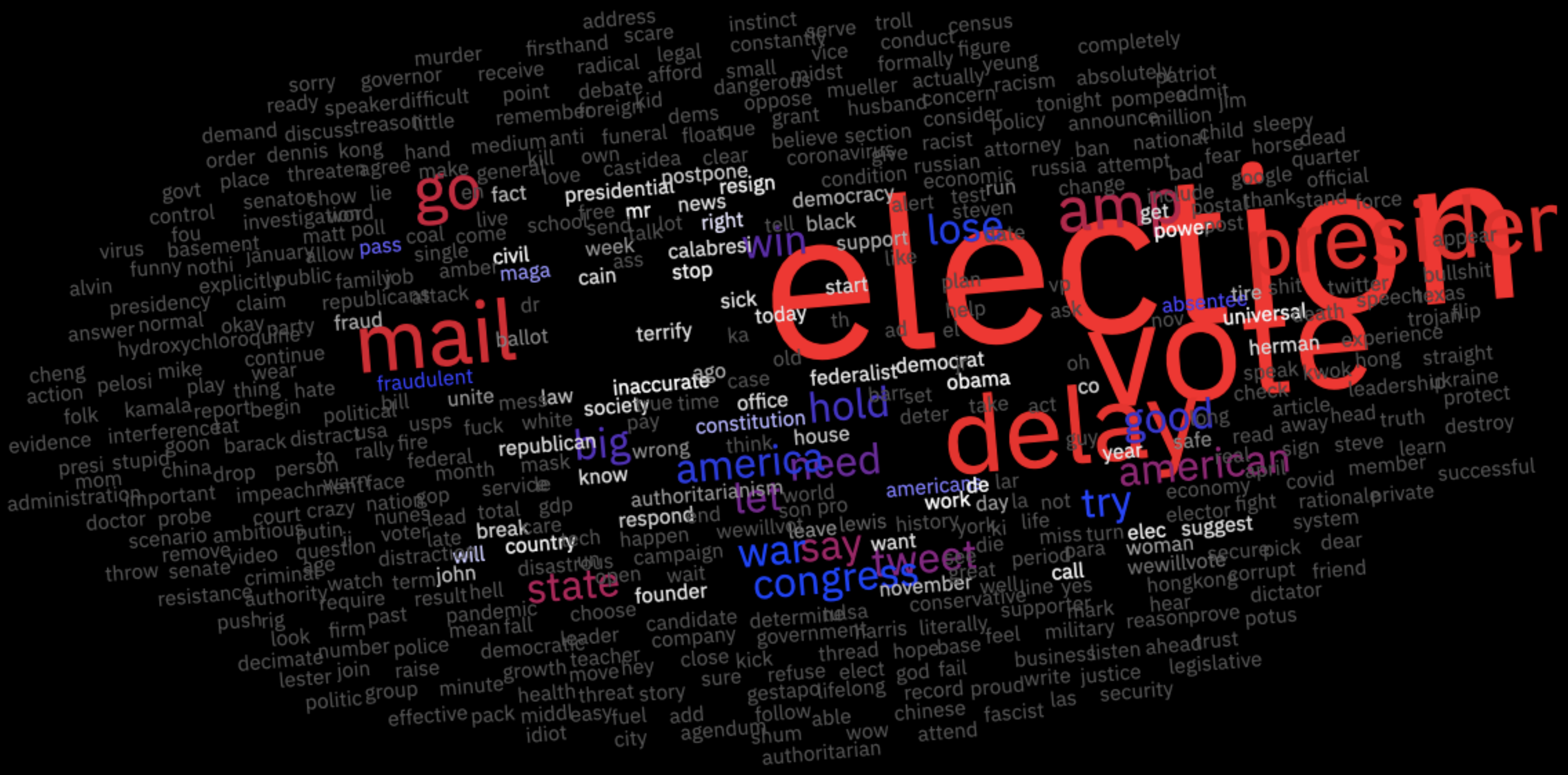
Tag cloud
(US Elections 2020 Tweets)



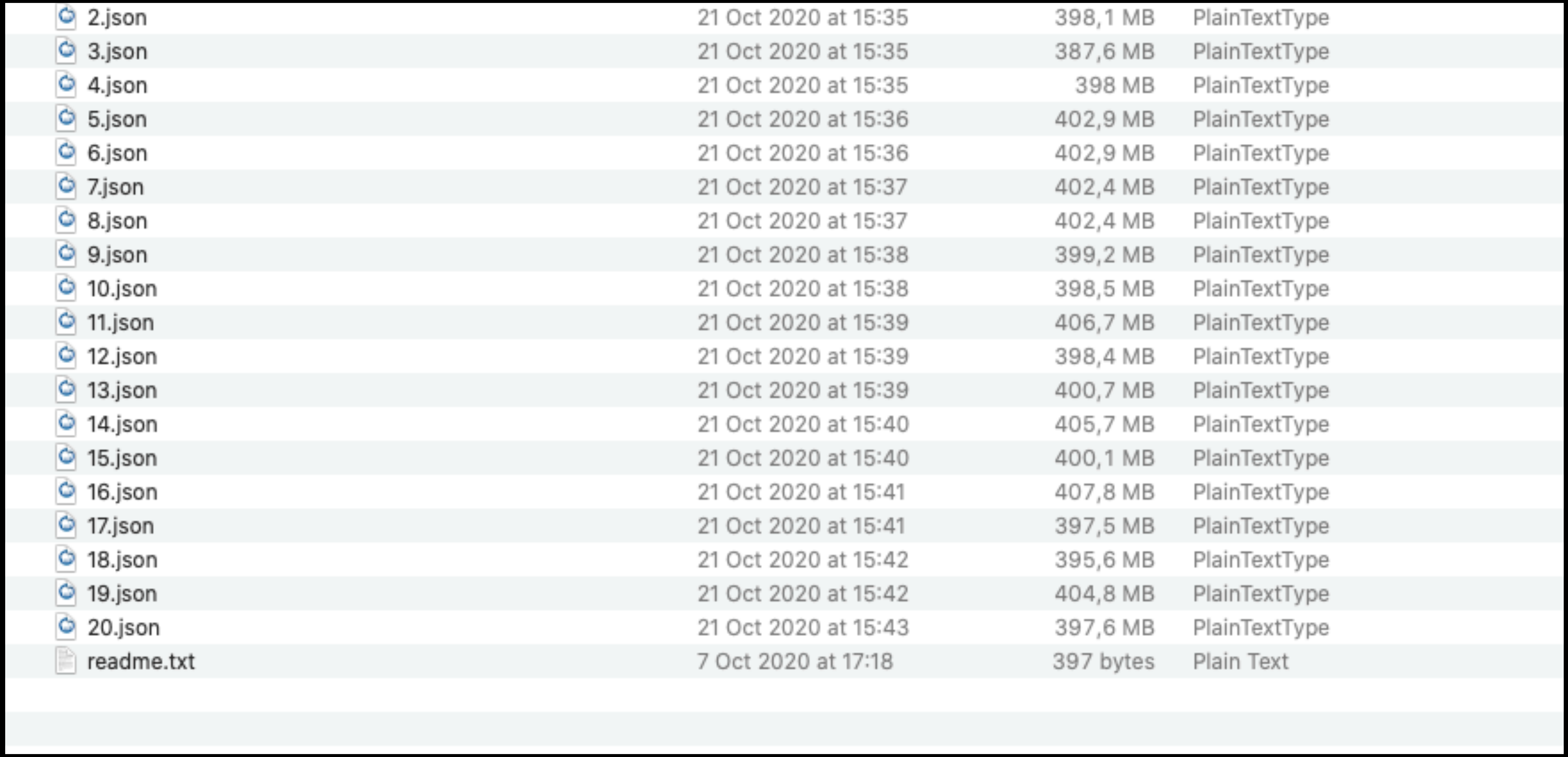
ROUSSEAU Alexy (21910036) - KESOURI Manil (21914480)

