Laboratory Assignment 3.1

Introduction

In this laboratory, we will cover the basic element of programming using a map-reduce methodology. For that purpose, we will be using Apache Spark as a reference, but bear in mind that similar frameworks exists and principles can be extrapolated.

Some concepts and facts

- Spark is a distributed computing platform that operates on a cluster. Like MPI, we expect that nodes does NOT share a memory space but they are connected in high-speed dedicated network. Distributed filesystems that work over the network are extremely useful.
- It is considered the next generation of previous map-reduce standard Apache Hadoop. Main difference is thought to be the use of the memory instead of disk for intermediate operations, but there are many more improvements.
- It is built on Java. Despite this, it can be programmed using Java, Scala, Python or R. The complete API can only be found in JVM-based languages but the most frequent one is PySpark, since people is reluctant to use JVM-based languages in data science. Indeed, since Hadoop was only available for Java, it is likely that some Java codes of Spark are adaptations of previous Hadoop codes.
- Resilient Distributed Dataset (RDD): the basic unit that is processed in Spark. Equivalent to a numpy array but distributed.
- RDD API usually exposes the low-level operations of Apache Spark, useful for preprocessing data but useless for data analytics
- For data analysis, Dataframe and Spark SQL is used. It relies on a pandas-alike API that even accepts SQL code (which may sound crazy and useless for developers, but many *old* data scientists and statisticians are really proficient in SQL but not in Python).

How to install Spark in colab.

```
spark_version = "3.5.0"
hadoop_version = "3"

!apt-get update
!apt-get install openjdk-8-jdk-headless -qq # Install JVM v8
!wget -O spark-{spark_version}-bin-hadoop{hadoop_version}.tgz -q https://downloads.apache.org/spark/spark-{spark_version}/spark-{spark_v
!tar xf spark-{spark_version}-bin-hadoop{hadoop_version}.tgz # Unzip
!pip install pyspark # Well, the library itself
import os
os.environ["JAVA_HOME"] = "/usr/lib/jvm/java-8-openjdk-amd64"
os.environ["SPARK_HOME"] = f"/content/spark-{spark_version}-bin-hadoop{hadoop_version}"
```

```
Hit:12 https://ppa.launchpadcontent.net/ubuntugis/ppa/ubuntu jammy InRelease
Get:13 http://archive.ubuntu.com/ubuntu jammy-updates/multiverse amd64 Packages [49.8 kB]
Get:14 http://archive.ubuntu.com/ubuntu jammy-updates/main amd64 Packages [1,473 kB]
Fetched 3,077 kB in 1s (2,272 kB/s)
Reading package lists... Done
Selecting previously unselected package libxtst6:amd64.
(Reading database ... 120880 files and directories currently installed.)
Preparing to unpack .../libxtst6_2%3a1.2.3-1build4_amd64.deb ...
Unpacking libxtst6:amd64 (2:1.2.3-1build4) ...
Selecting previously unselected package openjdk-8-jre-headless:amd64.
Preparing to unpack .../openjdk-8-jre-headless_8u382-ga-1~22.04.1_amd64.deb ...
Unpacking openjdk-8-jre-headless:amd64 (8u382-ga-1~22.04.1) ...
Selecting previously unselected package openjdk-8-jdk-headless:amd64.
Preparing to unpack .../openjdk-8-jdk-headless_8u382-ga-1~22.04.1_amd64.deb ...
Unpacking openjdk-8-jdk-headless:amd64 (8u382-ga-1~22.04.1) ...
Setting up libxtst6:amd64 (2:1.2.3-1build4) ...
Setting up openjdk-8-jre-headless:amd64 (8u382-ga-1~22.04.1) ...
update-alternatives: using /usr/lib/jvm/java-8-openjdk-amd64/jre/bin/orbd to provide /usr/bin/orbd (orbd) in auto mode
update-alternatives: using /usr/lib/jvm/java-8-openjdk-amd64/jre/bin/servertool to provide /usr/bin/servertool (servertool) in au
update-alternatives: using /usr/lib/jvm/java-8-openjdk-amd64/jre/bin/tnameserv to provide /usr/bin/tnameserv (tnameserv) in auto
Setting up openjdk-8-jdk-headless:amd64 (8u382-ga-1~22.04.1) ...
update-alternatives: using /usr/lib/jvm/java-8-openjdk-amd64/bin/clhsdb to provide /usr/bin/clhsdb (clhsdb) in auto mode
update-alternatives: using /usr/lib/jvm/java-8-openjdk-amd64/bin/extcheck to provide /usr/bin/extcheck (extcheck) in auto mode
update-alternatives: using /usr/lib/jvm/java-8-openjdk-amd64/bin/hsdb to provide /usr/bin/hsdb (hsdb) in auto mode
update-alternatives: using /usr/lib/jvm/java-8-openjdk-amd64/bin/idlj to provide /usr/bin/idlj (idlj) in auto mode
```

```
/sbin/ldconfig.real: /usr/local/lib/libtbbbind.so.3 is not a symbolic link

/sbin/ldconfig.real: /usr/local/lib/libtbbbind_2_5.so.3 is not a symbolic link

/sbin/ldconfig.real: /usr/local/lib/libtbbmalloc_proxy.so.2 is not a symbolic link

/sbin/ldconfig.real: /usr/local/lib/libtbb.so.12 is not a symbolic link

/sbin/ldconfig.real: /usr/local/lib/libtbbmalloc.so.2 is not a symbolic link

Collecting pyspark
  Downloading pyspark-3.5.0.tar.gz (316.9 MB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 316.9/316.9 MB 2.9 MB/s eta 0:00:00
  Preparing metadata (setup.py) ... done
Requirement already satisfied: py4j==0.10.9.7 in /usr/local/lib/python3.10/dist-packages (from pyspark) (0.10.9.7)
Building wheels for collected packages: pyspark
  Building wheel for pyspark (setup.py) ... done
  Created wheel for pyspark: filename=pyspark-3.5.0-py2.py3-none-any.whl size=317425344 sha256=491f0533662fb82ca85051ca99a7f0a3a2
  Stored in directory: /root/.cache/pip/wheels/41/4e/10/c2cf2467f71c678cfc8a6b9ac9241e5e44a01940da8fbb17fc
Successfully built pyspark
Installing collected packages: pyspark
```

How to initialize Spark

```
from pyspark.sql import SparkSession

APP_NAME = "CAP-lab3"
SPARK_URL = "local[*]"
spark = SparkSession.builder.appName(APP_NAME).master(SPARK_URL).getOrCreate()
sc = spark.sparkContext
```

First Part: RDDs

Basic operations

Parallelize & collect

It creates a RDD out of a list or array. Second argument indicates the number of pieces of the RDD

```
array = sc.parallelize([1,2,3,4,5,6,7,8,9,10], 2)
array
```

```
ParallelCollectionRDD[0] at readRDDFromFile at PythonRDD.scala:289
```

```
import numpy as np
randomSamples = sc.parallelize(np.random.randn(100))
randomSamples
```

```
ParallelCollectionRDD[1] at readRDDFromFile at PythonRDD.scala:289
```

Cool, RDDs can not be printed...

Of course, RDDs can not be printed unless they are reduced

```
print(array.collect())
print(randomSamples.collect())
```

```
[1, 2, 3, 4, 5, 6, 7, 8, 9, 10]
[2.0095827346650865, 0.5280512556951231, 0.018622308184655448, -1.0367695755158028, 0.22421941859055075, 0.25463585450681786, -0.226
```

Spark uses lazy operations for everything, this means that nothing is evaluated until an action, a reduce operation normally, is performed. The basic reduce operation is collect, which returns the whole RDD (i.e. no reduction is performed).

Other ways of loading data

```
import requests

request = requests.get("https://gist.githubusercontent.com/jsdario/6d6c69398cb0c73111e49f1218960f79/raw/8d4fc4548d437e2a7203a5aeeace5477f
with open("elquijote.txt", "wb") as f:
  f.write(request.content)
```

```
    f.write(request.content)

quijote = sc.textFile("elquijote.txt")
quijote.take(10)
```

```
['DON QUIJOTE DE LA MANCHA',
 'Miguel de Cervantes Saavedra',
 '',
 'PRIMERA PARTE',
 'CAPÍTULO 1: Que trata de la condición y ejercicio del famoso hidalgo D. Quijote de la Mancha',
 'En un lugar de la Mancha, de cuyo nombre no quiero acordarme, no ha mucho tiempo que vivía un hidalgo de los de lanza en
astillero, adarga antigua, rocín flaco y galgo corredor. Una olla de algo más vaca que carnero, salpicón las más noches, duelos y
quebrantos los sábados, lentejas los viernes, algún palomino de añadidura los domingos, consumían las tres partes de su hacienda.
El resto della concluían sayo de velarte, calzas de velludo para las fiestas con sus pantuflos de lo mismo, los días de entre
semana se honraba con su vellori de lo más fino. Tenía en su casa una ama que pasaba de los cuarenta, y una sobrina que no llegaba
a los veinte, y un mozo de campo y plaza, que así ensillaba el rocín como tomaba la podadera. Frisaba la edad de nuestro hidalgo
con los cincuenta años, era de complexión recia, seco de carnes, enjuto de rostro; gran madrugador y amigo de la caza. Quieren
decir que tenía el sobrenombre de Quijada o Quesada (que en esto hay alguna diferencia en los autores que deste caso escriben),
aunque por conjeturas verosímiles se deja entender que se llama Quijana; pero esto importa poco a nuestro cuento; basta que en la
narración dél no se salga un punto de la verdad. Es, pues, de saber, que este sobredicho hidalgo, los ratos que estaba ocioso (que
eran los más del año) se daba a leer libros de caballerías con tanta afición y gusto, que olvidó casi de todo punto el ejercicio de
la caza, y aun la administración de su hacienda; y llegó a tanto su curiosidad y desatino en esto, que vendió muchas hanegas de
tierra de sembradura, para comprar libros de caballerías en que leer; y así llevó a su casa todos cuantos pudo haber dellos; y de
todos ningunos le parecían tan bien como los que compuso el famoso Feliciano de Silva: porque la claridad de su prosa, y aquellas
intrincadas razones suyas, le parecían de perlas; y más cuando llegaba a leer aquellos requiebros y cartas de desafío, donde en
muchas partes hallaba escrito: la razón de la sinrazón que a mi razón se hace, de tal manera mi razón enflaquece, que con razón me
quejo de la vuestra fermosura, y también cuando leía: los altos cielos que de vuestra divinidad divinamente con las estrellas se
fortifican, y os hacen merecedora del merecimiento que merece la vuestra grandeza. Con estas y semejantes razones perdía el pobre
caballero el juicio, y desvelábase por entenderlas, y desentrañarles el sentido, que no se lo sacara, ni las entendiera el mismo
Aristóteles, si resucitara para sólo ello. No estaba muy bien con las heridas que don Belianis daba y recibía, porque se imaginaba
que por grandes maestros que le hubiesen curado, no dejaría de tener el rostro y todo el cuerpo lleno de cicatrices y señales; pero
con todo alababa en su autor aquel acabar su libro con la promesa de aquella inacabable aventura, y muchas veces le vino deseo de
tomar la pluma, y darle fin al pie de la letra como allí se promete; y sin duda alguna lo hiciera, y aun saliera con ello, si otros
mayores y continuos pensamientos no se lo estorbaran.',
 'Tuvo muchas veces competencia con el cura de su lugar (que era hombre docto graduado en Sigüenza), sobre cuál había sido mejor
caballero, Palmerín de Inglaterra o Amadís de Gaula; mas maese Nicolás, barbero del mismo pueblo, decía que ninguno llegaba al
caballero del Febo, y que si alguno se le podía comparar, era don Galaor, hermano de Amadís de Gaula, porque tenía muy acomodada
condición para todo; que no era caballero melindroso, ni tan llorón como su hermano, y que en lo de la valentía no le iba en
zaga.',
 'En resolución, él se enfrascó tanto en su lectura, que se le pasaban las noches leyendo de claro en claro, y los días de turbio
en turbio, y así, del poco dormir y del mucho leer, se le secó el cerebro, de manera que vino a perder el juicio. Llenósele la
fantasía de todo aquello que leía en los libros, así de encantamientos, como de pendencias, batallas, desafíos, heridas,
requiebros, amores, tormentas y disparates imposibles, y asentósele de tal modo en la imaginación que era verdad toda aquella
máquina de aquellas soñadas invenciones que leía, que para él no había otra',
 'historia más cierta en el mundo.',
 'Decía él, que el Cid Ruy Díaz había sido muy buen caballero; pero que no tenía que ver con el caballero de la ardiente espada,
que de sólo un revés había partido por medio dos fieros y descomunales gigantes. Mejor estaba con Bernardo del Carpio, porque en
Roncesvalle había muerto a Roldán el encantado, valiéndose de la industria de Hércules, cuando ahogó a Anteo, el hijo de la Tierra,
entre los brazos. Decía mucho bien del gigante Morgante, porque con ser de aquella generación gigantesca, que todos son soberbios y
descomedidos, él solo era afable y bien criado; pero sobre todos estaba bien con Reinaldos de Montalbán, y más cuando le veía salir
de su castillo y robar cuantos topaba, y cuando en Allende robó aquel ídolo de Mahoma, que era todo de oro, según dice su historia.
Diera él, por dar una mano de coces al traidor de Galalón, al ama que tenía y aun a su sobrina de añadidura.']
```

Here, you can see both a method to load a text file line per line and a another reduction operation.

```
quijote.take
```

```
<bound method RDD.take of elquijote.txt MapPartitionsRDD[3] at textFile at NativeMethodAccessorImpl.java:0>
```

Transformations

Let's review all the transformation that can be performed to data.

```
from numpy.random.mtrand import sample
charsPerLine = quijote.map(lambda s: len(s))
allWords = quijote.flatMap(lambda s: s.split())
allWordsNoArticles = allWords.filter(lambda a: a.lower() not in ["el", "la"])
allWordsUnique = allWords.map(lambda s: s.lower()).distinct()
sampleWords = allWords.sample(withReplacement=True, fraction=0.2, seed=666)
sampleUnique = allWordsUnique.sample(withReplacement=True, fraction=0.2, seed=666)
weirdSampling = sampleWords.union(sampleUnique)
```

```
# Map devuelve el mismo número de elementos que la entrada (K = N)
print(f"Quijote_count: {quijote.count()}\nChars_per_line: {charsPerLine.count()}")
```

```
Quijote_count: 2186
Chars_per_line: 2186
```

```
# Flatmap aumenta resultado por cada elemento de entrada (K >= N)
print(f"Quijote_count: {quijote.count()}\nAll_words: {allWords.count()}")
```

```
Quijote_count: 2186
All_words: 187018
```

```
# Filter filtra elementos por los que normalmente (K <= N, pero puede ser K = N o K = 0 si todos pasan o ninguno pasa)
print(f"All_words: {allWords.count()}\nAll_words_no_articles: {allWordsNoArticles.count()}")
```

```
All_words: 187018
All_words_no_articles: 178167
```

```
# Distinct borra elementos repetidos (K <= N o K = N si no hay)
print(f"All_words: {allWords.count()}\nAll_words_unique: {allWordsUnique.count()}")
```

```
All_words: 187018
All_words_unique: 22211
```

```
# Sample selecciona una muestra aleatoria (K <= N)
print(f"All_words: {allWords.count()}\nSample_words: {sampleWords.count()}")
```

```
All_words: 187018
Sample_words: 37251
```

```
# Union une dos RDDs (K = N1 + N2)
print(f"Sample_unique: {sampleUnique.count()}\nSample_words: {sampleWords.count()}\nWeirdSampling: {weirdSampling.count()}")
```

```
Sample_unique: 4430
Sample_words: 37251
WeirdSampling: 41681
```

```
allWordsUnique.take(10)
```

```
['don',
 'mancha',
 'saavedra',
 'primera',
 'parte',
 '1:',
 'que',
 'condición',
 'y',
 'del']
```

```
charsPerLine.take(10)
```

```
[24, 28, 0, 13, 94, 3080, 526, 595, 33, 918]
```

```
allWords.take(10)
```

```
['DON',
 'QUIJOTE',
 'DE',
 'LA',
 'MANCHA',
 'Miguel',
 'de',
 'Cervantes',
 'Saavedra',
 'PRIMERA']
```

```
allWordsNoArticles.take(10)
```

```
['DON',
 'QUIJOTE',
 'DE',
 'MANCHA',
 'Miguel',
 'de',
 'Cervantes',
 'Saavedra',
 'PRIMERA',
 'PARTE']
```

```
allWordsUnique.take(10)
```

```
['don',
 'mancha',
 'saavedra',
 'primera',
 'parte',
 '1:',
 'que',
```

```
 'condición',
 'y',
 'del']
```

```
sampleWords.take(10)
```

```
['DON', 'Que', 'ejercicio', 'del', 'D.', 'la', 'no', 'ha', 'hidalgo', 'de']
```

```
weirdSampling.take(10)
```

```
['DON', 'Que', 'ejercicio', 'del', 'D.', 'la', 'no', 'ha', 'hidalgo', 'de']
```

Assignment question

Explain the use and purpose of each operation above.

