoop\data\namen  
00000000000000 using no compression  
21/09/17 15:25:19 INFO namenode.FSImageFormatProtobuf: Image file C:\hadoop\data\namenode\cur  
0000000000 of size 400 bytes saved in 0 seconds .  
21/09/17 15:25:19 INFO namenode.NNStorageRetentionManager: Going to retain 1 images with txid  
21/09/17 15:25:19 INFO namenode.FSNamesystem: Stopping services started for active state  
21/09/17 15:25:19 INFO namenode.FSNamesystem: Stopping services started for standby state  
21/09/17 15:25:19 INFO namenode.FSImage: FSImageSaver clean checkpoint: txid=0 when meet shut  
21/09/17 15:25:19 INFO namenode.NameNode: SHUTDOWN_MSG:  
*****  
SHUTDOWN_MSG: Shutting down NameNode at DESKTOP-88T8HT4/192.168.56.1  
*****
```

```
C:\WINDOWS\system32>cd C:\hadoop\sbin
```

```
C:\hadoop\sbin>start-all.cmd  
This script is Deprecated. Instead use start-dfs.cmd and start-yarn.cmd  
starting yarn daemons
```

```
C:\hadoop\sbin>jps  
16160 NameNode  
3312 Jps  
11812 NodeManager  
12652 DataNode  
13148 ResourceManager
```

```
C:\hadoop\sbin>j
```

# Montar El DataSet En Hadoop

```
C:\hadoop\sbin>hadoop fs -mkdir /input
```

```
C:\hadoop\sbin>hadoop fs -put C:/dataset/DataSet.txt /input
```

```
C:\hadoop\sbin>fs -ls /input/
```

"fs" no se reconoce como un comando interno o externo,  
programa o archivo por lotes ejecutable.

```
C:\hadoop\sbin>hdfs fs -ls /input/
```

Error: no se ha encontrado o cargado la clase principal fs

```
C:\hadoop\sbin>hadoop fs -ls /input/
```

Found 1 items

```
-rw-r--r--    1 renat supergroup  251516822 2021-09-17 16:39 /input/DataSet.txt
```