We trough this project is simple…

spaCy

- Use spaCy as shown in TPs,
- Develop a Python loop,
- Count in a dictionary occurrences,

Et voilà !



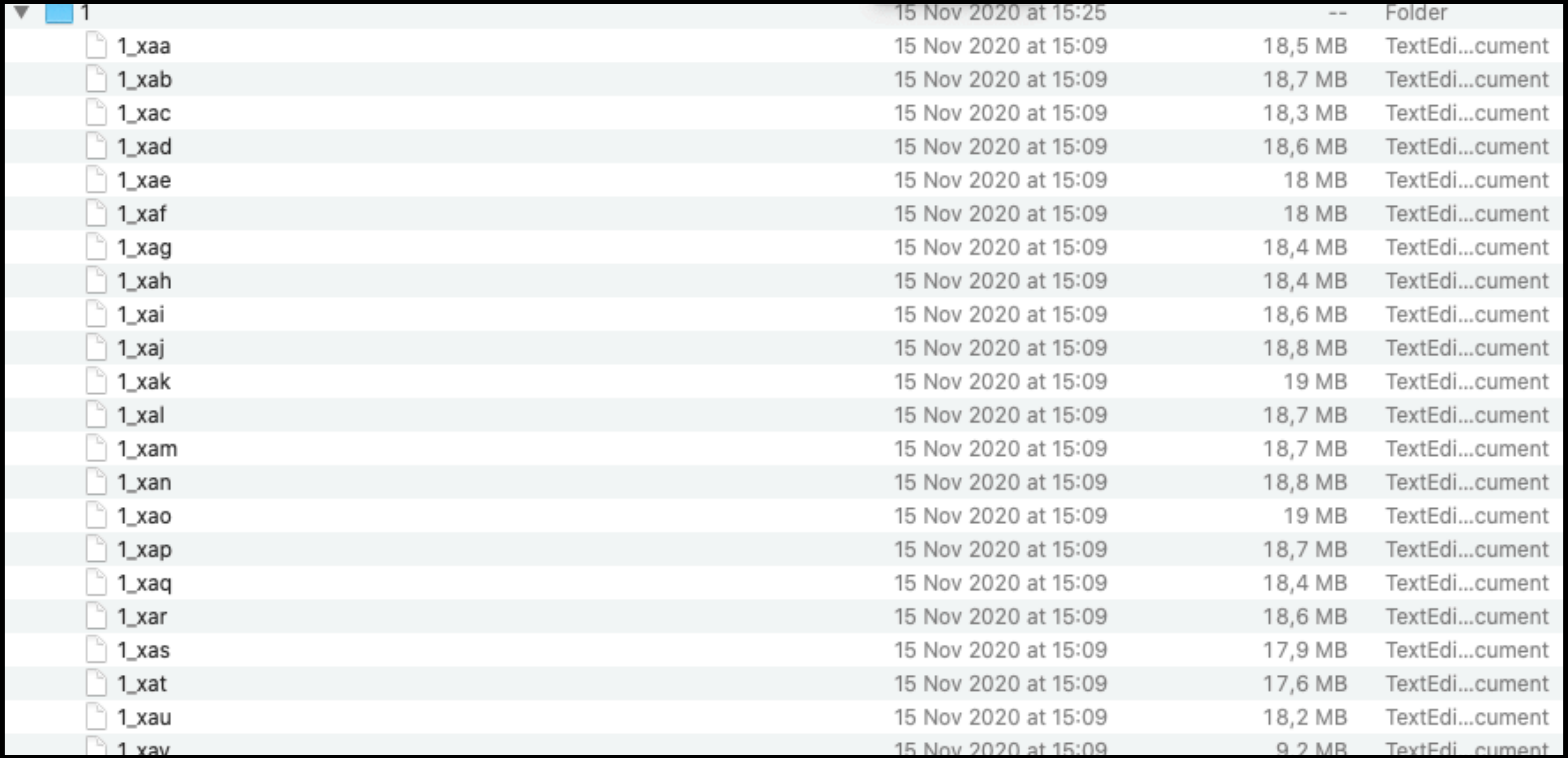
2.json	21 Oct 2020 at 15:35	398,1 MB	PlainTextType
3.json	21 Oct 2020 at 15:35	387,6 MB	PlainTextType
4.json	21 Oct 2020 at 15:35	398 MB	PlainTextType
5.json	21 Oct 2020 at 15:36	402,9 MB	PlainTextType
6.json	21 Oct 2020 at 15:36	402,9 MB	PlainTextType
7.json	21 Oct 2020 at 15:37	402,4 MB	PlainTextType
8.json	21 Oct 2020 at 15:37	402,4 MB	PlainTextType
9.json	21 Oct 2020 at 15:38	399,2 MB	PlainTextType
10.json	21 Oct 2020 at 15:38	398,5 MB	PlainTextType
11.json	21 Oct 2020 at 15:39	406,7 MB	PlainTextType
12.json	21 Oct 2020 at 15:39	398,4 MB	PlainTextType
13.json	21 Oct 2020 at 15:39	400,7 MB	PlainTextType
14.json	21 Oct 2020 at 15:40	405,7 MB	PlainTextType
15.json	21 Oct 2020 at 15:40	400,1 MB	PlainTextType
16.json	21 Oct 2020 at 15:41	407,8 MB	PlainTextType
17.json	21 Oct 2020 at 15:41	397,5 MB	PlainTextType
18.json	21 Oct 2020 at 15:42	395,6 MB	PlainTextType
19.json	21 Oct 2020 at 15:42	404,8 MB	PlainTextType
20.json	21 Oct 2020 at 15:43	397,6 MB	PlainTextType
readme.txt	7 Oct 2020 at 17:18	397 bytes	Plain Text

But it just doesn’t works…

- Files are too big to parse…
- We’ll finish the work in 1000 hours with this approach…

We need another way to compute

Is it possible to parallelise this lecture?
And how to do?