Comment also on the size of the resulting RDD in terms of the size of the original RDD, e.g. if original RDD is of size N, then rdd.filter() is of size $K \leq N$

Answer:

- map : Aplica sobre la totalidad de los elementos de un RDD una función, la cual se le pasa como argumento. El resultado es un nuevo RDD.
- flatMap : similar a la función map, pero en lugar de producir un solo valor para cada elemento de entrada, puede producir cero, uno o múltiples valores de salida por cada uno de entrada. Todos estos valores se aplanan en un solo conjunto de resultados.
- filter : Como dice su nombre, filtra los elementros del RDD y devuelve el RDD 'limpio'. El filtro a usar se especifica por parámetro.
- distinct : Borra todos los elementos repetidos dentro de un RDD
- sample : Adquiere del RDD una muestra aleatoria.
- union : El union logico de conjuntos, une un RDD con otro(s)

Actions

```
numLines = quijote.count()
numChars = charsPerLine.reduce(lambda a,b: a+b) # also charsPerLine.sum()
sortedWordsByLength = allWordsNoArticles.takeOrdered(10, key=lambda x: -len(x))
numLines, numChars, sortedWordsByLength
```

```
(2186,
 1036211,
 ['procuremos.Levántate,',
  'extraordinariamente,',
  'estrechísimamente,',
  'convirtiéndoseles',
  'entretenimientos,',
  'inadvertidamente.',
  'cortesísimamente',
  'Agredeciéronselo',
  'Pintiquiniestra,',
  'entretenimiento,'])
```

Assignment question

Explain the use and purpose of each action above.

Implement the count operation using reduce as the unique option. You can use transformations. Is it possible to achieve a solution without any transformation? Does it make sense?

Answer:

- count : Devuelve el numero de elementos del RDD
- reduce : Es una operación de acción que combina los elementos de un RDD utilizando una función de reducción y devuelve el resultado final
- takeOrdered : Es una operación de acción que devuelve los primeros n elementos de un RDD, según un orden específico. Puedes especificar el número 'n' y opcionalmente proporcionar una función de ordenación personalizada para determinar el criterio de orden.

```
count = quijote.map(lambda s: 1).reduce(lambda a,b: a+b)
count
```

```
2186
```

```
count = quijote.reduce(lambda a,b: a + b)
count
```

```
'DON QUIJOTE DE LA MANCHAMiguel de Cervantes SaavedraPRIMERA PARTECAPÍTULO 1: Que tr
ata de la condición y ejercicio del famoso hidalgo D. Quijote de la ManchaEn un luga
r de la Mancha, de cuyo nombre no quiero acordarme, no ha mucho tiempo que vivía un
hidalgo de los de lanza en astillero, adarga antigua, rocín flaco y galgo corredor.
Una olla de algo más vaca que carnero, salpicón las más noches, duelos y quebrantos
```

Key-Value RDDs

```
import re
allWords = allWords.flatMap(lambda w: re.sub("""[;|:|\.|,|-|–|"|'|\s"""," ", w.lower()).split(" ")).filter(lambda a: len(a)>0)
allWords2 = sc.parallelize(requests.get("https://gist.githubusercontent.com/jsdario/9d871ed773c81bf217f57d1db2d2503f/raw/585de69b0631c805
allWords2 = allWords2.flatMap(lambda w: re.sub("""[;|:|\.|,|-|–|"|'|\s"""," ", w.decode("utf8").lower()).split(" ")).filter(lambda a: len(
```

```
allWords.take(10), allWords2.take(10)
```

```
(['don',
  'quijote',
  'de',
  'la',
  'mancha',
  'miguel',
  'de',
  'cervantes',
  'saavedra',
  'primera'],
 ['don',
  'quijote',
  'de',
  'la',
  'mancha',
  'miguel',
  'de',
  'cervantes',
  'saavedra',
  'segunda'])
```

Next, we move to more interesting operations that involve key-value RDDs. Key-value RDDs are a special kind of RDDs where each is element is a tuple (K,V) where K is the key and V the value.

```
words = allWords.map(lambda e: (e,1))
words2 = allWords2.map(lambda e: (e,1))

words.take(10)
```

```
[('don', 1),
 ('quijote', 1),
 ('de', 1),
 ('la', 1),
 ('mancha', 1),
 ('miguel', 1),
 ('de', 1),
 ('cervantes', 1),
 ('saavedra', 1),
 ('primera', 1)]
```

How to manipulate K-V RDDs

```
frequencies = words.reduceByKey(lambda a,b: a+b)
frequencies2 = words2.reduceByKey(lambda a,b: a+b)
frequencies.takeOrdered(10, key=lambda a: -a[1])
```

```
[('que', 10705),
 ('de', 9033),
 ('y', 8668),
 ('la', 5015),
 ('a', 4815),
 ('en', 4046),
 ('el', 3857),
```

```
       ('no', 3083),
       ('se', 2382),
       ('los', 2148)]


res = words.groupByKey().takeOrdered(10, key=lambda a: -len(a))
res # To see the content, res[i][1].data
```

```
     [('don', <pyspark.resultiterable.ResultIterable at 0x79dd8fc71900>),
      ('mancha', <pyspark.resultiterable.ResultIterable at 0x79dd8fc72fe0>),
      ('saavedra', <pyspark.resultiterable.ResultIterable at 0x79dd8fc734f0>),
      ('primera', <pyspark.resultiterable.ResultIterable at 0x79dd8fc736d0>),
      ('parte', <pyspark.resultiterable.ResultIterable at 0x79dd8fc73670>),
      ('1', <pyspark.resultiterable.ResultIterable at 0x79dd8fc73700>),
      ('que', <pyspark.resultiterable.ResultIterable at 0x79dd8fc73790>),
      ('condición', <pyspark.resultiterable.ResultIterable at 0x79dd8fc737f0>),
      ('y', <pyspark.resultiterable.ResultIterable at 0x79dd8fc73850>),
      ('del', <pyspark.resultiterable.ResultIterable at 0x79dd8fc738b0>)]
```

```
# GroupByKey implementation
groupBK = words.groupByKey()
res2 = groupBK.takeOrdered(10, key=lambda a: -len(a))
res2
```

```
     [('don', <pyspark.resultiterable.ResultIterable at 0x79dd8fc73b50>),
      ('mancha', <pyspark.resultiterable.ResultIterable at 0x79dd8fc73550>),
      ('saavedra', <pyspark.resultiterable.ResultIterable at 0x79dd8fc731f0>),
      ('primera', <pyspark.resultiterable.ResultIterable at 0x79dd8fc73340>),
      ('parte', <pyspark.resultiterable.ResultIterable at 0x79dd8fc73250>),
      ('1', <pyspark.resultiterable.ResultIterable at 0x79dd8fc732e0>),
      ('que', <pyspark.resultiterable.ResultIterable at 0x79dd8fc73160>),
      ('condición', <pyspark.resultiterable.ResultIterable at 0x79dd8fc73100>),
      ('y', <pyspark.resultiterable.ResultIterable at 0x79dd8fc73520>),
      ('del', <pyspark.resultiterable.ResultIterable at 0x79dd8fc70c10>)]
```

```
joinFreq = frequencies.join(frequencies2)
joinFreq.take(10)
```

```
     [('don', (1072, 1606)),
      ('mancha', (50, 101)),
      ('saavedra', (2, 1)),
      ('primera', (39, 55)),
      ('parte', (178, 158)),
      ('1', (1, 1)),
      ('que', (10705, 10040)),
      ('condición', (33, 39)),
      ('y', (8668, 9650)),
      ('del', (1128, 1344))]
```

```
print(f"Words_count: {words.count()}\nFreq_count: {frequencies.count()}\nGroupByKey_count: {groupBK.count()}\n")

print(f"Freq_count: {frequencies.count()}\nFreq2_count: {frequencies2.count()}\nJoin_freqs_count: {joinFreq.count()}")
```

```
     Words_count: 187045
     Freq_count: 15262
     GroupByKey_count: 15262

     Freq_count: 15262
     Freq2_count: 17229
     Join_freqs_count: 8155
```

```
joinFreq.map(lambda e: (e[0], (e[1][0] - e[1][1])/(e[1][0] + e[1][1]))).takeOrdered(10, lambda v: -v[1]), joinFreq.map(lambda e: (e[0], (
```

```
     ([('bacía', 0.9393939393939394),
       ('venia', 0.9230769230769231),
       ('hermandad', 0.9),
       ('andrés', 0.8823529411764706),
       ('peña', 0.8823529411764706),
       ('micomicona', 0.8823529411764706),
       ('barca', 0.875),
       ('novela', 0.875),
       ('yerme', 0.875),
       ('acertó', 0.8666666666666667)],
      [('teresa', -0.9767441860465116),
       ('roque', -0.96),
       ('refranes', -0.9375),
       ('condesa', -0.9333333333333333),
       ('leones', -0.9333333333333333),
       ('gobernadores', -0.9166666666666666),
       ('lacayo', -0.9166666666666666),
       ('visorrey', -0.9130434782608695),
       ('antonio', -0.9076923076923077),
       ('zaragoza', -0.9047619047619048)])
```

Assignment question

Explain the use and purpose of each action above.

Implement the frequency with groupByKey and transformations.

Which of the two following cells is more efficient?

Answer:

- ReduceByKey : Combina los valores asociados con la misma clave en un RDD aplicando una función de reducción. La función de reducción debe ser asociativa y conmutativa
- GroupByKey : Agrupa los valores asociados con la misma clave en un RDD, creando una lista de valores para cada clave. Aunque agrupa los valores, no realiza ninguna operación de reducción en los grupos resultantes
- Join : Combina dos RDDs basándose en sus claves, similar al JOIN de SQL

```
joinFreq.map(lambda e: (e[0], (e[1][0] - e[1][1])/(e[1][0] + e[1][1]))).takeOrdered(10, lambda v: -v[1]), joinFreq.map(lambda e: (e[0],
```

```
([('bacía', 0.9393939393939394),
  ('venia', 0.9230769230769231),
  ('hermandad', 0.9),
  ('andrés', 0.8823529411764706),
  ('peña', 0.8823529411764706),
  ('micomicona', 0.8823529411764706),
  ('barca', 0.875),
  ('novela', 0.875),
  ('yerme', 0.875),
  ('acertó', 0.8666666666666667)],
 [('teresa', -0.9767441860465116),
  ('roque', -0.96),
  ('refranes', -0.9375),
  ('condesa', -0.9333333333333333),
  ('leones', -0.9333333333333333),
  ('gobernadores', -0.9166666666666666),
  ('lacayo', -0.9166666666666666),
  ('visorrey', -0.9130434782608695),
  ('antonio', -0.9076923076923077),
  ('zaragoza', -0.9047619047619048)])
```

```
result = joinFreq.map(lambda e: (e[0], (e[1][0] - e[1][1])/(e[1][0] + e[1][1])))
result.takeOrdered(10, lambda v: -v[1]), result.takeOrdered(10, lambda v: +v[1])
```

```
([('bacía', 0.9393939393939394),
  ('venia', 0.9230769230769231),
  ('hermandad', 0.9),
  ('andrés', 0.8823529411764706),
  ('peña', 0.8823529411764706),
  ('micomicona', 0.8823529411764706),
  ('barca', 0.875),
  ('novela', 0.875),
  ('yerme', 0.875),
  ('acertó', 0.8666666666666667)],
 [('teresa', -0.9767441860465116),
  ('roque', -0.96),
  ('refranes', -0.9375),
  ('condesa', -0.9333333333333333),
  ('leones', -0.9333333333333333),
  ('gobernadores', -0.9166666666666666),
  ('lacayo', -0.9166666666666666),
  ('visorrey', -0.9130434782608695),
  ('antonio', -0.9076923076923077),
  ('zaragoza', -0.9047619047619048)])
```

Optimizations and final notes

Optimizing the data movement around the cluster

One of the main issues could be that if data after an operation is not balanced, we may not be using the cluster properly. For that purpose, we have two operations

```
result.coalesce(numPartitions=2) # Avoids the data movement, so it tries to balance inside each machine
result.repartition(numPartitions=2) # We don't care about data movement, this balance the whole thing to ensure all machines are used
```

```
MapPartitionsRDD[82] at coalesce at NativeMethodAccessorImpl.java:0
```

Persistance for intermediate operations

In contrast to Hadoop, intermediate RDDs are not preserved, each time we use an action/reduction, the whole data pipeline is executed from the datasources. To avoid this:

```
result.take(10)
allWords.cache() # allWords RDD must  stay in memory after computation, we made a checkpoint (well, it's a best effort, so must might be
result.take(10)
```

```
[('don', -0.19940253920836445),
 ('mancha', -0.33774834437086093),
 ('saavedra', 0.3333333333333333),
 ('primera', -0.1702127659574468),
 ('parte', 0.05952380952380952),
 ('1', 0.0),
 ('que', 0.03205591708845505),
 ('condición', -0.08333333333333333),
 ('y', -0.05360847254067038),
 ('del', -0.08737864077669903)]
```

```
from pyspark import StorageLevel
# https://spark.apache.org/docs/2.2.0/rdd-programming-guide.html#rdd-persistence
allWords2.persist(StorageLevel.MEMORY_AND_DISK) # Now it will be preserved on disk also
```

```
PythonRDD[86] at RDD at PythonRDD.scala:53
```

```
allWords2_rep = allWords2.repartition(numPartitions=5)
```

```
allWords_coalesce1 = allWords2.coalesce(2, shuffle=True)
allWords_coalesce2 = allWords2.coalesce(4, shuffle=True)
```

```
allWords_coalesce1.count()
```

```
197777
```

```
!rm -rf palabras_parte2
allWords2.saveAsTextFile("palabras_parte2")
```

```
!ls palabras_parte2
```

```
part-00000  part-00001  _SUCCESS
```

Assignment question

Before saving with saveAsTextFile, use coalesce with different values. What's the difference in the previous ls?

Answer: La función coalesce parte el RDD para paralelizarlo y a parte, puede rendomizar los datos del mismo con el parametro shuffle.