```
C:\hadoop\sbin>
```

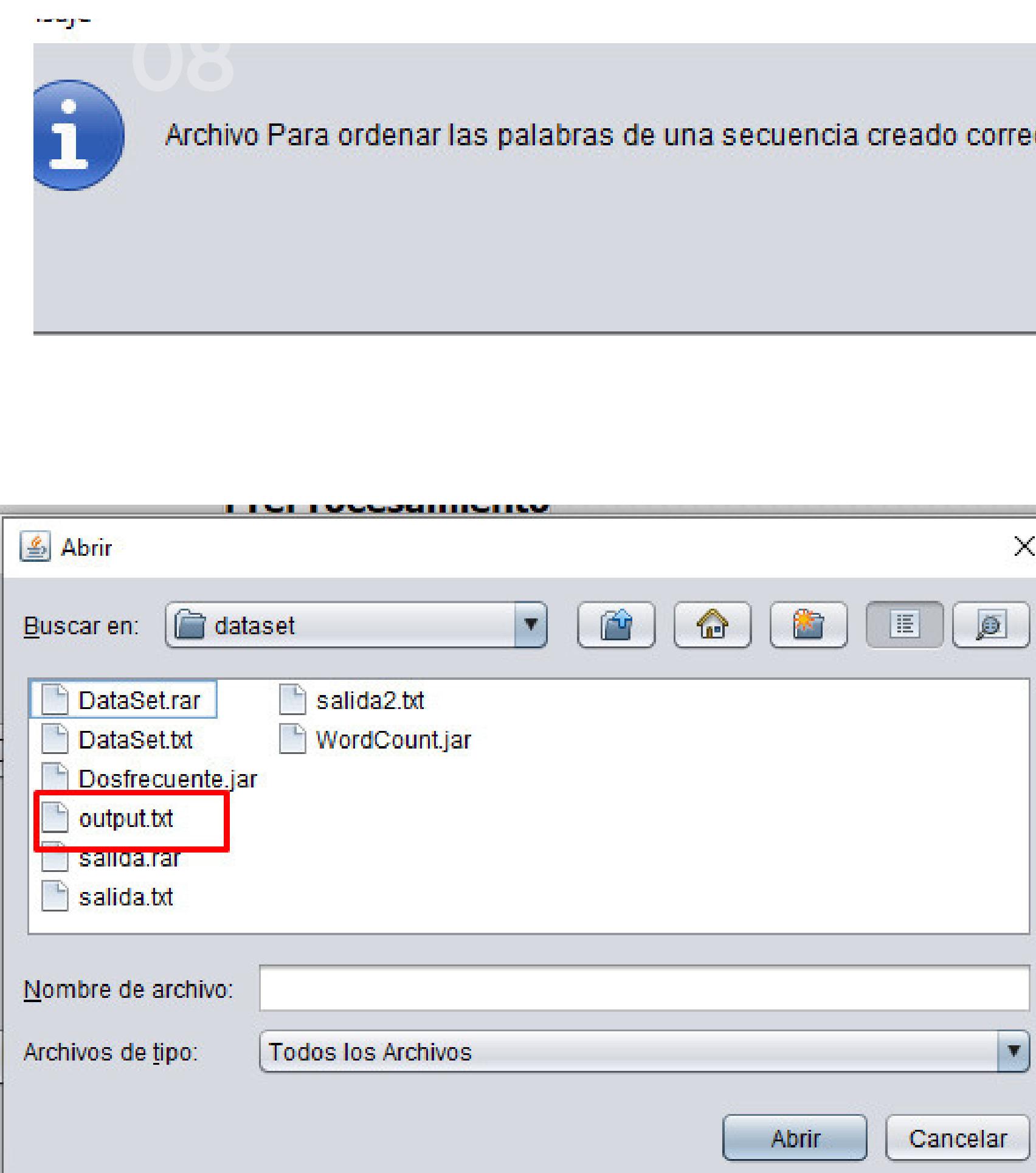
# Proceso de Map-Reduce

```
C:\hadoop\sbin>hadoop jar C:/dataset/WordCount.jar WordCount /input /out
21/09/17 16:45:07 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
21/09/17 16:45:08 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
21/09/17 16:45:09 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/renat/.staging/job_1631913970226_0001
21/09/17 16:45:09 INFO input.FileInputFormat: Total input files to process : 1
21/09/17 16:45:09 INFO mapreduce.JobSubmitter: number of splits:2
21/09/17 16:45:10 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1631913970226_0001
21/09/17 16:45:10 INFO mapreduce.JobSubmitter: Executing with tokens: []
21/09/17 16:45:10 INFO conf.Configuration: resource-types.xml not found
21/09/17 16:45:10 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
21/09/17 16:45:11 INFO impl.YarnClientImpl: Submitted application application_1631913970226_0001
21/09/17 16:45:11 INFO mapreduce.Job: The url to track the job: http://DESKTOP-88T8HT4:8088/proxy/application_1631913970226_0001/
21/09/17 16:45:11 INFO mapreduce.Job: Running job: job_1631913970226_0001
```

# Proceso de Map-Reduce