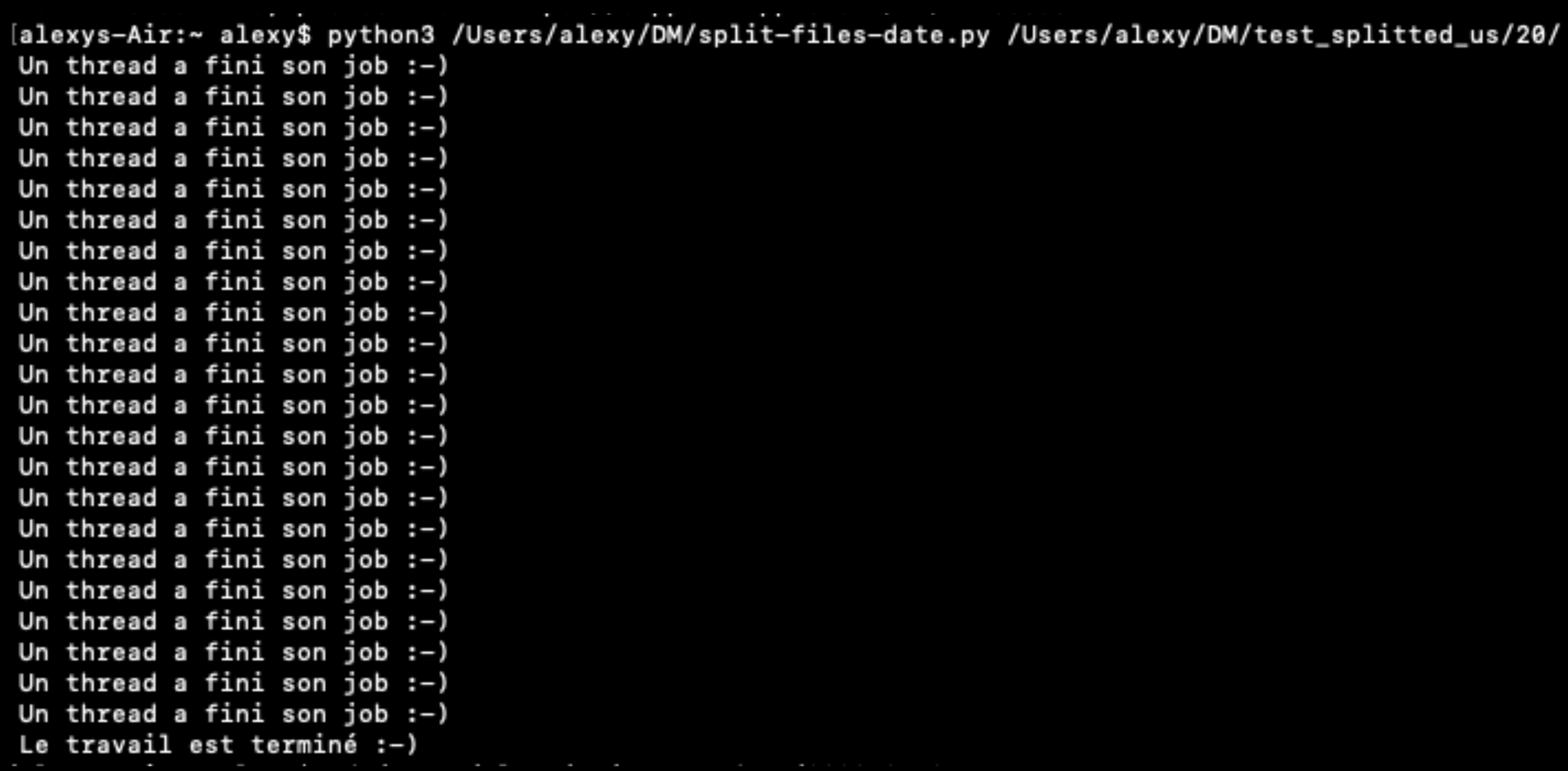


1	15 Nov 2020 at 15:25	--	Folder
1_xaa	15 Nov 2020 at 15:09	18,5 MB	TextEdi...cument
1_xab	15 Nov 2020 at 15:09	18,7 MB	TextEdi...cument
1_xac	15 Nov 2020 at 15:09	18,3 MB	TextEdi...cument
1_xad	15 Nov 2020 at 15:09	18,6 MB	TextEdi...cument
1_xae	15 Nov 2020 at 15:09	18 MB	TextEdi...cument
1_xaf	15 Nov 2020 at 15:09	18 MB	TextEdi...cument
1_xag	15 Nov 2020 at 15:09	18,4 MB	TextEdi...cument
1_xah	15 Nov 2020 at 15:09	18,4 MB	TextEdi...cument
1_xai	15 Nov 2020 at 15:09	18,6 MB	TextEdi...cument
1_xaj	15 Nov 2020 at 15:09	18,8 MB	TextEdi...cument
1_xak	15 Nov 2020 at 15:09	19 MB	TextEdi...cument
1_xal	15 Nov 2020 at 15:09	18,7 MB	TextEdi...cument
1_xam	15 Nov 2020 at 15:09	18,7 MB	TextEdi...cument
1_xan	15 Nov 2020 at 15:09	18,8 MB	TextEdi...cument
1_xao	15 Nov 2020 at 15:09	19 MB	TextEdi...cument
1_xap	15 Nov 2020 at 15:09	18,7 MB	TextEdi...cument
1_xaq	15 Nov 2020 at 15:09	18,4 MB	TextEdi...cument
1_xar	15 Nov 2020 at 15:09	18,6 MB	TextEdi...cument
1_xas	15 Nov 2020 at 15:09	17,9 MB	TextEdi...cument
1_xat	15 Nov 2020 at 15:09	17,6 MB	TextEdi...cument
1_xau	15 Nov 2020 at 15:09	18,2 MB	TextEdi...cument
1_xav	15 Nov 2020 at 15:09	9,2 MB	TextEdi...cument

First step: “cut the cake”

By using the shell command split
“split -l 100000 {fileName}.{ext}

Each part is in specific folder and prefixed by the name of the parent directory (it’ll be important).