Global variables

There are two kind of global variables, read-only and write-only.

```
articles = sc.broadcast(["el", "la"])
articles.value
```

```
['el', 'la']
```

Broadcast variables are read-only. They help us to avoid local variables of the closures (the functions we use inside map, reduce, ...) to be transferred in every single Spark operation. In that way, they are only transferred only once.

```
acc = sc.accumulator(0)
def incrementar(x):
  global acc
  acc += x

allWords.map(lambda l:1).foreach(incrementar)
acc
```

```
Accumulator<id=0, value=187045>
```

Write-only variables can be also declared and initialized, but they can not be read since reading will force a complete synchronization of the cluster.

Second part: Spark SQL

Next, we will do a short review of the high-level API of Spark

```
import pandas as pd

size = int(1e6)
def loadRedditToPandas(subreddit=None, size=size):
  if subreddit is not None:
    redditData = requests.get(f"https://api.pushshift.io/reddit/search/submission/?subreddit={subreddit}&sort=desc&sort_type=created_ut
  else:
    redditData = requests.get(f"https://api.pushshift.io/reddit/search/submission/?sort=desc&sort_type=created_utc&size={size:d}").jso
  print(redditData)
  return pd.DataFrame(redditData["data"])
!wget -O RS_2016_02_reduced http://mirai.ii.uam.es/RS_2016_02_reduced
```

```
--2023-11-19 19:35:04--  http://mirai.ii.uam.es/RS_2016_02_reduced
Resolving mirai.ii.uam.es (mirai.ii.uam.es)... 150.244.59.231
Connecting to mirai.ii.uam.es (mirai.ii.uam.es)|150.244.59.231|:80... connected.
HTTP request sent, awaiting response... 200 OK
Length: 176318960 (168M) [application/octet-stream]
Saving to: 'RS_2016_02_reduced'

RS_2016_02_reduced  100%[===================>] 168.15M  25.5MB/s    in 7.4s

2023-11-19 19:35:12 (22.9 MB/s) - 'RS_2016_02_reduced' saved [176318960/176318960]
```

```
#pdf = loadRedditToPandas() #Debido a los cambios en el API de reddit, ya no se pueden extraer datos sin hacer login
import json
import pandas as pd
def load():
  with open("RS_2016_02_reduced", "r") as f:
    for line in f:
      post = json.loads(line)
      yield post
pdf = pd.DataFrame(load())
pdf.head()
```

	from	created_utc	saved	num_comments	permali
0	NaN	1454284800	False	0	/r/Gear4Sale/comments/43lprb/wts_tech21_sansam
1	NaN	1454284800	False	1	/r/ImagesOfIllinois/comments/43lprc/fsft_mayac
2	NaN	1454284800	False	4	/r/customhearthstone/comments/43lprd/a_combina
3	NaN	1454284801	False	3	/r/Fireteams/comments/43lpre/ps4_lf2m_too_flaw
4	NaN	1454284801	False	2	/r/recordthis/comments/43lprg/feedback_the_sou

5 rows × 50 columns

```
attrs = ["author", "created_utc", "title", "subreddit", "selftext", "over_18"]
pdf[attrs].head()
```

	author	created_utc	title	subreddit
0	[deleted]	1454284800	[WTS] Tech21 SansAmp Bass Driver - $175 Shipped	Gear4Sale

```
df = spark.createDataFrame(pdf[attrs])
```

```
#pdf.selftext = pdf.selftext.apply(lambda e: str(e))


#df = spark.createDataFrame(pdf[attrs])
#df = spark.createDataFrame(df[attrs])
```

Basic operations

```
df.show()
df.createOrReplaceTempView("reddit")
```

```
+---------------+-----------+--------------------+----------------+--------------------+-------+
|         author|created_utc|               title|       subreddit|            selftext|over_18|
+---------------+-----------+--------------------+----------------+--------------------+-------+
|      [deleted]| 1454284800|[WTS] Tech21 Sans...|       Gear4Sale|           [deleted]|  false|
|     amici_ursi| 1454284800|"[FS/FT] Mayaca, ...| ImagesOfIllinois|                    |  false|
|        Buhadog| 1454284800|A combination of ...|customhearthstone|                    |  false|
|    DJOlYoungin| 1454284801|[PS4] LF2M ToO Fl...|       Fireteams|Leave psn and I'l...|  false|
|BrightenQuintin| 1454284801|[FEEDBACK] The So...|      recordthis|https://www.casti...|  false|
|         jptx82| 1454284802|The gloves bother...|mildlyinfuriating|                    |  false|
|       kahlil88| 1454284802|"questions" for m...|      Shoe0nHead|                    |  false|
|        OverSol| 1454284802|Visual Comparison...| BleachBraveSouls|After seeing the ...|  false|
|      [deleted]| 1454284802|hes tryin to turn...|       TheBarons|           [deleted]|  false|
|      [deleted]| 1454284802|If Charles Martel...|      metacanada|                    |  false|
|    Shadowwoods| 1454284802|      The Rocketship|        Noobtube|                    |  false|
|      [deleted]| 1454284802|Eat like your gra...|          canada|           [deleted]|  false|
|      [deleted]| 1454284802|[PS4] KF HM War p...|       Fireteams|           [deleted]|  false|
|       hxnterrr| 1454284802|I wish I could ca...|       blackops3|                    |  false|
| LoveGaugedGals| 1454284802|  wow what a phrame!|           ssbbw|                    |   true|
|      [deleted]| 1454284803|[Question] What i...|       jailbreak|           [deleted]|  false|
|MaconCountyLine| 1454284803|(Friendly Reminde...|        Sherlock|It will have been...|  false|
|      [deleted]| 1454284803|Kyber Crystals us...|        StarWars|Is there any cano...|  false|
|      [deleted]| 1454284804|DOWNLOAD BOOK "Wi...|  TekkitCreations|     Rebecca Jimenez|  false|
|       potzdamn| 1454284804|                Cruz|    FuckYeahWierd|                    |   true|
+---------------+-----------+--------------------+----------------+--------------------+-------+
only showing top 20 rows
```

Filtering

```
df.filter(~df.over_18).show()
```

```
+---------------+-----------+--------------------+----------------+--------------------+-------+
|         author|created_utc|               title|       subreddit|            selftext|over_18|
+---------------+-----------+--------------------+----------------+--------------------+-------+
|      [deleted]| 1454284800|[WTS] Tech21 Sans...|       Gear4Sale|           [deleted]|  false|
|     amici_ursi| 1454284800|"[FS/FT] Mayaca, ...| ImagesOfIllinois|                    |  false|
|        Buhadog| 1454284800|A combination of ...|customhearthstone|                    |  false|
|    DJOlYoungin| 1454284801|[PS4] LF2M ToO Fl...|       Fireteams|Leave psn and I'l...|  false|
|BrightenQuintin| 1454284801|[FEEDBACK] The So...|      recordthis|https://www.casti...|  false|
|         jptx82| 1454284802|The gloves bother...|mildlyinfuriating|                    |  false|
|       kahlil88| 1454284802|"questions" for m...|      Shoe0nHead|                    |  false|
|        OverSol| 1454284802|Visual Comparison...| BleachBraveSouls|After seeing the ...|  false|
|      [deleted]| 1454284802|hes tryin to turn...|       TheBarons|           [deleted]|  false|
|      [deleted]| 1454284802|If Charles Martel...|      metacanada|                    |  false|
|    Shadowwoods| 1454284802|      The Rocketship|        Noobtube|                    |  false|
|      [deleted]| 1454284802|Eat like your gra...|          canada|           [deleted]|  false|
|      [deleted]| 1454284802|[PS4] KF HM War p...|       Fireteams|           [deleted]|  false|
|       hxnterrr| 1454284802|I wish I could ca...|       blackops3|                    |  false|
|      [deleted]| 1454284803|[Question] What i...|       jailbreak|           [deleted]|  false|
|MaconCountyLine| 1454284803|(Friendly Reminde...|        Sherlock|It will have been...|  false|
|      [deleted]| 1454284803|Kyber Crystals us...|        StarWars|Is there any cano...|  false|
|      [deleted]| 1454284804|DOWNLOAD BOOK "Wi...|  TekkitCreations|     Rebecca Jimenez|  false|
|           onrv| 1454284804|Foo Fighters - Fe...|   Musicthemetime|                    |  false|
|        ranalog| 1454284805|Monthly 'Self Pro...|          analog|This thread is fo...|  false|
+---------------+-----------+--------------------+----------------+--------------------+-------+
only showing top 20 rows
```

```
spark.sql("select * from reddit where not over_18").show()
```

```
+---------------+-----------+--------------------+----------------+--------------------+-------+
|         author|created_utc|               title|       subreddit|            selftext|over_18|
+---------------+-----------+--------------------+----------------+--------------------+-------+
|      [deleted]| 1454284800|[WTS] Tech21 Sans...|       Gear4Sale|           [deleted]|  false|
|     amici_ursi| 1454284800|"[FS/FT] Mayaca, ...| ImagesOfIllinois|                    |  false|
|        Buhadog| 1454284800|A combination of ...|customhearthstone|                    |  false|
|    DJOlYoungin| 1454284801|[PS4] LF2M ToO Fl...|       Fireteams|Leave psn and I'l...|  false|
|BrightenQuintin| 1454284801|[FEEDBACK] The So...|      recordthis|https://www.casti...|  false|
|         jptx82| 1454284802|The gloves bother...|mildlyinfuriating|                    |  false|
|       kahlil88| 1454284802|"questions" for m...|       Shoe0nHead|                    |  false|
|        OverSol| 1454284802|Visual Comparison...| BleachBraveSouls|After seeing the ...|  false|
|      [deleted]| 1454284802|hes tryin to turn...|       TheBarons|           [deleted]|  false|
|      [deleted]| 1454284802|If Charles Martel...|      metacanada|                    |  false|
|    Shadowwoods| 1454284802|      The Rocketship|        Noobtube|                    |  false|
|      [deleted]| 1454284802|Eat like your gra...|          canada|           [deleted]|  false|
|      [deleted]| 1454284802|[PS4] KF HM War p...|       Fireteams|           [deleted]|  false|
|       hxnterrr| 1454284802|I wish I could ca...|       blackops3|                    |  false|
|      [deleted]| 1454284803|[Question] What i...|       jailbreak|           [deleted]|  false|
|MaconCountyLine| 1454284803|(Friendly Reminde...|        Sherlock|It will have been...|  false|
|      [deleted]| 1454284803|Kyber Crystals us...|        StarWars|Is there any cano...|  false|
|      [deleted]| 1454284804|DOWNLOAD BOOK "Wi...| TekkitCreations|     Rebecca Jimenez|  false|
|           onrv| 1454284804|Foo Fighters - Fe...|  Musicthemetime|                    |  false|
|        ranalog| 1454284805|Monthly 'Self Pro...|          analog|This thread is fo...|  false|
+---------------+-----------+--------------------+----------------+--------------------+-------+
only showing top 20 rows
```

```
df.where(~df.over_18).show()
```

```
+---------------+-----------+--------------------+----------------+--------------------+-------+
|         author|created_utc|               title|       subreddit|            selftext|over_18|
+---------------+-----------+--------------------+----------------+--------------------+-------+
|      [deleted]| 1454284800|[WTS] Tech21 Sans...|       Gear4Sale|           [deleted]|  false|
|     amici_ursi| 1454284800|"[FS/FT] Mayaca, ...| ImagesOfIllinois|                    |  false|
|        Buhadog| 1454284800|A combination of ...|customhearthstone|                    |  false|
|    DJOlYoungin| 1454284801|[PS4] LF2M ToO Fl...|       Fireteams|Leave psn and I'l...|  false|
|BrightenQuintin| 1454284801|[FEEDBACK] The So...|      recordthis|https://www.casti...|  false|
|         jptx82| 1454284802|The gloves bother...|mildlyinfuriating|                    |  false|
|       kahlil88| 1454284802|"questions" for m...|       Shoe0nHead|                    |  false|
|        OverSol| 1454284802|Visual Comparison...| BleachBraveSouls|After seeing the ...|  false|
|      [deleted]| 1454284802|hes tryin to turn...|       TheBarons|           [deleted]|  false|
|      [deleted]| 1454284802|If Charles Martel...|      metacanada|                    |  false|
|    Shadowwoods| 1454284802|      The Rocketship|        Noobtube|                    |  false|
|      [deleted]| 1454284802|Eat like your gra...|          canada|           [deleted]|  false|
|      [deleted]| 1454284802|[PS4] KF HM War p...|       Fireteams|           [deleted]|  false|
|       hxnterrr| 1454284802|I wish I could ca...|       blackops3|                    |  false|
|      [deleted]| 1454284803|[Question] What i...|       jailbreak|           [deleted]|  false|
|MaconCountyLine| 1454284803|(Friendly Reminde...|        Sherlock|It will have been...|  false|
|      [deleted]| 1454284803|Kyber Crystals us...|        StarWars|Is there any cano...|  false|
|      [deleted]| 1454284804|DOWNLOAD BOOK "Wi...| TekkitCreations|     Rebecca Jimenez|  false|
|           onrv| 1454284804|Foo Fighters - Fe...|  Musicthemetime|                    |  false|
|        ranalog| 1454284805|Monthly 'Self Pro...|          analog|This thread is fo...|  false|
+---------------+-----------+--------------------+----------------+--------------------+-------+
only showing top 20 rows
```

```
spark.sql("select * from reddit where not over_18").show()
```

```
+---------------+-----------+--------------------+----------------+--------------------+-------+
|         author|created_utc|               title|       subreddit|            selftext|over_18|
+---------------+-----------+--------------------+----------------+--------------------+-------+
|      [deleted]| 1454284800|[WTS] Tech21 Sans...|       Gear4Sale|           [deleted]|  false|
|     amici_ursi| 1454284800|"[FS/FT] Mayaca, ...| ImagesOfIllinois|                    |  false|
|        Buhadog| 1454284800|A combination of ...|customhearthstone|                    |  false|
|    DJOlYoungin| 1454284801|[PS4] LF2M ToO Fl...|       Fireteams|Leave psn and I'l...|  false|
|BrightenQuintin| 1454284801|[FEEDBACK] The So...|      recordthis|https://www.casti...|  false|
|         jptx82| 1454284802|The gloves bother...|mildlyinfuriating|                    |  false|
|       kahlil88| 1454284802|"questions" for m...|       Shoe0nHead|                    |  false|
|        OverSol| 1454284802|Visual Comparison...| BleachBraveSouls|After seeing the ...|  false|
|      [deleted]| 1454284802|hes tryin to turn...|       TheBarons|           [deleted]|  false|
|      [deleted]| 1454284802|If Charles Martel...|      metacanada|                    |  false|
|    Shadowwoods| 1454284802|      The Rocketship|        Noobtube|                    |  false|
|      [deleted]| 1454284802|Eat like your gra...|          canada|           [deleted]|  false|
|      [deleted]| 1454284802|[PS4] KF HM War p...|       Fireteams|           [deleted]|  false|
|       hxnterrr| 1454284802|I wish I could ca...|       blackops3|                    |  false|
|      [deleted]| 1454284803|[Question] What i...|       jailbreak|           [deleted]|  false|
|MaconCountyLine| 1454284803|(Friendly Reminde...|        Sherlock|It will have been...|  false|
|      [deleted]| 1454284803|Kyber Crystals us...|        StarWars|Is there any cano...|  false|
|      [deleted]| 1454284804|DOWNLOAD BOOK "Wi...| TekkitCreations|     Rebecca Jimenez|  false|
|           onrv| 1454284804|Foo Fighters - Fe...|  Musicthemetime|                    |  false|
|        ranalog| 1454284805|Monthly 'Self Pro...|          analog|This thread is fo...|  false|
+---------------+-----------+--------------------+----------------+--------------------+-------+
only showing top 20 rows
```

```
df.where("not over_18").show() # SQL syntax
```

```
+--------------+-----------+--------------------+----------------+--------------------+-------+
|        author|created_utc|               title|       subreddit|            selftext|over_18|
+--------------+-----------+--------------------+----------------+--------------------+-------+
|     [deleted]| 1454284800|[WTS] Tech21 Sans...|       Gear4Sale|           [deleted]|  false|
|    amici_ursi| 1454284800|"[FS/FT] Mayaca, ...| ImagesOfIllinois|                    |  false|
|       Buhadog| 1454284800|A combination of ...|customhearthstone|                    |  false|
|   DJOlYoungin| 1454284801|[PS4] LF2M ToO Fl...|       Fireteams|Leave psn and I'l...|  false|
|BrightenQuintin| 1454284801|[FEEDBACK] The So...|      recordthis|https://www.casti...|  false|
|         jptx82| 1454284802|The gloves bother...|mildlyinfuriating|                    |  false|
|       kahlil88| 1454284802|"questions" for m...|       Shoe0nHead|                    |  false|
|        OverSol| 1454284802|Visual Comparison...| BleachBraveSouls|After seeing the ...|  false|
|     [deleted]| 1454284802|hes tryin to turn...|       TheBarons|           [deleted]|  false|
|     [deleted]| 1454284802|If Charles Martel...|      metacanada|                    |  false|
|   Shadowwoods| 1454284802|      The Rocketship|        Noobtube|                    |  false|
|     [deleted]| 1454284802|Eat like your gra...|          canada|           [deleted]|  false|
|     [deleted]| 1454284802|[PS4] KF HM War p...|       Fireteams|           [deleted]|  false|
|       hxnterrr| 1454284802|I wish I could ca...|       blackops3|                    |  false|
|     [deleted]| 1454284803|[Question] What i...|       jailbreak|           [deleted]|  false|
|MaconCountyLine| 1454284803|(Friendly Reminde...|        Sherlock|It will have been...|  false|
|     [deleted]| 1454284803|Kyber Crystals us...|        StarWars|Is there any cano...|  false|
|     [deleted]| 1454284804|DOWNLOAD BOOK "Wi...| TekkitCreations|     Rebecca Jimenez|  false|
|          onrv| 1454284804|Foo Fighters - Fe...|  Musicthemetime|                    |  false|
|        ranalog| 1454284805|Monthly 'Self Pro...|          analog|This thread is fo...|  false|
+--------------+-----------+--------------------+----------------+--------------------+-------+
only showing top 20 rows
```

```
spark.sql("select * from reddit where not over_18").show()
```

```
+--------------+-----------+--------------------+----------------+--------------------+-------+
|        author|created_utc|               title|       subreddit|            selftext|over_18|
+--------------+-----------+--------------------+----------------+--------------------+-------+
|     [deleted]| 1454284800|[WTS] Tech21 Sans...|       Gear4Sale|           [deleted]|  false|
|    amici_ursi| 1454284800|"[FS/FT] Mayaca, ...| ImagesOfIllinois|                    |  false|
|       Buhadog| 1454284800|A combination of ...|customhearthstone|                    |  false|
|   DJOlYoungin| 1454284801|[PS4] LF2M ToO Fl...|       Fireteams|Leave psn and I'l...|  false|
|BrightenQuintin| 1454284801|[FEEDBACK] The So...|      recordthis|https://www.casti...|  false|
|         jptx82| 1454284802|The gloves bother...|mildlyinfuriating|                    |  false|
|       kahlil88| 1454284802|"questions" for m...|       Shoe0nHead|                    |  false|
|        OverSol| 1454284802|Visual Comparison...| BleachBraveSouls|After seeing the ...|  false|
|     [deleted]| 1454284802|hes tryin to turn...|       TheBarons|           [deleted]|  false|
|     [deleted]| 1454284802|If Charles Martel...|      metacanada|                    |  false|
|   Shadowwoods| 1454284802|      The Rocketship|        Noobtube|                    |  false|
|     [deleted]| 1454284802|Eat like your gra...|          canada|           [deleted]|  false|
|     [deleted]| 1454284802|[PS4] KF HM War p...|       Fireteams|           [deleted]|  false|
|       hxnterrr| 1454284802|I wish I could ca...|       blackops3|                    |  false|
|     [deleted]| 1454284803|[Question] What i...|       jailbreak|           [deleted]|  false|
|MaconCountyLine| 1454284803|(Friendly Reminde...|        Sherlock|It will have been...|  false|
|     [deleted]| 1454284803|Kyber Crystals us...|        StarWars|Is there any cano...|  false|
|     [deleted]| 1454284804|DOWNLOAD BOOK "Wi...| TekkitCreations|     Rebecca Jimenez|  false|
|          onrv| 1454284804|Foo Fighters - Fe...|  Musicthemetime|                    |  false|
|        ranalog| 1454284805|Monthly 'Self Pro...|          analog|This thread is fo...|  false|
+--------------+-----------+--------------------+----------------+--------------------+-------+
only showing top 20 rows
```