```
226_0002/
21/09/17 16:48:54 INFO mapreduce.Job: Running job: job_1631913970226_0002
21/09/17 16:49:23 INFO mapreduce.Job: Job job_1631913970226_0002 running in uber mode : false
21/09/17 16:49:23 INFO mapreduce.Job: map 0% reduce 0%
21/09/17 16:49:42 INFO mapreduce.Job: map 33% reduce 0%
21/09/17 16:49:49 INFO mapreduce.Job: map 45% reduce 0%
21/09/17 16:49:55 INFO mapreduce.Job: map 58% reduce 0%
21/09/17 16:50:01 INFO mapreduce.Job: map 67% reduce 0%
21/09/17 16:50:03 INFO mapreduce.Job: map 83% reduce 0%
21/09/17 16:50:06 INFO mapreduce.Job: map 100% reduce 0%
21/09/17 16:50:13 INFO mapreduce.Job: map 100% reduce 100%
21/09/17 16:50:23 INFO mapreduce.Job: Job job_1631913970226_0002 completed successfully
21/09/17 16:50:23 INFO mapreduce.Job: Counters: 54
  File System Counters
    FILE: Number of bytes read=61605754
    FILE: Number of bytes written=89281773
    FILE: Number of read operations=0
    FILE: Number of large read operations=0
    FILE: Number of write operations=0
    HDFS: Number of bytes read=251521126
    HDFS: Number of bytes written=18219787
    HDFS: Number of read operations=11