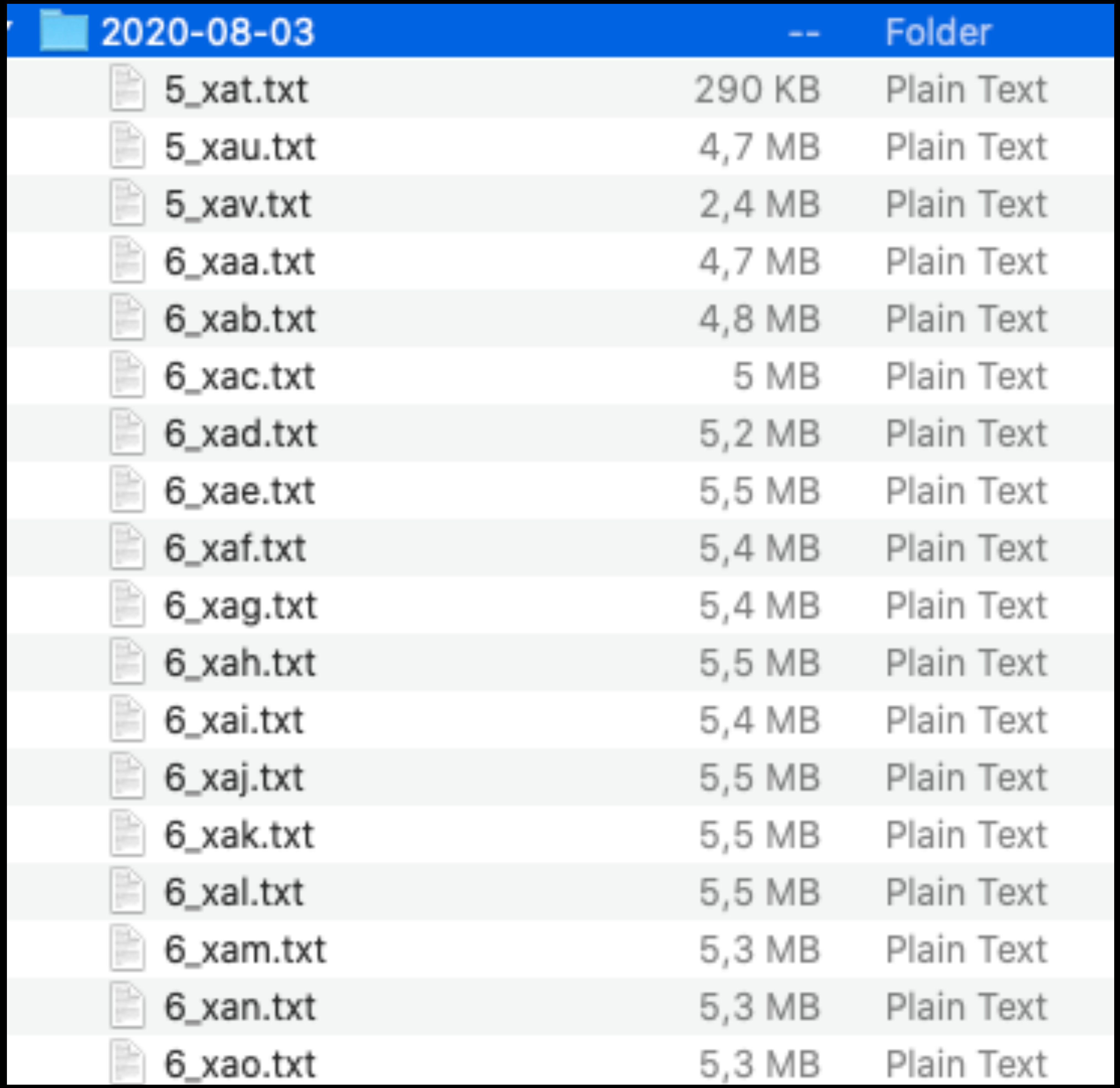
```
[alexys-Air:~ alexy$ python3 /Users/alexy/DM/split-files-date.py /Users/alexy/DM/test_splitted_us/20/
Un thread a fini son job :-)
Un thread a fini son job :-)
Un thread a fini son job :-)
Un thread a fini son job :-)
Un thread a fini son job :-)
Un thread a fini son job :-)
Un thread a fini son job :-)
Un thread a fini son job :-)
Un thread a fini son job :-)
Un thread a fini son job :-)
Un thread a fini son job :-)
Un thread a fini son job :-)
Un thread a fini son job :-)
Un thread a fini son job :-)
Un thread a fini son job :-)
Un thread a fini son job :-)
Un thread a fini son job :-)
Un thread a fini son job :-)
Un thread a fini son job :-)
Un thread a fini son job :-)
Un thread a fini son job :-)
Un thread a fini son job :-)
Le travail est terminé :-)
```

Second step: Launch python

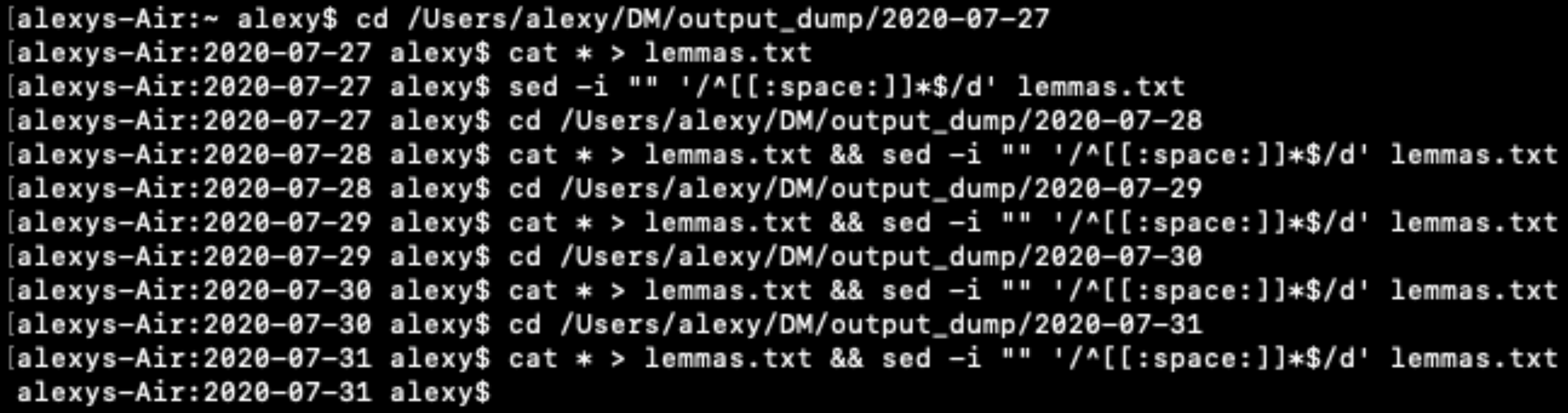
- Launch one time per folder (multiprocessing)
- The script creates one thread per splitted part
- In this example we had 20 process and 22 threads

The script works like that

- Each process reads its tweets
- In the first loop the script verifies if the folder of the day exists.
- If the folder exists it writes on it, else it creates it before writing.
- Off course we are using spaCy to lemmatise the text before writing it in a file.
- Remember prefixes before parts names : with them we'll not encounter a conflict with two files with the same name for the same day !
- Example : we can have 1_xaa et 2_xaa in the folder 2020-07-01 without any problem!