Operations

```
df.select(df.created_utc * 2).show()
```

```
+-----------------+
|(created_utc * 2)|
+-----------------+
|       2908569600|
|       2908569600|
|       2908569600|
|       2908569602|
|       2908569602|
|       2908569604|
|       2908569604|
|       2908569604|
|       2908569604|
|       2908569604|
|       2908569604|
|       2908569604|
|       2908569604|
|       2908569604|
|       2908569604|
|       2908569604|
|       2908569606|
|       2908569606|
|       2908569606|
|       2908569608|
|       2908569608|
+-----------------+
only showing top 20 rows
```

```
spark.sql("select created_utc * 2 from reddit").show()
```

```
+-----------------+
|(created_utc * 2)|
+-----------------+
|       2908569600|
|       2908569600|
|       2908569600|
|       2908569602|
|       2908569602|
|       2908569604|
|       2908569604|
|       2908569604|
|       2908569604|
|       2908569604|
|       2908569604|
|       2908569604|
|       2908569604|
|       2908569604|
|       2908569606|
|       2908569606|
|       2908569606|
|       2908569608|
|       2908569608|
+-----------------+
only showing top 20 rows
```

```
from pyspark.sql.functions import log
df.select(log(df.created_utc * 2)).show()
```

```
+---------------------+
|ln((created_utc * 2))|
+---------------------+
|   21.790927250889528|
|   21.790927250889528|
|   21.790927250889528|
|   21.790927251577152|
|   21.790927251577152|
|   21.790927252264776|
|   21.790927252264776|
|   21.790927252264776|
|   21.790927252264776|
|   21.790927252264776|
|   21.790927252264776|
|   21.790927252264776|
|   21.790927252264776|
|   21.790927252264776|
|   21.790927252952397|
|   21.790927252952397|
|   21.790927252952397|
|    21.79092725364002|
|    21.79092725364002|
+---------------------+
only showing top 20 rows
```

```
spark.sql("select LN(created_utc * 2) from reddit").show()
```

```
+---------------------+
|ln((created_utc * 2))|
+---------------------+
|   21.790927250889528|
|   21.790927250889528|
|   21.790927250889528|
|   21.790927251577152|
|   21.790927251577152|
|   21.790927252264776|
|   21.790927252264776|
|   21.790927252264776|
|   21.790927252264776|
|   21.790927252264776|
|   21.790927252264776|
|   21.790927252264776|
|   21.790927252264776|
|   21.790927252264776|
|   21.790927252952397|
|   21.790927252952397|
|   21.790927252952397|
|    21.79092725364002|
|    21.79092725364002|
+---------------------+
```

```
only showing top 20 rows
```

Aggregations

```
df.where("not over_18").groupby(["author", df.subreddit]).count().toPandas()
```

	author	subreddit	count
0	dedbrane	YoutubeRevolution	1
1	mkguyote	MakeupAddiction	1
2	godofalex	NewTubers	1
3	gtbishop83	comics	1
4	Josefina14	showbox	1
...	...	...	...
66422	inninara	nova	1
66423	alice_shadow	AnimalGIFs	1
66424	[deleted]	fyrehoz	1
66425	[deleted]	lifeofanintern	1
66426	Munchux2	listentothis	1

66427 rows × 3 columns

```
spark.sql("select author, subreddit, count(*) from reddit where not over_18 group by author, subreddit").toPandas()
```

	author	subreddit	count(1)
0	dedbrane	YoutubeRevolution	1
1	mkguyote	MakeupAddiction	1
2	godofalex	NewTubers	1
3	gtbishop83	comics	1
4	Josefina14	showbox	1
...	...	...	...
66422	inninara	nova	1
66423	alice_shadow	AnimalGIFs	1
66424	[deleted]	fyrehoz	1
66425	[deleted]	lifeofanintern	1
66426	Munchux2	listentothis	1

66427 rows × 3 columns

Custom functions

```
from pyspark.sql.functions import length

df = df.withColumn("length", length(df.selftext)) # This adds a column

df.where("length > 1000").toPandas()
```

	author	created_utc	title	subreddit	selftext	c
0	cruyff8	1454284808	How much money could the Australian government...	unitsd8u	Summary generated by [cruyff8's autosummarizer...	

```
from pyspark.sql.functions import udf

def splitWords(e):
  return e.split(" ")

splitWords = udf(splitWords)
df.select(splitWords(df.selftext)).show()
```

```
+--------------------+
|splitWords(selftext)|
+--------------------+
|         [[deleted]]|
|                  []|
|                  []|
|[Leave, psn, and,...|
|[https://www.cast...|
|                  []|
|                  []|
|[After, seeing, t...|
|         [[deleted]]|
|                  []|
|                  []|
|         [[deleted]]|
|         [[deleted]]|
|                  []|
|                  []|
|         [[deleted]]|
|[It, will, have, ...|
|[Is, there, any, ...|
|  [Rebecca, Jimenez]|
|                  []|
+--------------------+
only showing top 20 rows
```

```
df.groupby(["author",df.subreddit]).count().toPandas()
```

	author	subreddit	count
0	dedbrane	YoutubeRevolution	1
1	mkguyote	MakeupAddiction	1
2	godofalex	NewTubers	1
3	gtbishop83	comics	1
4	Josefina14	showbox	1
...	...	...	...
70322	inninara	nova	1
70323	alice_shadow	AnimalGIFs	1
70324	[deleted]	fyrehoz	1
70325	[deleted]	lifeofanintern	1
70326	Munchux2	listentothis	1

70327 rows × 3 columns

```
spark.sql("select author,subreddit,count(*) from reddit group by author, subreddit").toPandas()
```

	author	subreddit	count(1)
0	dedbrane	YoutubeRevolution	1
1	mkguyote	MakeupAddiction	1
2	godofalex	NewTubers	1

```
spark.sql("select author, count(*) as B from reddit group by author having B > 1000 order by B DESC").toPandas()
```

	author	B
0	[deleted]	20037

Assignment question

Obtain the users who have posted in reddit more than 1k posts in any subreddit

Answer:

spark.sql("select author, count(*) as B from reddit group by author having B > 1000 order by B DESC").toPandas()

-0 [deleted] 20037

SQL operations

How to declare a view from a Dataframe

```
df.createOrReplaceTempView("reddit")
```

```
spark.sql("select * from reddit limit 10").show()
```

```
+---------------+-----------+--------------------+----------------+--------------------+-------+------+
|         author|created_utc|               title|       subreddit|            selftext|over_18|length|
+---------------+-----------+--------------------+----------------+--------------------+-------+------+
|      [deleted]| 1454284800|[WTS] Tech21 Sans...|       Gear4Sale|           [deleted]|  false|     9|
|     amici_ursi| 1454284800|"[FS/FT] Mayaca, ...| ImagesOfIllinois|                    |  false|     0|
|        Buhadog| 1454284800|A combination of ...|customhearthstone|                    |  false|     0|
|    DJOlYoungin| 1454284801|[PS4] LF2M ToO Fl...|       Fireteams|Leave psn and I'l...|  false|    31|
|BrightenQuintin| 1454284801|[FEEDBACK] The So...|      recordthis|https://www.casti...|  false|   105|
|         jptx82| 1454284802|The gloves bother...|mildlyinfuriating|                    |  false|     0|
|       kahlil88| 1454284802|"questions" for m...|      Shoe0nHead|                    |  false|     0|
|        OverSol| 1454284802|Visual Comparison...| BleachBraveSouls|After seeing the ...|  false|   552|
|      [deleted]| 1454284802|hes tryin to turn...|       TheBarons|           [deleted]|  false|     9|
|      [deleted]| 1454284802|If Charles Martel...|      metacanada|                    |  false|     0|
+---------------+-----------+--------------------+----------------+--------------------+-------+------+
```

```
spark.sql("Select author from reddit where length(selftext)>1000 group by author").toPandas()
```

Assignment question

Obtain the users who have posted in reddit more than 1k characters in any subreddit with SQL (without using any column named length)

Answer:

spark.sql("Select author from reddit where length(selftext)>1000 group by author").toPandas()

como vemos arriba salen 3885 filas de resultados.

3881 Cpt_Insaneo_82

Other libraries

3885 rows × 1 columns

Beyond dataframes, we can find other libraries that also rely on Spark...

```
!pip install koalas
```

```
Collecting koalas
  Downloading koalas-0.32.0-py3-none-any.whl (593 kB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 593.2/593.2 kB 7.4 MB/s eta 0:00:00
Requirement already satisfied: pandas>=0.23.2 in /usr/local/lib/python3.10/dist-packages (from koalas) (1.5.3)
Requirement already satisfied: pyarrow>=0.10 in /usr/local/lib/python3.10/dist-packages (from koalas) (9.0.0)
Requirement already satisfied: numpy>=1.14 in /usr/local/lib/python3.10/dist-packages (from koalas) (1.23.5)
Requirement already satisfied: matplotlib>=3.0.0 in /usr/local/lib/python3.10/dist-packages (from koalas) (3.7.1)
Requirement already satisfied: contourpy>=1.0.1 in /usr/local/lib/python3.10/dist-packages (from matplotlib>=3.0.0->koalas) (1.2.0)
Requirement already satisfied: cycler>=0.10 in /usr/local/lib/python3.10/dist-packages (from matplotlib>=3.0.0->koalas) (0.12.1)
Requirement already satisfied: fonttools>=4.22.0 in /usr/local/lib/python3.10/dist-packages (from matplotlib>=3.0.0->koalas) (4.44.3
Requirement already satisfied: kiwisolver>=1.0.1 in /usr/local/lib/python3.10/dist-packages (from matplotlib>=3.0.0->koalas) (1.4.5
Requirement already satisfied: packaging>=20.0 in /usr/local/lib/python3.10/dist-packages (from matplotlib>=3.0.0->koalas) (23.2)
Requirement already satisfied: pillow>=6.2.0 in /usr/local/lib/python3.10/dist-packages (from matplotlib>=3.0.0->koalas) (9.4.0)
Requirement already satisfied: pyparsing>=2.3.1 in /usr/local/lib/python3.10/dist-packages (from matplotlib>=3.0.0->koalas) (3.1.1)
Requirement already satisfied: python-dateutil>=2.7 in /usr/local/lib/python3.10/dist-packages (from matplotlib>=3.0.0->koalas) (2.8
Requirement already satisfied: pytz>=2020.1 in /usr/local/lib/python3.10/dist-packages (from pandas>=0.23.2->koalas) (2023.3.post1)
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.10/dist-packages (from python-dateutil>=2.7->matplotlib>=3.0.0->ko
Installing collected packages: koalas
Successfully installed koalas-0.32.0
```

```
import collections
import collections.abc
collections.Iterable = collections.abc.Iterable
collections.Mapping = collections.abc.Mapping
collections.MutableSet = collections.abc.MutableSet
collections.MutableMapping = collections.abc.MutableMapping
collections.Callable = collections.abc.Callable

import databricks.koalas as ks
import pandas as pd

# Create a Koalas DataFrame from pandas DataFrame
kdf = ks.from_pandas(pdf[attrs])

kdf.head()
```

```
/usr/local/lib/python3.10/dist-packages/databricks/koalas/internal.py:1005: FutureWar
  [
/usr/local/lib/python3.10/dist-packages/databricks/koalas/internal.py:1012: FutureWar
  for name, col in reset_index.iteritems():
```