    HDFS: Number of large read operations=0
    HDFS: Number of write operations=2
    HDFS: Number of bytes read erasure-coded=0
```

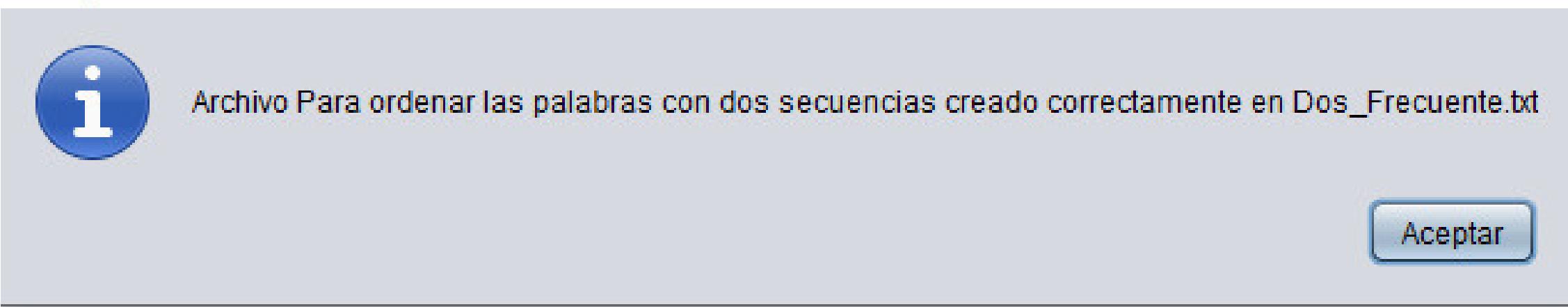
# Proceso de Ordenado - Un Frecuente



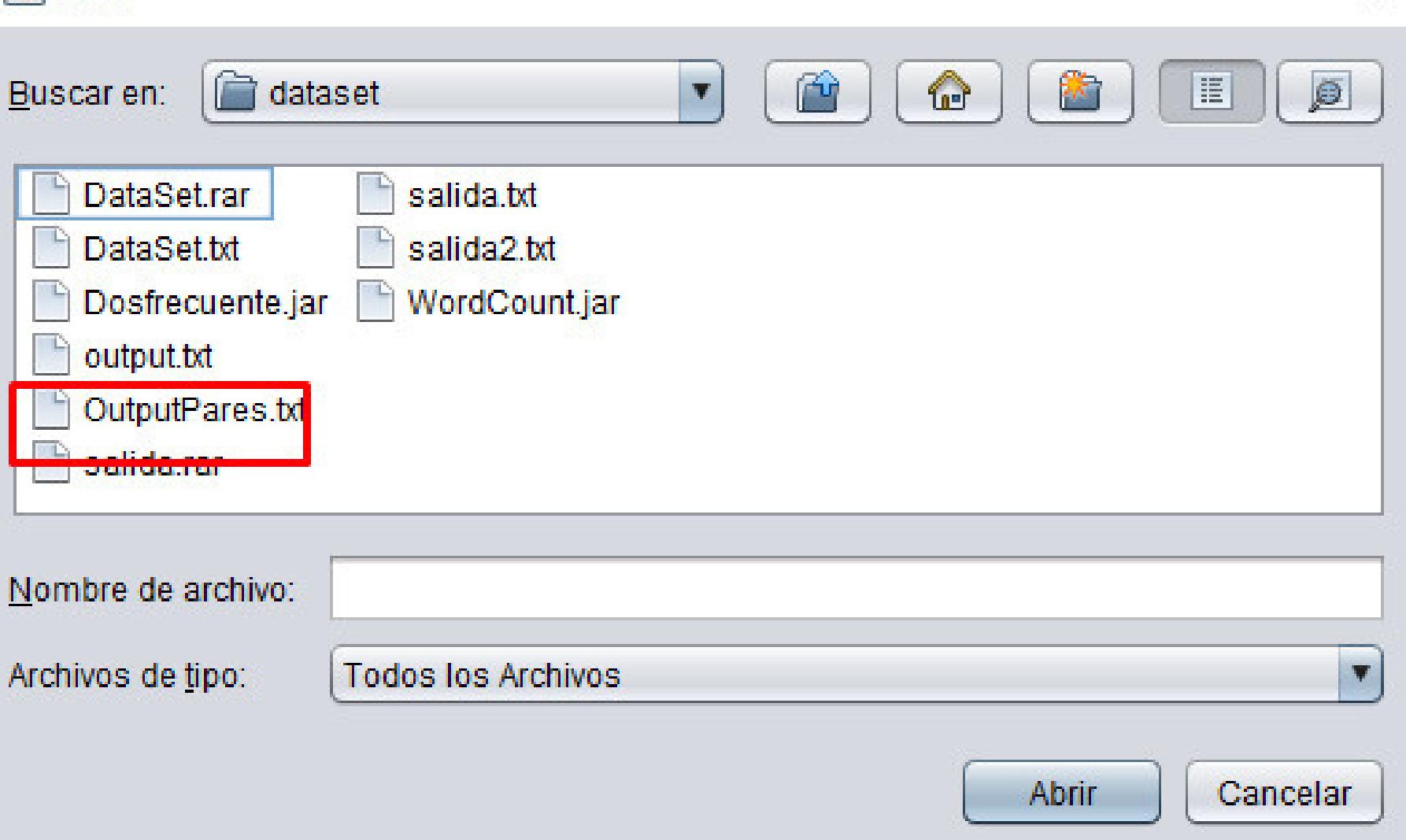
Un\_frecuente.txt: Bloc de notas  
Archivo Edición Formato Ver Ayuda  
year 40266  
call 40449  
last 40940  
earnings 42972  
price 44072  
want 45014  
week 48614  
long 49496  
options 50301  
way 50511  
today 51363  
bought 53575  
got 57178  
still 58189  
see 61098  
sell 62176  
back 64036  
day 65829  
stock 66822  
make 66881  
right 67478  
people 73110  
time 77874  
good 78542  
go 79217  
one 79795  
market 84994  
going 88732  
puts 89708  
now 91863  
think 94177  
buy 97646  
removed 108434  
calls 114642  
money 115268

# Proceso de Ordenado - Dos Frecuente

Mensaje



Abrir



Dos\_Frecuente.txt: Bloc de notas

Archivo Edición Formato Ver Ayuda

dgaff says 5061

fartbiscuit dgaff 5061

tastes marshmallows 5061

says tastes 5062

concerns moderator 5294

moderators questions 5367

automatically please 5370

contact moderators 5371