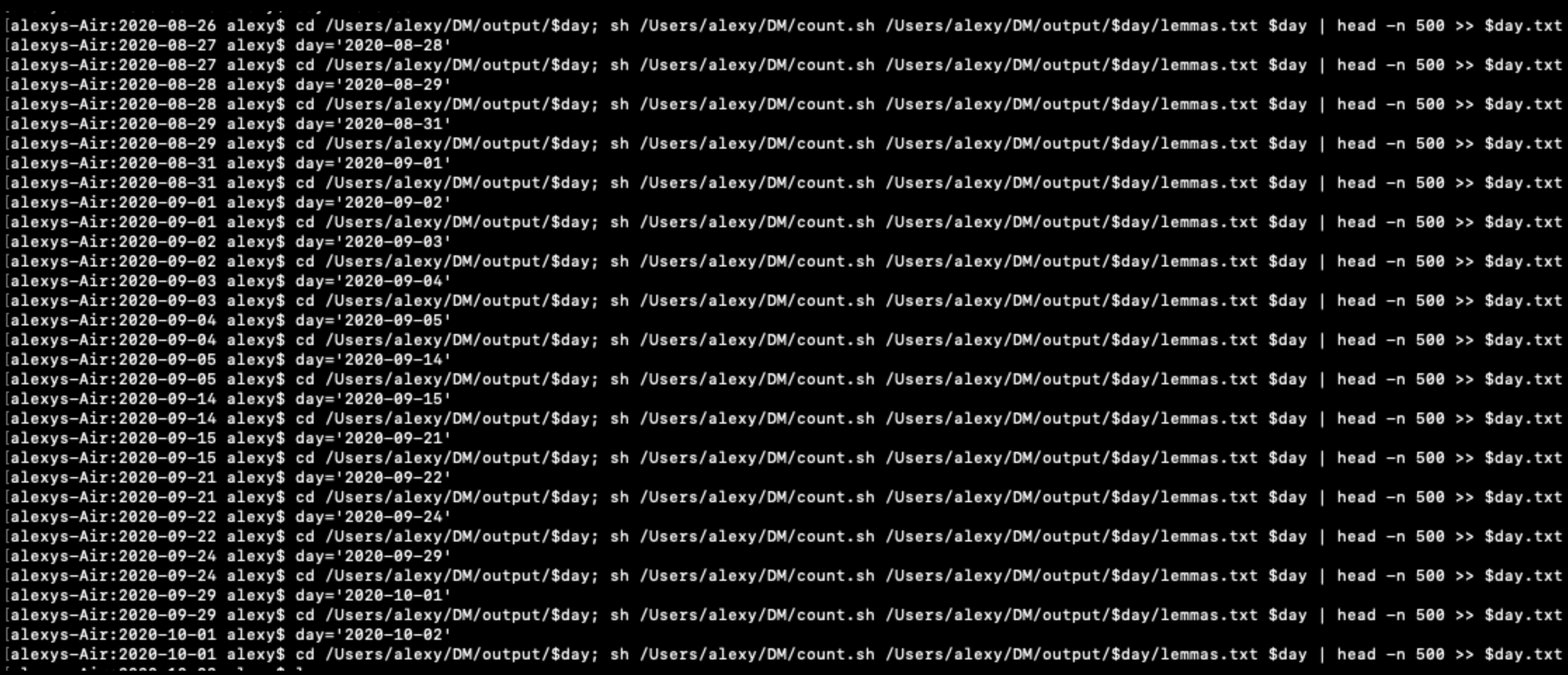
2020-08-03	--	Folder
5_xat.txt	290 KB	Plain Text
5_xau.txt	4,7 MB	Plain Text
5_xav.txt	2,4 MB	Plain Text
6_xaa.txt	4,7 MB	Plain Text
6_xab.txt	4,8 MB	Plain Text
6_xac.txt	5 MB	Plain Text
6_xad.txt	5,2 MB	Plain Text
6_xae.txt	5,5 MB	Plain Text
6_xaf.txt	5,4 MB	Plain Text
6_xag.txt	5,4 MB	Plain Text
6_xah.txt	5,5 MB	Plain Text
6_xai.txt	5,4 MB	Plain Text
6_xaj.txt	5,5 MB	Plain Text
6_xak.txt	5,5 MB	Plain Text
6_xal.txt	5,5 MB	Plain Text
6_xam.txt	5,3 MB	Plain Text
6_xan.txt	5,3 MB	Plain Text
6_xao.txt	5,3 MB	Plain Text



```
[alexys-Air:~ alexy$ cd /Users/alexy/DM/output_dump/2020-07-27
[alexys-Air:2020-07-27 alexy$ cat * > lemmas.txt
[alexys-Air:2020-07-27 alexy$ sed -i "" '/^[[:space:]]*$/d' lemmas.txt
[alexys-Air:2020-07-27 alexy$ cd /Users/alexy/DM/output_dump/2020-07-28
[alexys-Air:2020-07-28 alexy$ cat * > lemmas.txt && sed -i "" '/^[[:space:]]*$/d' lemmas.txt
[alexys-Air:2020-07-28 alexy$ cd /Users/alexy/DM/output_dump/2020-07-29
[alexys-Air:2020-07-29 alexy$ cat * > lemmas.txt && sed -i "" '/^[[:space:]]*$/d' lemmas.txt
[alexys-Air:2020-07-29 alexy$ cd /Users/alexy/DM/output_dump/2020-07-30
[alexys-Air:2020-07-30 alexy$ cat * > lemmas.txt && sed -i "" '/^[[:space:]]*$/d' lemmas.txt
[alexys-Air:2020-07-30 alexy$ cd /Users/alexy/DM/output_dump/2020-07-31
[alexys-Air:2020-07-31 alexy$ cat * > lemmas.txt && sed -i "" '/^[[:space:]]*$/d' lemmas.txt
 alexys-Air:2020-07-31 alexy$
```