	author	created_utc	title	subreddit
0	[deleted]	1454284800	[WTS] Tech21 SansAmp Bass Driver - $175 Shipped	Gear4Sale
1	amici_ursi	1454284800	"[FS/FT] Mayaca, Crypt Wendtii, Java Fern, and	ImagesOfIllinois

```
kdf["sumChars"] = kdf.selftext.str.len()
res = kdf.groupby(["author", "subreddit"]).sum()
res[res.sumChars > 1000]
```

author	subreddit	created_utc	sumChars
Ughsmash	leagueoflegends	2908574611	1091
izzaberra	Paranormal	1454286411	7302
AMuseXD	PurplePillDebate	1454287017	4118
LittleMarMan	codes	1454287420	1301
branflakes	PersonalFinanceCanada	1454290057	1301
nefhithiel	IDontWorkHereLady	1454292348	1029
blackiechan99	tifu	1454298298	1244
LoveMeACrazyDPP	dirtypenpals	1454299093	1692
Pugway	Xcom	1454296498	2387
[deleted]	buildapc	65443507022	5142
AutoModerator	onlinemovieplaylists	17451666968	10140
HeartlessOne123	GlobalOffensiveTrade	1454285920	3444
[deleted]	SchoolIdolFestival	7271496567	1603
bmxmavericks	keto	1454288660	1024
Tnarg_Helped_Us	Fitness	1454288947	1095
Red_3_Standing_By	AVexchange	1454288984	1988
wreckingcanon	Metroid	2908588844	1519
TvviztedJezter	electronic_cigarette	1454291726	2522
vitreosity	flicks	1454293630	1182
lostgirlthrowawayy	relationships	1454293764	2431
Unclassified1	starbucks	1454293835	1460
Gear5th	Anki	1454294719	3954
zombiegrey	gratefuldoe	2908590699	3195
Wake-o-sleeper	dirtypenpals	1454297271	2041
Thereptilia	ecigclassifieds	1454297976	1484
beef_salad	Construction	1454299759	1553
Di_Bello	Mywhynot	2908603703	63901
[deleted]	zen	13088664701	3495
snowballcupcake	Accounting	1454285227	1086
amt897	buildapc	1454285616	2645
foxinyourbox	hockey	10180021017	2552
ydoowoody	retrogameswap	1454290082	1111
Spywin	UnknownTradeCo	1454293748	3742
ancientmelodies	hockey	1454293761	1973
CaymanG	Debate	1454294723	3996
DanknugzBlazeit420	DestinyTheGame	1454294997	1069
scooby_dooooo	india	2908599320	1750
sweattosuccess	loseit	1454300190	1069
triemers	ASU	1454300246	1878
fuqudouches	relationships	1454301368	4035
AutoModerator	powersofmiddleearth	4362855555	1889
lightshades	nfsnolimits	1454285767	3682
23andcounting	Herpes	1454287836	1510
Urall5150	OnePiece	1454289303	1923
JvIonden	relationships	1454289493	5344
Netscape9	GGFreeForAll	2908598805	2307
savannah_allie	pokemontrades	1454291759	1309
ZaneWuzHere	roblox	1454293162	1070

tthrowaway0001	infj	1454296379	1487
2806421200	tipofmytongue	1454298738	1993
probably_another1	r4r	1454301482	1594
[deleted]	GlobalPowers	1454301198	3675
henlow1i	nflprob	1454285027	7297
ldjarmin	HomeNetworking	1454285797	1186
Tixoz	FashionReps	1454290531	1757
YourPersonalDM	buildapc	1454290638	1289
Randolpho	SandersForPresident	1454290689	1587
driderr	Trimps	1454293937	4404
whoopsidied	bisexual	1454295934	1488
caesar45	swtor	1454298692	1242
repressedandconfused	offmychest	1454300694	1870
Poetic_Anarchy	SuicideWatch	1454301474	1751
gRntaus	gamedev	1454286102	1715
garysully1986	gameofthrones	1454286387	1254
PixieProblems	needadvice	1454288443	1749
viet4sh0	summonerswar	1454288805	3436
gonoworlater987	Career_Advice	1454289764	1237
Dgameman1	asm	1454293874	2379
sparkingthefire	legaladvice	1454293958	2512
KingOfMay	MakeupAddiction	1454293968	2464
helimx	cigars	1454294112	3732
soloans	personalfinance	1454295629	1760
Sandstone	CrystalGemRP	1454297129	1769
hi_Im_Meg	TBI	1454299993	1515
Cuddle-King	relationships	1454300772	2330
SlappaDaBassMahn	leagueoflegends	2908614527	1320
telestrial	Vaping101	1454284965	2527
Psychnerd12	Insurance	1454286833	1947
Ichigowins	Ichigowins	1454287635	1223
duckforceone	thedivision	1454289290	1415
randomlagger54	JhinMains	1454291358	6386
grappler0000	MakingaMurderer	1454292215	1519
illogic2	DarkNetMarkets	1454292717	4559
ManicFrizz	Fibromyalgia	1454293462	1134
tomorrowsanewday45	ibs	1454296034	2089
yajimari21	personalfinance	1454296930	1674
Named_after_color	ColoredInk	1454297796	4646
NotYouTu	BBQ	1454284983	1315
shannabuns	MakeupAddiction	1454285301	2078
Thief39	TheDescendantsOfRome	1454285802	2468
DogeyDogeDoge	buildapc	1454287234	1441
lexmasta	WingmanMiami	1454289066	4647
Strollo	DarkSouls2	1454291549	1015
pimpuls	Advice	1454292984	1829
Jakecoop2822	GlobalOffensiveTrade	1454293253	2172
aznscourge	pkmntcgtrades	1454293362	1618
derrickisdp	buildapc	1454294205	4941
goodnewsjimdotcom	Christianity	1454297892	1918

chillvilletilt	C25K	1454297681	1180
willyoung1999	techsupport	1454285994	1762
Oohsoangsty69xxx	teenagers	1454288457	1019
sexyfrenchboy93	buildapc	1454290641	3890
LohengrammRL	darkestdungeon	4362894260	1995
OGraffe	CFB	1454295776	1534
throwawaytrust12	offmychest	1454295831	1205
Taygai	WeissSchwarz	1454296965	1260
lukezach716	relationships	1454300307	2402
tl8719	WallpaperRequests	1454301331	2047
sludj5	Guitar	1454286535	1207
[deleted]	GGFreeForAll	1454286868	3368
littlesnapdragon	alittleolder	1454290887	1305
benitocamelas	Entrepreneur	1454293684	1151
spmcewen	iphone	1454297676	1662
grantn2000	GlobalOffensiveTrade	1454298351	3065
Xovius	dndnext	1454298958	1610
[deleted]	askscience	55263510603	1055
Saya93	GlobalOffensiveTrade	2908591316	1154
KeniyukiNC	PS2Cobalt	1454288146	2057
bittles_	exmormon	1454291652	4941
hotpequod	teenagers	2908589436	3245
Mr_Inconsequential	ROCD	1454297036	2246
StrexCorp	Smite	1454298377	1323
pietapang	Glitch_in_the_Matrix	1454298573	1082
AvoidingTheENT	AskDocs	1454288179	2507
Paraclipse	GlobalOffensiveTrade	4362872136	1533
color178924	Archery	1454285962	2221
m0narx	GlobalOffensiveTrade	1454288047	2128
TA6767	relationships	1454292496	1763
RTAdams89	NFA	1454297912	1135
matthewhugh84	nosleep	1454299109	14044
fourcheers	UnresolvedMysteries	1454301076	1174
joeybroadmoor	AgeplayPenPals	1454301491	1071
MikeBackAtYou	buildapc	1454285246	2493
garybracket	WiiHacks	1454288193	1086
LaboratoryOne	EliteRacers	1454288487	1112
beholdmycape	security	1454289172	1323
forescience	drones	1454291427	1041
Onlyhereforthelaughs	MLPLounge	1454292290	1520
Joetruck229	DetroitPistons	1454292544	4971
TW-Account	tifu	1454300807	1414
NotAnotherRando	loseit	1454300931	2151
Toadleclipse	PuzzleAndDragons	1454300955	1222
onrv	Musicthemetime	2908577799	8985
Raziel313	beertrade	1454284996	1053
King_of_the_Nerdth	SandersForPresident	1454286150	9372
Provoke-rs	Provoke_Adv_Log	24723210210	1095
slowlygoinground	homestuck	1454286575	1871
TakenakaHanbei	IronThroneRP	1454288143	6125

AutoModerator	Megaten	1454288573	1976
MTP_DER	cordcutters	1454288713	1003
blindpringles	Militaryfaq	1454292540	1388
Cinnypoo	parrots	1454295111	1102
bigblackskateboard	GlobalOffensiveTrade	2908601822	2601
dammitmang	trees	1454295683	3848
OpportunityKnox	pkmntcg	1454297932	1680
TheHynusofTime	Pikmin	1454299368	1110
CptnPants	buildapc	1454300154	2524
[deleted]	fivenightsatfreddys	20360129686	6689
trekbette	MealPrepSunday	1454285901	1426
ashimara	pennystocks	1454286369	3185
AnotherFewMore	SCCM	1454291255	1213
oeterb	Rainbow6	1454292180	1444
gunsonreddit	guns	1454298728	1740
SashaTheFireGypsy	tax	1454300918	1191
hippehpanda	relationship_advice	1454299415	1908
cptn__	GlobalOffensiveTrade	1454300752	1094
[deleted]	r4r	125069693924	2276
caseigl	GoRVing	1454285191	1193
MARS4597	spikes	1454285438	2216
Illsonmedia	DragonsDogma	1454285466	1001
lemlemx	thedivision	1454287831	3640
throwawaygradapp	GradSchool	1454288631	1266
yzmonker	Dreams	1454290046	1584
NotRuPaul	rupaulsdragrace	1454299853	2400
[deleted]	blackops3	49446206031	1459
idontknowanymore521	offmychest	1454286394	2305
SombodyThatUDontKnow	Agario	2908591736	1248
KingOfAllDogz	UHCMatches	1454290764	1040
BananaHamSalad	flashlight	1454293500	1995
angryco1	democrats	1454297914	2147
Modshroom128	halo	2908603349	1834
[deleted]	bladeandsoul	50900647355	2086
B3qui	LongDistance	1454286146	3556
naptownhayday	relationships	1454287382	2785
Escaflowne	escafrisky	1454291350	1529
justtotossit	mentalhealth	1454293305	3207
[deleted]	benzodiazepines	2908589876	2198
Bardock1233	gameswap	1454296574	1142
sealegss	atheism	1454297192	1429
JohniiMagii	DnD	1454297462	2217
sweet_Smolder_tank	stopdrinking	1454298651	2714
CreamPieSexCouple	CreamPieSexCouple	1454300753	3900
[deleted]	Showerthoughts	286497479456	2051
cygnae	GiftofGames	1454284866	1585
LK3000and1	opieandanthony	4362924748	1018
terribleswag	GlobalOffensiveTrade	2908590231	1834
notarealgobby	woweconomy	1454286965	1290
techyg	Multicopter	1454287296	3984

kittysdoormat	NarcissisticAbuse	2908582601	29123
darksweetrevnge	MakeNewFriendsHere	1454290691	1610
evilrick94	love	1454290866	1167
DBarrow15	squadup	1454293429	1803
skiwithpete	mechmarket	1454293922	2059
AnimeMusicLove	Advice	1454297535	1186
MyMonochromeLife	russian	1454300555	1135
cruyff8	unitsd8u	120705747519	92622
Maliw4n2015	buildapcforme	1454286503	2264
Turtle_Jedi	Fitness	1454288054	1439
craycatlay	NarcissisticAbuse	1454288168	1912
reactology	wow	1454289417	1035
FloridaFriend9000	r4r	1454291602	1676
Bennguins	NHLHUT	1454293489	4232
fartquestionthrowout	funnyfartstories	1454293722	2130
LegoMech	cthulhutech	1454293957	2591
Naman_Mehrotra	GlobalOffensiveTrade	2908595539	2260
Frikster	Laptop	1454296194	1536
GMCsonoma4x4	NoFap	1454298689	2435
KristenCampbell	BloodGulchRP	1454299977	6016
KirbyATK48	baconcraftia	2908614287	3071
Hannahlebanana	DogCare	1454300219	3069
eckodota	dota2tutor	1454299436	1923
five_of_five	tales	1454287150	1165
ZaBanpaiaNeko	NewTubers	2908594315	2205
pmurphh	GlobalOffensiveTrade	1454293722	1345
sandturtle024	Drugs	1454294434	3672
max225	GlobalOffensiveTrade	2908596660	1494
Zatain	GlobalOffensiveTrade	1454296501	1041
toobuku15	KingkillerChronicle	1454298627	1820
AutoModerator	aldub	2908606949	3587
otterplay	democrats	1454299542	2973
nat_r	translator	1454300388	1216
Night_Thastus	oblivionmods	2908598826	3419
[deleted]	videos	482828431558	2943
	thedivision	82895125226	1572
ZeroVibes	VanillaHCF	1454288877	1265
JoeyPrak736	dirtypenpals	1454289523	1399
DontAskhowIknow	dirtypenpals	1454291991	3099
AudaciTEA	mcservers	1454295628	4248
IlliterateEngineer	tipofmyjoystick	1454297636	1724
thebigbawk	Drugs	1454285163	1190
steakish	Maplestory	1454286352	1235
Shallan1	infj	2908583315	1917
undercoverwaffles	mac	1454287843	4237
andy33theone	NoFap	1454288334	1926
tristanwallin	LosAngelesRams	1454289139	1153
MowseChao	KDCGameGrumps	1454289774	1820
SpiderHack	noveltranslations	2908596200	1264
merrikatnip	nosleep	1454291520	16425

Timorelle	Bestbuy	1454292884	1548
cis-lunar	dndnext	1454293872	6389
Fa11enAngeLIV	XenobladeChroniclesX	1454293874	1620
SgtFlexxx	Rainbow6	1454295251	1947
gasolinetrain	dirtypenpals	1454295833	1919
Benjenzo	anime	1454296121	1040
reylee	bravefrontier	1454299086	19369
jagerbombastic93	tifu	1454301204	1298
[deleted]	DirtySnapchat	126524136163	1138
AIsnake55	DvZ	1454285542	1146
Krabbas	HamptonRoads	1454287524	2059
ec0402	GlobalOffensiveTrade	1454290611	3476
TristanL33	GlobalOffensiveTrade	4362881154	3830
LazyDinosaur	nosleep	1454291325	4159
mrmoncriefman	anime	1454292788	5639
GeneralPsycho5	cigars	1454293072	1495
NikkiP0P	cancer	1454293775	1098
writingtoss	SandersForPresident	2908578919	3116
gottagetover	relationships	1454292871	3560
NotSoFatHA	bigdickproblems	1454298598	1266
skdgldgajdgkj	nyiurmelambai	4362900360	23628
thundercatsh00	Seattle	1454299095	1281
RandomPhantom	jobs	1454300010	1471
J3Tisgod	respectthreads	1454285100	3636
RigasUT	kotor	2908600989	1490
NokkonWud	blackops3	1454285329	1386
RightiousToast	Ircast	1454285549	2247
TheDingusJr	NuclearThrone	1454285594	2392
FerrisTriangle	SandersForPresident	1454286344	1598
RecoveryJournal	pornfree	1454287001	1045
LegendaryPatMan	buildapc	1454289182	5267
Thresser	boardgames	1454289229	1377
imEFFE	LawSchool	1454293603	1068
KWsonar	GetMotivated	1454295472	1257
ticklestuff	spacex	1454296897	4832
SethrySethMcD	lostinwriting	1454297796	4987
Trinculoisdead	DnD	1454285082	1184
indariver	GlobalOffensiveTrade	4362900423	1185
Xmortus	SandersForPresident	1454288348	1740
m00Cat	raisedbynarcissists	1454289419	1352
manbare	CompetitiveEDH	1454291291	5004
AllHawkeyesGoToHell	CFB	1454294049	1770
AVG_AMERICAN_MALE	awardtravel	1454295679	1590
cunt48	buildapc	1454296228	2358
sabertoothduck	diyelectronics	1454296616	1242
KurodaRS	KurodaRS	39266439730	1754
PM_me_chubby_women	circlejerk	1454297563	39841
Karl_Marxxx	latterdaysaints	1454298280	2277
justneedtoletitgo	SuicideWatch	1454299285	1143
dspeyer	HFY	1454301522	7804