questions concerns 5374

please contact 5395

performed automatically 5526

bot action 5530

action performed 5531

stock market 5597

wsbvotebot robot 5708

robot janitor 5713

spaghetti spaghetti 5812

short term 6168

october mwq 6198

somehow lost 6236

puts october 6251

good luck 6612

money puts 6750

buy calls 6794

buy puts 7823

lose money 8105

last week 8158

lost money 8574

1 ec 8710

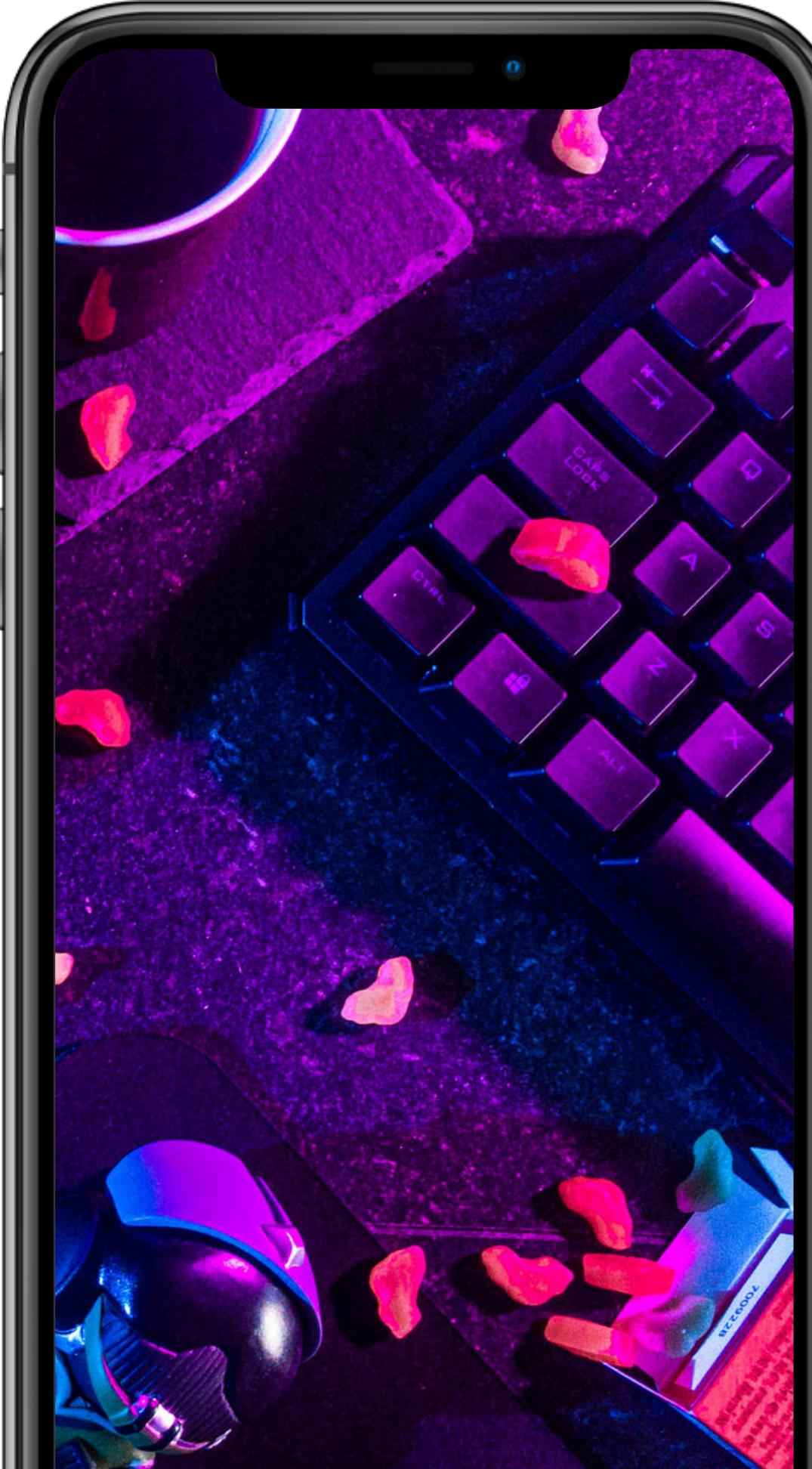
long term 8927

next week 9523

make money 11176

avgazn retard 12834

right now 21977



# Top de Un Frecuente

1. money 115268
2. calls 114642
3. buy 97646
4. puts 89708
5. market 84994
6. stock 66822
7. sell 62176
8. bought 53575
9. earnings 42972
10. company 33580



# Interesantes de Un Frecuente

AMD 28689

ILL 2241

SPY 17894

Tesla 23681

Friday 11657

FB 10428

profit 24124

Trump 20513

China 15811

Amazon 12053

Apple 21128

Robinhood 9561



# Top de Dos Frecuente

1. right now 21977
2. make money 11176
3. next week 9523
4. long term 8927
5. lost money 8574
6. last week 8158
7. buy puts 7823
8. buy calls 6794
9. money puts 6750
10. good luck 6612



Gracias Por  
Su Atención