Third step : merge (reduce) files of the day

We remove any useless “\n” in the merged file

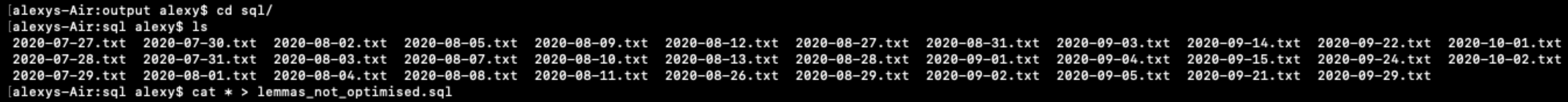


```
[alexys-Air:2020-08-26 alexy$ cd /Users/alexy/DM/output/$day; sh /Users/alexy/DM/count.sh /Users/alexy/DM/output/$day/lemmas.txt $day | head -n 500 >> $day.txt
[alexys-Air:2020-08-27 alexy$ day='2020-08-28'
[alexys-Air:2020-08-27 alexy$ cd /Users/alexy/DM/output/$day; sh /Users/alexy/DM/count.sh /Users/alexy/DM/output/$day/lemmas.txt $day | head -n 500 >> $day.txt
[alexys-Air:2020-08-28 alexy$ day='2020-08-29'
[alexys-Air:2020-08-28 alexy$ cd /Users/alexy/DM/output/$day; sh /Users/alexy/DM/count.sh /Users/alexy/DM/output/$day/lemmas.txt $day | head -n 500 >> $day.txt
[alexys-Air:2020-08-29 alexy$ day='2020-08-31'
[alexys-Air:2020-08-29 alexy$ cd /Users/alexy/DM/output/$day; sh /Users/alexy/DM/count.sh /Users/alexy/DM/output/$day/lemmas.txt $day | head -n 500 >> $day.txt
[alexys-Air:2020-08-31 alexy$ day='2020-09-01'
[alexys-Air:2020-08-31 alexy$ cd /Users/alexy/DM/output/$day; sh /Users/alexy/DM/count.sh /Users/alexy/DM/output/$day/lemmas.txt $day | head -n 500 >> $day.txt
[alexys-Air:2020-09-01 alexy$ day='2020-09-02'
[alexys-Air:2020-09-01 alexy$ cd /Users/alexy/DM/output/$day; sh /Users/alexy/DM/count.sh /Users/alexy/DM/output/$day/lemmas.txt $day | head -n 500 >> $day.txt
[alexys-Air:2020-09-02 alexy$ day='2020-09-03'
[alexys-Air:2020-09-02 alexy$ cd /Users/alexy/DM/output/$day; sh /Users/alexy/DM/count.sh /Users/alexy/DM/output/$day/lemmas.txt $day | head -n 500 >> $day.txt
[alexys-Air:2020-09-03 alexy$ day='2020-09-04'
[alexys-Air:2020-09-03 alexy$ cd /Users/alexy/DM/output/$day; sh /Users/alexy/DM/count.sh /Users/alexy/DM/output/$day/lemmas.txt $day | head -n 500 >> $day.txt
[alexys-Air:2020-09-04 alexy$ day='2020-09-05'
[alexys-Air:2020-09-04 alexy$ cd /Users/alexy/DM/output/$day; sh /Users/alexy/DM/count.sh /Users/alexy/DM/output/$day/lemmas.txt $day | head -n 500 >> $day.txt
[alexys-Air:2020-09-05 alexy$ day='2020-09-14'
[alexys-Air:2020-09-05 alexy$ cd /Users/alexy/DM/output/$day; sh /Users/alexy/DM/count.sh /Users/alexy/DM/output/$day/lemmas.txt $day | head -n 500 >> $day.txt
[alexys-Air:2020-09-14 alexy$ day='2020-09-15'
[alexys-Air:2020-09-14 alexy$ cd /Users/alexy/DM/output/$day; sh /Users/alexy/DM/count.sh /Users/alexy/DM/output/$day/lemmas.txt $day | head -n 500 >> $day.txt
[alexys-Air:2020-09-15 alexy$ day='2020-09-21'
[alexys-Air:2020-09-15 alexy$ cd /Users/alexy/DM/output/$day; sh /Users/alexy/DM/count.sh /Users/alexy/DM/output/$day/lemmas.txt $day | head -n 500 >> $day.txt
[alexys-Air:2020-09-21 alexy$ day='2020-09-22'
[alexys-Air:2020-09-21 alexy$ cd /Users/alexy/DM/output/$day; sh /Users/alexy/DM/count.sh /Users/alexy/DM/output/$day/lemmas.txt $day | head -n 500 >> $day.txt
[alexys-Air:2020-09-22 alexy$ day='2020-09-24'
[alexys-Air:2020-09-22 alexy$ cd /Users/alexy/DM/output/$day; sh /Users/alexy/DM/count.sh /Users/alexy/DM/output/$day/lemmas.txt $day | head -n 500 >> $day.txt
[alexys-Air:2020-09-24 alexy$ day='2020-09-29'
[alexys-Air:2020-09-24 alexy$ cd /Users/alexy/DM/output/$day; sh /Users/alexy/DM/count.sh /Users/alexy/DM/output/$day/lemmas.txt $day | head -n 500 >> $day.txt
[alexys-Air:2020-09-29 alexy$ day='2020-10-01'
[alexys-Air:2020-09-29 alexy$ cd /Users/alexy/DM/output/$day; sh /Users/alexy/DM/count.sh /Users/alexy/DM/output/$day/lemmas.txt $day | head -n 500 >> $day.txt
[alexys-Air:2020-10-01 alexy$ day='2020-10-02'
[alexys-Air:2020-10-01 alexy$ cd /Users/alexy/DM/output/$day; sh /Users/alexy/DM/count.sh /Users/alexy/DM/output/$day/lemmas.txt $day | head -n 500 >> $day.txt
```