KustyTheKlown	skiing	2908577733	2507
RileyAbelAlt	InfamousSecondRP	1454289099	1639
XXXCheckmate	streetfightercj	7271489326	1916
unreplaced	dcrp	1454293917	3261
JES2140	buildapcforme	1454296475	2088
RomanNumeralII	conlangs	1454296613	2122
Lexielovekiss	Anxiety	1454301345	1266
612Shane	bodyweightfitness	1454301611	1040
2much2know	MakingaMurderer	1454296391	1037
Jafoob	relationships	1454284814	4262
slvrplme	GlobalOffensiveTrade	2908590541	1794
razurite	Roleplay	1454285164	1773
crabsongs	indiegameswap	1454285282	1669
idmonfish	vainglorygame	1454285582	6889
lolretkj	nicegains	45083251321	2424
moegli	DIY	1454289340	1812
i4k20z3	legaladvice	1454290533	2084
BloodyRahu	playarkservers	2908616724	9846
shingalated	aquaponics	1454292043	2148
maflickner	knifeclub	1454293568	3381
LacieIE	ExNoContact	1454299941	1117
SilentKnight333	IowaForSanders	1454292414	1605
minclo	personalfinance	1454296276	2126
melodyknighton	NoSleepOOC	1454285498	1248
Zulu95	IronThronePowers	1454286697	1374
SquallGunBlade	SuggestALaptop	1454287143	2531
Ulimit200	Ulimit200RsRSSfeed	61080589320	3587
agesrust	playrust	1454288300	1348
aaodi	investing	1454288430	1996
cephii2	GlobalOffensiveTrade	2908590117	1530
LizzyH-S	dirtypenpals	1454291538	1658
harzach	DasOhrIstDerWeg	1454291874	2187
Ebessan	SquaredCircle	1454292431	1267
Unr3alGamer	gravityfalls	1454292536	1388
Rocket_Sciencetist	spheremasterrace	2908590638	1032
fitzjack	Monitors	1454292905	1503
flameoguy	CivilizatonExperiment	1454294369	1488
MeatLover66	ACT	1454297717	1357
rsarector	ExNoContact	1454297740	1525
BigBooty-Milf	dirtypenpals	1454301195	2648
[deleted]	opiates	17451647106	1059
lethalcup	YewRS	30540372095	1408
GattDayum2	tifu	1454288674	3893
Kelawesome	ultrahardcore	2908583750	1113
telephonoscope	xxfitness	1454293689	1638
Polishkitten	neuro	1454294157	1152
LSteel4	HFY	1454295926	11227
blogg10	SuggestALaptop	1454299304	1293
baxteria	FantasyWarTactics	1454284813	1990
bountyxhunted	Guiltygear	1454285676	1035
[illegible]	[illegible]	[illegible]	[illegible]

Jay211	GlobalOffensiveTrade	2908589181	1375
dgrace97	learnjava	1454286711	1547
Nestle_ambrosian	OpTicGaming	1454286945	1061
AkaashMaharaj	pbsspacetime	1454288729	2105
waywardwoodwork	StarWarsBattlefront	1454288971	1530
marchpisces	childfree	1454291487	1726
All_Luck_No_Skill	leagueoflegends	1454291927	1052
metropolic3	eu4	1454292590	1293
DeadManWONDER	pokemon	1454293345	7075
heyRaxa	GlobalOffensiveTrade	2908609603	4045
techno_mage	DistantWorlds	1454299614	1300
a_not_so_random_name	mountandblade	1454300272	1181
SgtFlexxx	Warframe	2908580264	1301
johnnyquestNY	SandersForPresident	5817188158	2542
skiwithpete	MechanicalKeyboards	2908581765	2609
foreverguiltyanon	rapecounseling	1454288161	9329
danyzuko	buildmeapc	1454288673	1237
TurbidWarrior	HaloOnline	1454289567	5081
vizuelconquerer	gaystoriesgonewild	1454290763	8674
songbirddancing	ChronicPain	1454290905	1615
ACIDLIF3	depression	1454291732	3507
Selphade	kpop	1454291826	2489
daemonseeker	conspiracy	1454293166	4922
melveal	UCSC	1454295192	1179
Talaquen	personalfinance	1454296164	1175
IngloreuzBstrd	skyrimrequiem	1454297945	2733
2dP_rdg	malefashionadvice	1454299087	1033
klebermo	shield	1454285530	1102
GrizzlyS	NoFapChristians	1454287263	1026
TOKYO-SLIME	MetalGearPatriots	1454288896	3197
wotheli	Anxiety	1454289011	2739
RicFlairGangsta	Music	1454289547	1766
Dracious	dotaimba	1454289716	2000
suckitifly	bikewrench	1454289716	1510
Hillsmills	OSVR	2908585206	1038
COUTS_132	summonerschool	1454290808	1237
76SUP	tfour	1454291344	1370
Robobeepboop	buildapc	1454291751	1899
doodledays	depression	1454292161	1599
Tnpf	CitiesSkylines	1454294274	1252
Heartoplease	dogs	1454300012	4198
TooTyrnt	mgo	2908604203	1090
josephmurielgellar	r4r	1454290254	1086
NarrowElf	ProduceMyScript	1454292722	1069
kpthrowaway8290	AskDocs	1454295933	1868
DarkLorde117	RWBY	2908608408	3078
rin_shinobu	wormrp	1454298377	4436
ADaddyHansen	techsupport	1454299623	2879
IVlattEndureFort	DIY	1454299948	1507
Voltairious	relationships	1454289111	1175
AutoModerator	summonerschool	1454299580	4269

AutoModerator	summonerschool	1454295980	4269
manowarp	tipofmytongue	1454286990	1158
CR4allthethings	running	1454287986	2076
BayesianJudo	csshelp	1454288664	1528
edgar193	help	1454288735	1082
TeddyBdaGOAT	FanTheories	1454290925	1927
LeBeauMonde	AskVet	1454292811	1083
MissionaryControl	RandomActsOfBlowJob	1454295020	1102
SoulsPedia	bloodborne	1454297418	1106
tmtreat	CitiesSkylines	1454294195	1517
AutoModerator	Romania	1454285023	1970
Blee10	SquaredCircle	4362891417	2705
Thekiddsgood	summonerschool	2908572631	1182
acrimony87	battlefield_4	1454286837	1908
SirWookieeChris	DnDBehindTheScreen	1454287080	1399
chivnz	SteamGameSwap	1454288838	1648
sircumsizemeup	Naruto	1454289077	1491
exmographer	exmormon	1454292023	7311
AutoModerator	Dogtraining	1454292135	2087
tburke40	gis	1454292585	2945
Buzzismydog	trees	1454292756	1951
daftalchemist	ftm	1454295148	1307
Throw1010Away	DeadBedrooms	1454296600	1584
Quihatzin	alcohol	1454297019	1040
Ombliguitoo	stunfisk	1454297808	2500
CondorCalabasas	GGFreeForAll	2908596899	1510
CrumblyButterMuffins	socialism	1454298046	1078
Ninjapahnda	lonely	1454298985	1671
gemnight	depression	1454300278	2105
acheyshakey	AskGayMen	1454284906	2070
MoreSkindredPlz	Skindred	1454285525	2497
xveg	VictoriaBC	1454286258	1714
Hammeredmantis	relationships	1454286801	1578
Ginaguo411	3FVAPE	1454292632	1867
exmointhecloset	exmormon	1454296372	3258
Ochaosnine	ecigclassifieds	1454297560	1379
SvenBTB	r4r	1454297791	1633
Scubastevie00	pcmasterrace	2908602419	2236
jlovisa	Bitcoin	1454300782	1770
tentends1	stopdrinking	1454300809	1219
Vicebit	pkmntcg	1454300956	1134
Kenbobb	hillaryclinton	1454298787	2307
Jak3theD0G	yugioh	1454285450	1136
r4x	sysadmin	1454287215	1223
spike_africa	tdi	1454287262	3273
Nyalloyd	Buddhism	1454287932	1558
twoiko	dirtybombconfigs	1454288836	4124
gorillakitty	GorillaRecipes	1454289692	1155
AlliedKhajiit	Guildwars2	1454291593	1322
Welden10	thedivision	1454292097	1508
Envimea	mcservers	1454292329	1090

NotTri	GlobalOffensiveTrade	4362889870	2687
Ahnaful1994	Ahnaf	23268825405	1039
Legoman1357	buildapc	1454296782	3086
Xdoctor	buildapc	1454298315	3382
ps6000	smoking	1454298490	1274
blaewen	Journaling	1454299623	1003
ArkSurvivalEast	playarkservers	1454299998	3152
DigitalSiren	FFXIVRECRUITMENT	1454286875	2377
AlwaysBeNice	awakened	1454285506	1826
somethingsassy	dirtypenpals	1454285730	1652
cannotfindanamee	IronThronePowers	2908575026	8171
SweetRissa	relationships	1454287179	5685
andrew650	techsupport	1454289297	2245
YorjYefferson	lexington	1454292426	2848
zipKill_FRAG	wRedditStreams	1454293217	1631
mandemscomin	nosleep	1454294121	5316
uinstoncharchil	NeverBeGameOver	1454296348	1994
ShbeblyTheGreat	pokemon	1454298105	1034
frost-shock	buildapcforme	1454298842	2565
panicATC	Anxiety	1454287549	1840
anikan1297	javahelp	1454288135	2283
sbpotdbot	sportsbook	1454288424	1187
Elijah_Abels	exmormon	4362899573	17909
AaronWithAQ	UHCMatches	2908580208	1079
Xerte	bravefrontier	1454289616	3551
OpenInTheory	polyamory	2908592939	5917
garysully1986	asoiaf	1454290327	1358
SwagBacon	LetsNotMeet	1454291863	2373
Epic_MC	streetwear	2908587234	1351
sasago	AskDocs	1454293321	1030
reversebottle	Music	1454294483	1079
brand0n	SteamGameSwap	2908595158	3588
DeeSchro	buildapc	1454297925	2614
maxyg1234	NHLHUT	1454300695	1403
emotionaltightrope	LegalAdviceUK	1454285599	2490
Doctor-Swoosh	loseit	1454285654	7013
cs_quest123	cscareerquestions	1454286825	1197
Mutant_Llama1	whowouldwin	2908576245	1673
throwaway49416	DeadBedrooms	1454288442	1091
showni	GlobalOffensiveTrade	2908608511	4915
derpaholicsanonymous	buildapc	1454290694	1155
DMeville	gameDevClassifieds	1454293450	2552
wer66	HFY	1454295102	10556
PiwwowPants	buildapc	1454295996	4695
Dr_Hydra	thedivision	1454297651	2226
l0an3rthrow	legaladvice	1454300650	1217
Drunk_Conlangs	conlangs	1454301366	1772
Eunovation	PuzzleAndDragons	1454284990	1094
AutisticJaffaCake	depression	1454285226	1104
sloopdoop	CabaloftheBuildsmiths	1454285541	1920

iguanajm	jailbreak	1454286302	1025
PM_ME_PSN_CODE	teenagers	2908592925	2407
GuyNoirPI	asoiafcirclejerk	1454295807	1167
victoriasbitter	Shoestring	1454301098	1731
Elivaras	LeagueofAngelsMobile	1454284837	5666
mickeywickey	AskWomen	1454287033	1307
tekproxy	playrustservers	1454287890	1285
yainfp	askgaybros	1454289098	2685
Prezombie	TheWitness	2908588902	1771
inkexit	smallbusiness	1454290112	1895
teamjennacide	TalesFromRetail	1454291026	4469
asmodeus81	summonerswar	1454292320	1162
AnonymousTurker	electronic_cigarette	1454292588	1075
jj4sanders	SandersForPresident	1454293658	1882
imasharpay	Sasquatch	1454294332	1055
Panhead369	ModelUSGov	1454297515	1695
Phrave	cscareerquestions	1454298641	1350
dontgetsad	makeupexchange	1454287769	5176
littletinykeys	nosleep	1454289879	8899
fetishsam	dirtypenpals	1454290010	1809
qayum	optometry	1454291207	2063
[deleted]	ForeverAloneWomen	4362902592	1329
flipton	facebook	1454298540	1102
Chromi0	TheWitness	1454298657	1227
nafedaykin	premed	1454286313	1878
Louchlyn	exmormon	1454286871	1379
blackdog314	gratefuldead	2908582475	1181
STCVKR	NoFap	1454289022	2779
pingasthrowaway	summonerschool	2908611140	3097
flippinkatie	relationships	1454292081	1021
JGPH	AndroidQuestions	1454293599	1345
SearchingTheTrue	NeverBeGameOver	1454294120	1066
MrCoolCol	Militaryfaq	1454297387	1582
Xkimberxkae	TalesFromRetail	1454297924	1400
ProfessorKag	college	1454284853	1006
HolidayNick	relationships	2908572114	1606
rotorschnee	Lovecraft	1454286005	1212
mypersonalshit	AskDocs	1454287952	2257
JakeVanderArkWriter	Spanishhelp	1454291071	4148
EpicDildo	Homebrewing	1454292017	1916
Prettychilledoutguy	summonerschool	1454292080	2876
ThieF60	flashlight	1454292512	1208
TheSauceyNoodle	loseit	1454294755	1931
5ft0lady	relationships	1454294808	2107
AllisGreat	GlobalOffensiveTrade	1454296397	1448
aceburninator	running	1454297702	2405
Sportfreunde	hockey	1454299258	1245
DrawerFullOfDicks	rapecounseling	1454300328	1702
NightcrawlerKing	DC_Cinematic	1454284989	1023
removalbot	removalbot	74169632192	23210

cantstopthinkingso	actuallesbians	1454288716	1026
TrotsTwats	mountandblade	1454291367	8841
Throwawaybabee00	actuallesbians	1454292573	2946
ORBGaming	GlobalOffensiveTrade	4362920651	48270
feelbadthrowaway1234	relationships	1454293729	2540
nflman2117	buildapc	1454294668	3174
oden619	yugioh	1454296039	2711
_teslaTrooper	learnpython	1454296099	2941
stcordova	chemistry	1454300767	1340
TeamMcLovin	fakeid	1454286155	1669
segIda	buildapc	1454287924	2893
DarthTator	ClarkCountyAirsoft	1454288566	1268
Confusedlover1234	relationships	1454289747	1296
h4venz	GlobalOffensiveTrade	1454290421	2895
deltaprogress	dailydraw	1454290456	1346
ImRickyBobby	opiates	1454290833	1427
ScarlettRose20	AgeplayPenPals	1454291408	1045
laughing_cat	SandersForPresident	1454292800	1196
Midori_Kun	moronarmy	1454294370	2690
44davi	Glitch_in_the_Matrix	1454294768	1428
midasmcfunk	techsupport	1454298044	1188
VitaGod	hcteams	1454290630	2170
[deleted]	cars	27631776899	1391
dgafkt	AsianBeauty	1454285787	1191
MasterLJ	cscareerquestions	1454291171	1183
Skellyton5	Askasurvivor	1454291419	1170
juriah121	oppopiopi	65443160455	65340
trentbat	Undertem	1454294908	1873
LIATG	BestOfOutrageCulture	1454297721	1929
throwaway368000	sex	1454299093	1278
cactoidjane	skyrim	1454300927	1463
brendanlim	KeybaseProofs	1454301298	2539
Griffin_Throwaway	dirtypenpals	1454298312	3546
Dexter87	IronThronePowers	1454284854	2394
[deleted]	relationships	162881729464	12235
Thedarkskinnedbrit	Fitness	1454285127	1835
xXdeathstar101Xx	teenagers	2908574094	1356
[deleted]	Parenting	2908578996	1159
JmodTracker	jmodtracker	8725891276	1915
Cenki	Fitness	1454288923	2384
BigMik_PL	thedivision	4362903640	2320
RichHardLemons	marvelstudios	1454289132	1321
Nogarda	thedivision	1454290381	2514
Weedfreelifestyle	leaves	5817243246	1757
HoyaSaxons	loseit	1454292395	1329
Smallchelle	cigars	1454296246	1259
StreakUHC	UHCMatches	2908597423	3036
grains_r_us	wallstreetbets	1454298984	4541
milky9311	consulting	1454299423	1114
reslifelifer	xxketo	1454300911	1763