Fourth step : count occurrences

And grab 500 most used of the day in a sql file

Final step : merge sql files



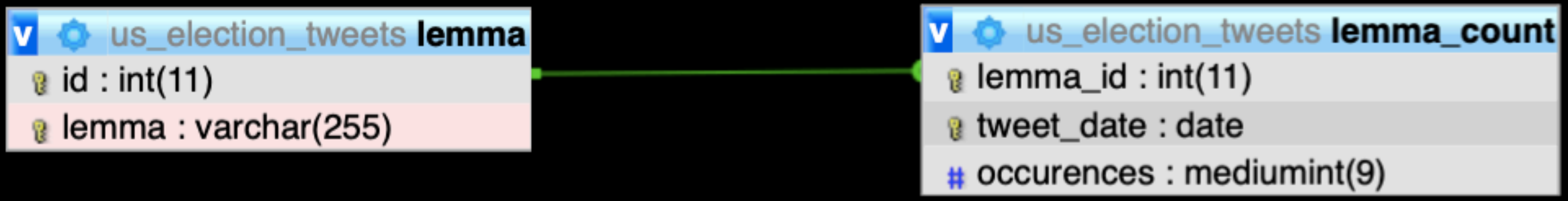
```
[alexys-Air:output alexy$ cd sql/
[alexys-Air:sql alexy$ ls
 2020-07-27.txt  2020-07-30.txt  2020-08-02.txt  2020-08-05.txt  2020-08-09.txt  2020-08-12.txt  2020-08-27.txt  2020-08-31.txt  2020-09-03.txt  2020-09-14.txt  2020-09-22.txt  2020-10-01.txt
 2020-07-28.txt  2020-07-31.txt  2020-08-03.txt  2020-08-07.txt  2020-08-10.txt  2020-08-13.txt  2020-08-28.txt  2020-09-01.txt  2020-09-04.txt  2020-09-15.txt  2020-09-24.txt  2020-10-02.txt
 2020-07-29.txt  2020-08-01.txt  2020-08-04.txt  2020-08-08.txt  2020-08-11.txt  2020-08-26.txt  2020-08-29.txt  2020-09-02.txt  2020-09-05.txt  2020-09-21.txt  2020-09-29.txt
[alexys-Air:sql alexy$ cat * > lemmas_not_optimised.sql
```

Little things to say

- We optimised the storage in the SQL database by creating 2 tables one for each lemma.

- One other for occurrences with a foreign key to avoid the repetition of lemma.

- We have 17500 records and 2938 different lemmas.



us_election_tweets **lemma**
id : int(11)
lemma : varchar(255)

us_election_tweets **lemma_count**
lemma_id : int(11)
tweet_date : date
occurences : mediumint(9)

Little things to say

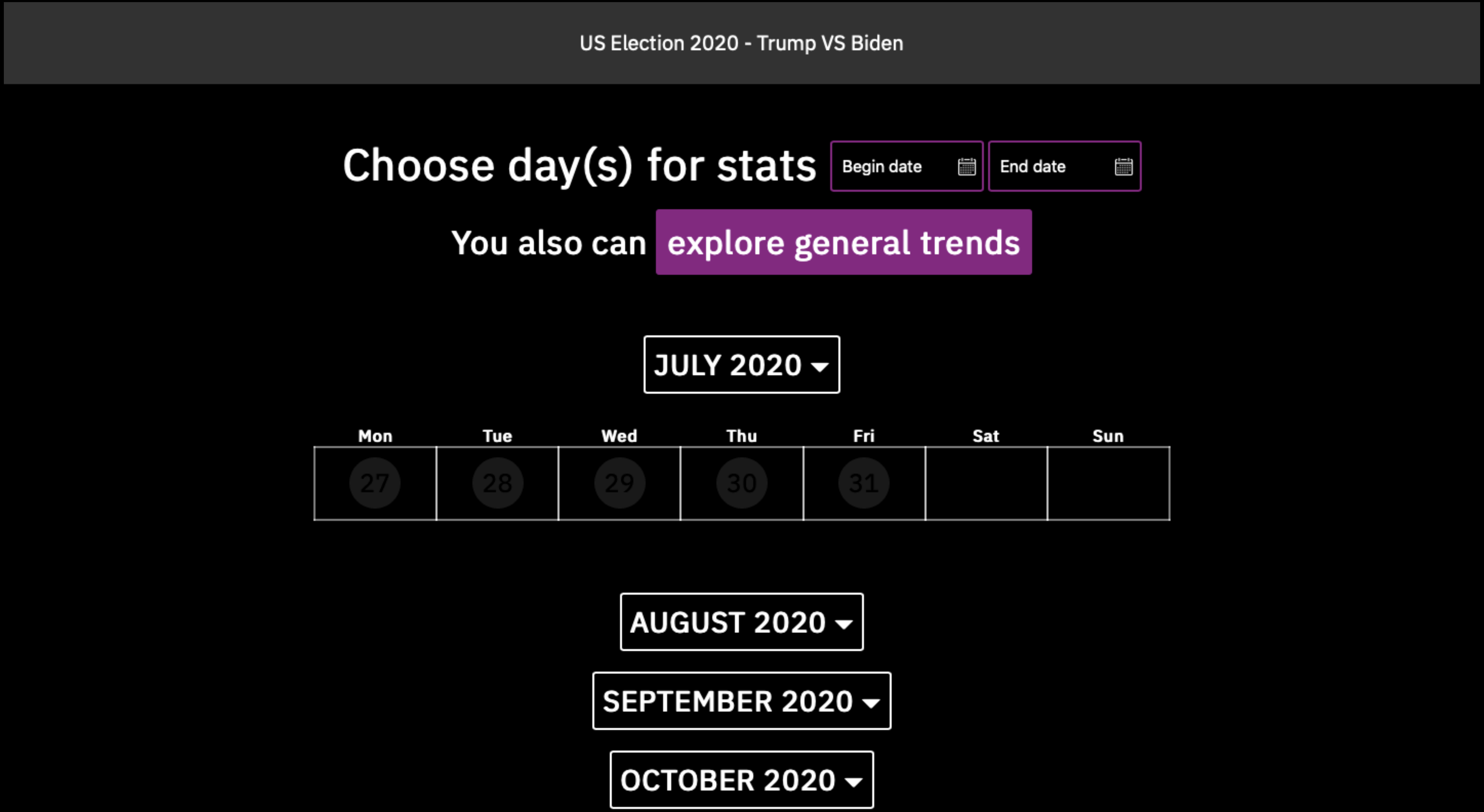
- We removed unused parts of spaCy in the pipeline in order to have a faster analysis of words.

- We removed noise with the 100 tops words used in English plus the frequent words in tweets about elections.

- We also removed tweets with less than 3 characters before lemmatisation.

- We could produce stats for day and hour in order to show the progress of each top word but the deadline was too short to do that with precision. The principle is the same, we just need to separate files per day and hour in folders and produce a 24 times larger table for occurrences.