ChipperbrownXO	explainlikeimfive	2908579025	1832
[deleted]	melbourne	21814624580	2336
DarkGenesis327	MonsterMusume	2908602166	1328
gingerattacks	succulents	1454290567	1093
ScaredShitlessLds	exmormon	2908596419	1329
[deleted]	myalog	26177514964	1068
Myusha123	Undertale	1454294356	13784
DarkRedTwist	NoFap	1454297142	2621
r-a-f-s	BreakUp	1454300663	1521
brocopina	FEGaiden	1454301339	1549
Targren	tipofmytongue	1454301446	1116
Gentle_Beard	bjj	1454295354	1757
Prone1	relationships	2908571482	1754
Pjd1986	fixmydiet	1454285563	1337
stairfaller	dragonage	1454286760	1426
nebulouscho	youtubegaming	1454289736	1042
KarmicEnigma	xxketo	1454290397	1499
vulcanfury12	buildapc	1454292931	1248
CitizenJac	ChicagoNWside	1454293585	1381
22Anonymous22	creepypasta	1454297936	1419
IWillNotLie	Grimdawn	1454298754	1305
DJ4Bernie2016	SandersForPresident	1454298825	1111
roqlord	DawnPowers	1454300318	2738
UniTe_CSGO	Fallout4Builds	1454301209	1597
Feasinde	tolkienfans	1454301267	1775
[deleted]	Guitar	29086000794	2111
	GlobalOffensiveTrade	836224887734	6573
Flatoftheblade	CK2GameOfthrones	2908572908	1524
foxdubois	AskDocs	1454285998	1316
cat_puke_shoes	applehelp	1454286597	2329
angrypenguin625	legaladvice	1454287529	2106
Sage_Musa	Naruto	1454287674	1005
_Amateur_Hour_	painting	1454287938	1274
liftedyota	MechanicAdvice	1454288293	1156
nicerackoflamb	SandersForPresident	1454288444	3569
shda5582	excel	1454288454	1298
sstrees	buildapc	1454288517	4061
sewhigh	microgrowery	1454289529	1567
theultmatecad	TheRedPill	1454290319	2815
MaizeZea	tax	1454300194	2163
daaatgekko	birthcontrol	1454300899	1119
hdtv003	sompsma	1454285079	2231
ScienceShawn	trees	1454288077	3016
Orl_lease	orlando	1454291130	1498
Spiraticus	summonerschool	1454292153	1077
wubalubadubduub	depression	1454293626	1068
hugs_not_ughs	sex	1454293957	1030
Little_Party	summonerswar	2908590255	1356
mrtman327	CasualConversation	1454294907	2546
seventhward	hometheater	1454296914	4063

justhisguyouknow	Vaping101	1454285088	1095
NanchoMan	SSBM	2908572005	4512
CClossus	LyricalWriting	1454286409	1702
narcolepsythrowaway1	SuicideWatch	1454287304	1999
absolutjorts	personalfinance	1454290107	2466
TruthandNature	SandersForPresident	1454291796	1166
HockRivers	Advice	1454292832	2007
Legofan2676	DvZ	1454293864	1298
happy509500	happy509500	4362900357	6911
splitshot	ar15	1454299699	3225
El_Nahual	Cuckold	1454301581	3126
Michaeltatio	fakeid	1454285509	2397
OneEarth3	DragonsDogma	1454285592	1333
Zorceror44	ExploreFiction	1454286282	1139
Marksman5147	PS4Planetside2	1454287527	1145
cocojambooo	GlobalOffensiveTrade	2908603574	10403
FerralOne	gamedev	1454289736	1238
cjjmfg	nosleep	1454291019	8037
sorenayrie	Pathfinder_RPG	1454291529	1068
smegko	BasicIncome	1454296512	5857
ZeosPantera	Zeos	1454297891	8309
FetusMonkey33	buildapcforme	1454298013	1644
RiskyDriver	self	1454300109	1239
Beatnik_Exploit	keurig	1454289874	1551
The-Vale	bindingofisaac	1454286986	1114
Borne2Run	empirepowers	2908581905	1489
x_yellowbird	TwoXChromosomes	1454290499	1618
ZalbagMC	CivAquila	1454292845	2065
aphistic	RATS	1454293579	1092
renascentangel	Guildwars2	1454293922	1526
Warmain	darkestdungeon	1454293942	5768
confusedmofoofreddit	relationships	1454294314	4839
Nevergoingback11	OCD	1454297745	1767
[deleted]	OpiatesRecovery	1454299443	1762
MasterHistorian	tifu	1454285925	16484
[deleted]	OkCupid	30540269987	1342
heywood_jablomeh	buildapc	1454286600	2554
Krabbas	NewportNews	1454287572	1935
Outmodeduser	whowouldwin	1454288925	2633
MushroomPlanet	buildapc	2908590371	8231
BrinkBreaker	EnorousPact	1454289896	2626
T3mpt	Mortgages	1454290354	1404
Humminglady	CasualPokemonTrades	2908592355	1605
The-Magic-Sword	UnearthedArcana	1454296924	2281
therealmenox	wowguilds	1454297023	1238
CraterCat	TalesFromRetail	1454297924	1474
lyssinflannel	Stutter	1454298971	4991
GayWarden	FanTheories	1454298979	2026
scheelio	marvelheroes	1454284921	3312
KahulaMatata	alcohol	1454285607	1235

gyfaglover4	HeyManga	59626629087	4729
fghfffffff	relationships	1454286442	2151
lmallagog	buildapcforme	1454287467	1821
QuotedMC	hcteams	1454287965	1143
lopgan121	SuggestALaptop	1454288688	1276
thejudicialpenis	TalesFromRetail	1454288830	2316
Odilon616	thedivision	1454289920	1892
Trainsman4	leagueoflegends	1454290937	1496
markt09	Adelaide	1454291100	2360
GuyMontagz	spikes	1454293415	2439
lmipolex42	ProsePorn	1454293701	1052
neongreenpurple	Random_Acts_Of_Amazon	2908606999	1948
Flewtea	breakingmom	1454298657	1301
Reksew_Trebla	RWBY	1454299471	1689
drunkwithanxiety	coys	1454301387	1520
ddeef	thinkpad	1454286461	2381
Seventhghost	talesfromtamriel	4362864775	69690
babycakes101	metalgearsolid	1454288622	3081
borrrden	Tokyo	1454289080	1710
Thecougarkid	hometheater	1454290649	1616
tysrak	birthcontrol	1454291134	1282
PayneGreyWolf	askseddit	1454292272	1095
Bicyclestories	sex	1454292923	1650
tretrebs	buildapc	1454293528	2529
kangkungking	buildapc	1454294933	3089
Achierius	HistoricalWorldPowers	2908617373	1848
actuallyfromcanada	DeadBedrooms	1454295349	1722
sadbat	boardgames	1454297297	1261
creativextent	buildapc	1454297754	1633
DrakeH44	GlobalOffensiveTrade	1454298597	2198
TheIncandescentAbyss	bleach	1454300441	2375
nasil2nd	GlobalOffensiveTrade	1454284900	2482
ohnoimrunningoutofsp	HaloStory	1454285097	3553
heavncentt	fitbit	1454286159	1004
aanthonyz	learnprogramming	1454286660	1817
[deleted]	dating_advice	18905918979	2540
Jorlung	gradadmissions	1454287918	1378
Iridar51	Planetside	1454288734	7829
[deleted]	korea	7271504463	5200
FallenRenegad3	empirepowers	2908585808	1065
leaopeng	tyuityuityu	78531836786	78408
fassla	dirtypenpals	1454293732	1885
MasterNeo27	SandersForPresident	1454296007	1488
Halfayear	relationships	1454296329	7269
WinterRays	dirtypenpals	1454297012	3619
melon2020head	relationship_advice	1454298147	1512
[deleted]	Jokes	81440860054	1249
ulkesh12	ShitRedditSays	1454285405	6429
setichi	canada	1454288308	1030
Cole-train99	elderscrollsonline	1454289765	1322

kdz13	Scotch	2908586351	1018
pm_me_your_kindwords	bullcity	1454293407	1028
chrispy294	PKMNRedditLeague	1454294943	2010
talltree2011	cigars	1454298449	1428
DoctorCheerio	gamemaker	1454298594	1193
jack_skellington	Pathfinder_RPG	1454300451	3349
[deleted]	xxfitness	1454285346	1898
theez35	GlobalOffensiveTrade	2908577702	1050
AverageJoeAudiophile	BudgetAudiophile	1454287035	1101
Meronomus	Advice	1454287211	1311
OnmyShoulders	selfimprovement	1454287243	2663
tom3838	GGFreeForAll	1454288600	1323
[deleted]	teenagers	40720437543	1518
VeroV133	Roleplaykik	2908585669	6760
NHKEasyNewsBot	NHKEasyNews	7271522639	2581
tooomine	Cooking	1454294912	1360
hachisugoi	makeupexchange	1454295315	1621
addythrowawayacct12	adderall	1454295590	1522
damonstea	tabletopgamedesign	1454297102	1226
Notalzac	MilitaryStories	4362930907	9205
ranalog	analog	4362862228	1532
Saxaphones	depression	4362864300	1204
emre23	LiverpoolFC	2908571883	1773
4benny2lava0	malehairadvice	1454289240	1020
HowDoYouLikeMyName	GlobalOffensiveTrade	1454289395	1478
flufthedude	relationships	1454289798	1034
xFarside	TownofSalemgame	1454289875	1148
ZaBanpaiaNeko	YouTube_startups	2908594298	2204
KleenexVII	xboxone	1454293827	1375
Vexwyf	infertility	2908590560	5493
NicoleMitchell	Spanish	1454295422	1504
nightwolf16a	DFO	1454298726	2077
DancingPear	TryingForABaby	1454299857	1150
masterrucker	AndroidQuestions	1454301240	4046
upads	investing	1454298142	2992
competitiveslacker	Anxiety	1454300318	1367
[deleted]	proED	4362918613	1689
tenac1ty	GlobalOffensiveTrade	2908595246	1180
Ahnedonia	raisedbynarcissists	1454291018	2533
enlambdment	leaves	1454291739	6022
Airum	photoshop	1454293316	1064
EvenFlowX93	SquaredCircle	1454294673	1191
TheNedben	whowouldwin	1454297093	1423
pleasehelpme232321	Insurance	1454299119	4795
Bennators	NHLHUT	7271471258	1442
BeingMC	UHCMatches	2908590058	3253
darkhado14	SuggestALaptop	1454297705	1825
CelineHagbard	C_S_T	1454299917	2774
starfruitcake	noveltranslations	1454298773	1682
funkyfalcon34	Csgotrading	1454287051	2970

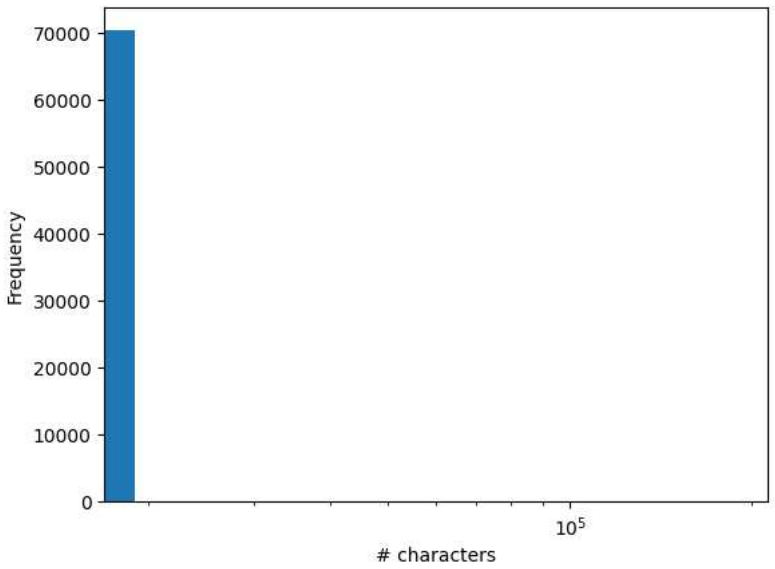
lyncblaze	battlefield_4	1454287731	1349
MadKingConnor	pcmasterrace	5817196783	1777
Dark_Element75	buildapc	1454292455	1264
Jumpinjoe123	FindAUnit	1454293459	1459
what2dowit	relationships	1454298423	6392
clementaiden	guildrecruitment	1454300414	1080
Nick2the4reaper7	stunfisk	1454286404	1005
silec	Diablo	1454286485	2135
lol_r_amiibo	Flipping	1454287100	3153
porksmash	django	1454287150	2314
Connorcpc	Drugs	1454287487	3647
nooblord111	leagueoflegends	2908608476	1237
SarahtheSlayer37	AgeplayPenPals	1454295097	10187
Feezec	xcom2mods	2908601109	4336
ivankasta	changemyview	1454299787	1846
avngr	Netherlands	1454301489	1016
TransmissionBuilder	NoContract	1454284969	1201
[deleted]	pcmasterrace	84349387524	3388
Beythoven	GlobalOffensiveTrade	1454286195	9992
Krawdd	ClashOfClansRecruit	1454287414	1173
Antetokounmpo	MkeBucks	1454287851	1597
raymanfan1	Metroid	1454288537	1278
vic6string	woweconomy	1454291070	1446
Chapperion	XWingTMG	1454291209	1132
xxglossxx	SMITEGODCONCEPTS	5817252390	6185
camafu	AirForce	1454294044	1728
[deleted]	quittingkratom	1454294891	1370
JarnabyBones	SandersForPresident	1454296004	1199
Garrod_Ran	Gunpla	1454297715	1170
mare0987	nosleep	1454300378	8753
questionablysober	leagueoflegends	1454285508	1343
DohTellMehDat	GlobalOffensiveTrade	1454286012	1027
missymia161	femalefashionadvice	1454289242	2848
DradDog0419	personalfinance	1454291250	1320
Stev_Zarr	whowouldwin	1454293807	2696
rsvpism1	TrueOffMyChest	1454295142	5780
hombredesangre	offmychest	1454295741	1071
Castleofwolves64	furryrp	1454296388	1029
axeupon	thedivision	1454296554	1712
ChiBlock	magicTCG	1454297212	1631
microninja162	airsoft	1454298748	1212
skdgldgajdgkj	sumberrejeki	2908601748	15793
DrTestBender	linuxmint	1454300797	1417
Vonragnier	GlobalOffensiveTrade	1454296624	1083
23458273465982736	relationships	1454284872	1577
supa_troopa1	ffxiv	1454288816	1158
gwyn15	weddingplanning	1454289990	4656
Liz1978	VPNTorrents	1454291715	1416
TaisiaTuaMagia	walmart	1454291910	1365
memoglobin	DotA2	1454292691	1481

Redleg61	PoliticalDiscussion	1454293966	1167
AltMom	ankeny	1454295598	1010
imnaked0	Steam	1454296659	2026
yliv3	SuicideWatch	1454297204	2847
Tokenofhon	nrlrll3	1454297947	1123
420Throw69away420	tifu	1454299195	1385
MeatLover66	chanceme	1454300319	1081
Mahousite	raisedbynarcissists	1454298201	1388
PerpetualDiet	whole30	1454287818	1409
zingsla	buildapc	1454288791	3112
l33tSpeak	tattoos	1454288920	1151
lcronics	hcteams	1454289207	1293
AtomicCoyote	TalesFromRetail	1454289454	1843
TheWizland	OverwatchHeroConcepts	1454290590	3338
SomeDumbHaircut	leaves	1454290730	1407
acloudrift	conspiracy	1454293267	1732
TheDandyLion	AnimeDeals	1454293583	2190
Editorgirl2617	relationships	1454296345	2212
betterthanprozac	kratom	1454297413	1033
19832012	occult	1454299923	1162
DearDarlingDearling	raisedbynarcissists	1454286311	8260
[deleted]	CasualConversation	31994718099	1239
ewhetstone	legaladvice	1454287848	1153
R0man1ac	leaves	1454289173	2044
kardkoach	buildapc	2908606247	4040
NaughtyNina69	actuallesbians	1454290391	1059
shittyfuckinthrowawa	relationships	1454291310	2563
jenhai	exchristian	1454294639	1885
RayVicario	microgrowery	1454294740	1361
shimmydance	tipofmytongue	1454294906	1353
Screamin_STEMI	ems	1454295774	2662
HSAHughes	DvZ	1454298462	2659
BigThickAss	dirtypenpals	1454300531	1805
amierchery	SandersForPresident	1454300631	1725
TTTonster	lordsoflootlog	47991785317	2268
InsaneOne8977	GiftofGames	1454288421	2539
wkf1114	OnePieceTC	1454289130	1526
horrorgal98	Paranormal	1454291225	3348
aeternuseternus	DotA2	2908593646	1948
zCiver	minerapocalypse	1454294050	1126
Three_If_By_TARDIS	SandersForPresident	1454294601	4047
nexeus	magicTCG	1454295532	1702
MrsSaffronReynolds	100movies365days	1454296545	1452
spikernum1	DotA2	1454296552	1992
aaaaa123throwaway	pettyrevenge	1454297357	1230
Chrispytoast123	MHOC	1454298559	3875
KennyLovesYou	woodworking	1454299313	1258
closetdork	breakingmom	1454300378	1555
Robman24	cscareerquestions	1454300524	1791
AboutNerf	Nerf	1454301624	1891
Cyarkos	PokemonTurfWars	1454286474	1399

CyarKos	PokemonTurnWars	1454286474	1399
ydtm	btc	4362883049	10021
DPP4711	dirtypenpals	1454291253	1986
Sedirex_KR	KamenRider	1454296356	1230
soulwyvern	SVExchange	8725805892	5496
Yesnothrway	relationships	1454299228	1062
PizzaCompiler	HomeNetworking	1454299567	2432
xkha0z	sportsbook	1454300034	3360
tony_chen0227	GlobalOffensiveTrade	5817152941	2152
[deleted]	gonewild	244322795710	1395
TrojanArmor	GlobalOffensiveTrade	2908575268	2587
Jorlung	EngineeringStudents	1454287277	1380
teefletch	SSBPM	1454289008	1083
niggerbernie	GGFreeForAll	2908580257	4144
tiggerbounc	RunUntilLikeMyself	1454290495	1002
HabsGirl	PersonalFinanceCanada	1454291806	1218
_STOP_YELLING_AT_ME_	Christianity	1454293676	2081
hello_company	xxketo	1454294329	1723
C0mpass	AyyMD	1454297647	3023
vineyau	malelivingspace	1454298401	1386
GunSizeMatter	buildapc	2908582218	3139
ferixchen	GlobalOffensiveTrade	2908588646	1558
owned_at_worms	MaddenUltimateTeam	1454300432	1122
Saranjello	Stronglifts5x5	1454284954	1270
csgonews	GlobalOffensive	1454285354	1778
DogaLover	GiftofGames	2908595323	1108
Z3ROWOLF1	blackops3	1454299098	1465
Vexedzero	DestinyTheGame	1454300021	1370
[deleted]	4yogurt	1454284997	17309
__SoL__	FalloutMods	1454285177	2164
DaB0mbb	relationships	1454285525	1053
darkhairthrowaway	FancyFollicles	1454286608	1018
Jelre	GlobalOffensiveTrade	1454286749	1271
teffdawg	ArtBuddy	1454288919	1303
trevboss124	GlobalOffensiveTrade	2908605556	4248
DMTsmoke	starcitizen	4362915317	1672
TaGeule	Rainbow6	1454294840	1141
erer1243	every15min	66898034657	2096
colsanders37	legaladvice	1454285789	1267
Mimihop	ACTrade	1454290620	2699
grandkgamer	GlobalOffensiveTrade	2908616730	1004
c135qet	learnprogramming	1454299134	1953
psychospacecow	customhearthstone	1454300695	1565
ukuleleteacher	ukulele	1454285567	1310
colinrhyshill	SandersForPresident	2908581623	2111
I_Luv_Oreo	hearthstone	1454285659	1339
cleanmindhappymind	latterdaysaints	1454285724	1772
ccmanga	AnimeFlu	17451651372	1686
confusedfat	TwoXChromosomes	1454287866	2037
TheMayaEdits	GlobalOffensiveTrade	2908582201	4324
bodecker	halo	1454290329	2523

redditfellow	GlobalOffensiveTrade	2908595480	3570
[deleted]	investing	7271516926	1413

```
import matplotlib.pyplot as plt
plt.hist(res.sumChars.to_numpy())
plt.xlabel("# characters")
plt.ylabel("Frequency")
plt.xscale("log")
```

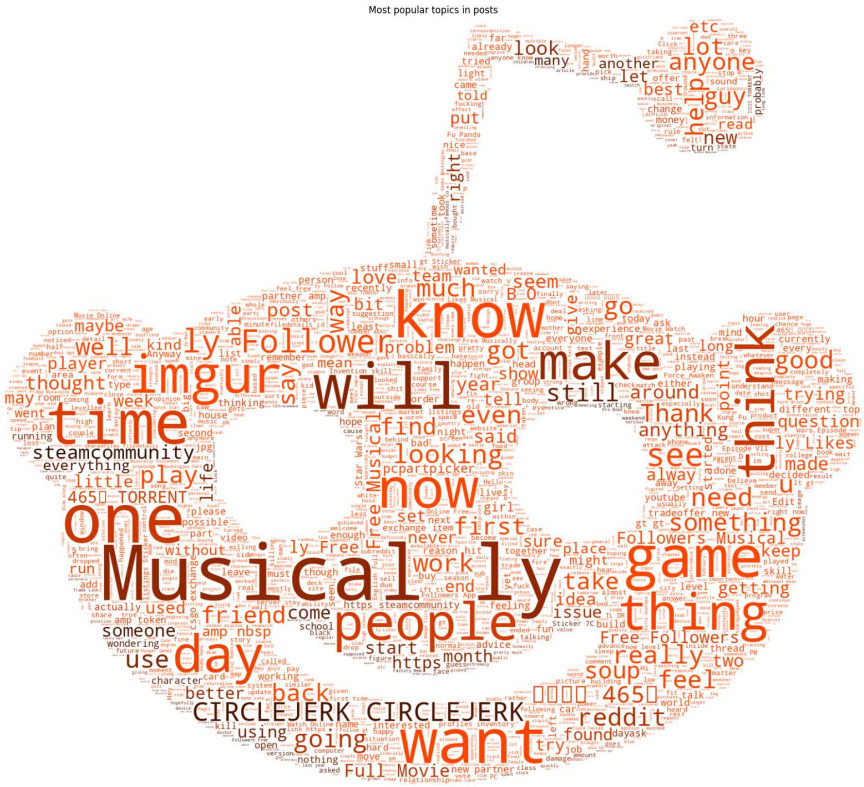



```
!curl https://2.bp.blogspot.com/-eGskF3n8_Ag/XE7F3P_de2I/AAAAAAAAHU8/WJw0un2nHqMGA8cFVtv_yFfpBVQJSYyVACK4BGAYYCw/s1600/Icon-Reddit.png >
from wordcloud import WordCloud, ImageColorGenerator
from PIL import Image

mask = np.array(Image.open("reddit.png"))
text = " ".join([i for i in kdf.selftext.to_numpy() if len(i) > 0 and i != "[removed]" and i!="[deleted]"])
```

```
  % Total    % Received % Xferd  Average Speed   Time    Time     Time  Current
                                 Dload  Upload   Total   Spent    Left  Speed
100 76154  100 76154    0     0   161k      0 --:--:-- --:--:-- --:--:--  162k
```

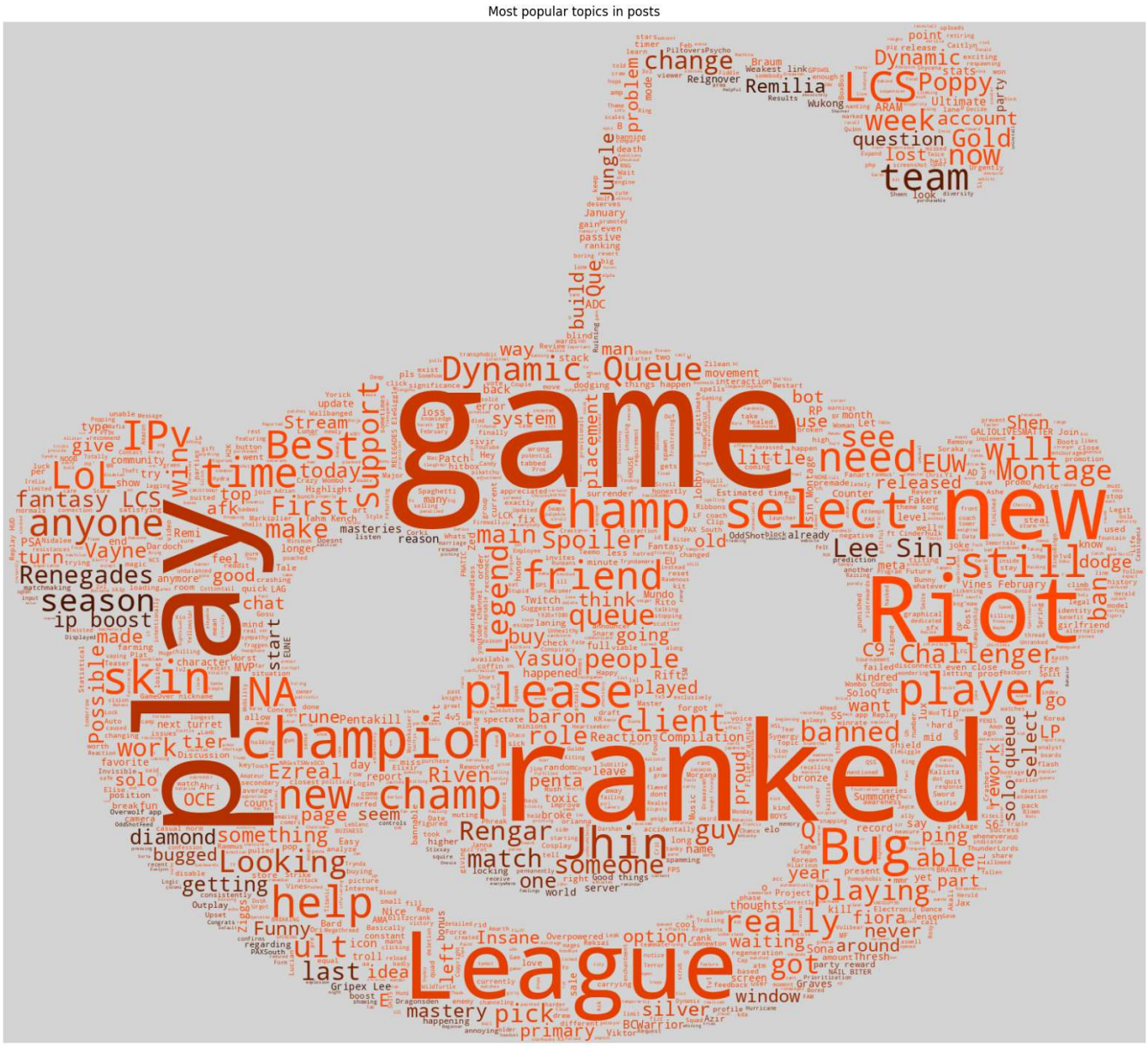
```
text = " ".join([i for i in kdf.selftext.to_numpy() if len(i) > 0 and i != "[removed]" and i!="[deleted]"])
wordcloud = WordCloud(max_words=5000, mask=~mask[:,:,0], background_color="white", mode="RGBA").generate(text)
# create coloring from image
image_colors = ImageColorGenerator(mask)
plt.figure(figsize=(20,20))
plt.imshow(wordcloud.recolor(color_func=image_colors), interpolation="bilinear")
plt.title("Most popular topics in posts")
plt.axis("off");
```




```
text = " ".join([i for i in kdf.title.to_numpy() if len(i) > 0 and i != "[removed]" and i!="[deleted]"])
wordcloud = WordCloud(max_words=5000, mask=~mask[:,:,0], background_color="white", mode="RGBA").generate(text)
# create coloring from image
image_colors = ImageColorGenerator(mask)
plt.figure(figsize=(20,20))
plt.imshow(wordcloud.recolor(color_func=image_colors), interpolation="bilinear")
plt.title("Most popular topics in title")
plt.axis("off");
```


```
data = kdf[kdf['subreddit'] == 'leagueoflegends']

text = " ".join([i for i in data["title"].to_numpy() if len(i) > 0 and i != "[removed]" and i!="[deleted]"])
wordcloud = WordCloud(max_words=5000, mask=~mask[:,:,0], background_color="lightgray", mode="RGBA").generate(text)
# create coloring from image
image_colors = ImageColorGenerator(mask)
plt.figure(figsize=(20,20))
plt.imshow(wordcloud.recolor(color_func=image_colors), interpolation="bilinear")
plt.title("Most popular topics in posts")
plt.axis("off");
```



Assignment question

Choose a subreddit you like and build a worldcloud using Koalas. Feel free to change the mask or the colors...

Answer:

Como podemos observar en la casilla de encima, hemos filtrado los resultados a coger solamente el texto contenido dentro de los titulos del subreddit "leagueoflegends", pero cambiando esa variable funcionaria con cualquiera, y los resultados nos quedan tal y como vemos en la imagen superior. A demas hemos probado a cambiar el color para verlo algo distinto, lo cual tambien ha funcionado correctamente.