The Website (homepage)



US Election 2020 - Trump VS Biden

Choose day(s) for stats

You also can

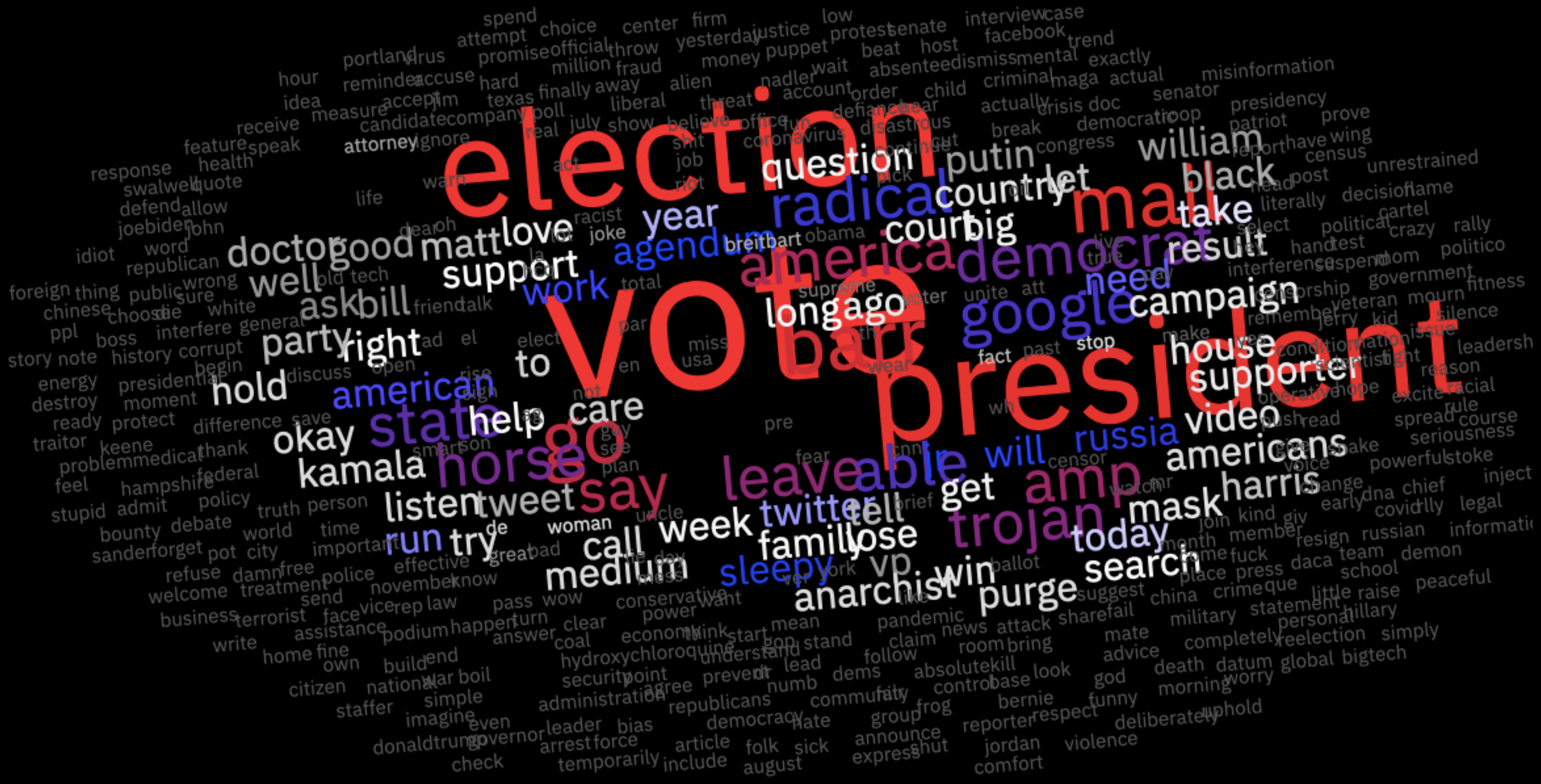
Mon	Tue	Wed	Thu	Fri	Sat	Sun
27	28	29	30	31		

The Website (results)

5 Most used words (2020-07-29)

Lemma	Occurences	% of usage
vote	197056	100
election	119026	60
president	119024	60
mail	73748	37
barr	61546	31

Tagcloud


vote
election
president
mail
barr
democrat
america
radical
google
question
putin
country
let
court
big
william
black
take
result
campaign
house
supporter
video
americans
harris
mask
today
search
purge
win
anarchist
trojan
amp
get
will
russia
able
leave
twitter
tell
family
lose
vp
sleepy
week
call
medium
say
horse
go
state
american
help
care
okay
kamala
listen
tweet
run
try
hold
party
fight
ask
bill
well
doctor
good
matt
love
support
work
year
agenda
breitbart
longago
need
to
spend
choice
center
firm
low
interview
case
facebook
trend
portland
virus
attempt
promise
official
throw
yesterday
justice
protest
senate
beat
host
absentee
dismiss
mental
exactly
hour
remind
accuse
hard
million
fraud
money
puppet
wait
idea
measure
accept
finally
away
alien
nadler
account
order
child
criminal
maga
actual
misinformation
receive
candidate
company
poll
texas
liberal
threat
senator
presidency
feature
speak
attorney
ignore
real
july
show
believe
office
coronavirus
crisis
doc
democratic
troop
patriot
prove
response
health
disastrous
break
congress
report
have
wing
census
swalwell
quote
unrestrained
defend
allow
life
decision
flame
joebiden
john
racist
joke
literally
cartel
political
crazy
rally
idiot
word
obama
hand
test
politico
republican
interference
suspend
mom
government
foreign
thing
wrong
supreme
fitness
chinese
public
sure
unite
veteran
mourn
choose
white
silence
ppl
boss
interfere
general
fact
past
stop
leadership
story
note
history
corrupt
begin
discuss
open
miss
usa
reason
energy
president
moment
racial
destroy
difference
save
spread
course
ready
protect
seriousness
traitor
keene
medicare
thank
smart
son
fear
powerful
stoke
problem
feel
hampshire
federal
plan
inject
stupid
admit
policy
truth
person
early
dna
chief
covid
lly
legal
bounty
debate
world
time
woman
kind
member
information
sander
forget
pot
city
important
great
bad
resign
russian
team
demon
refuse
damn
free
effective
fuck
press
daca
school
welcome
treatment
police
november
know
place
crime
que
little
raise
peaceful
send
pass
wow
conservative
want
china
statement
hillary
business
terrorist
face
vice
rep
law
power
suggest
share
fail
military
personal
assistance
podium
happen
turn
clear
pandemic
news
attack
mate
completely
reelection
simply
write
home
fine
answer
coal
economy
think
start
mean
gop
stand
claim
room
bring
advice
death
datum
global
bigtech
own
build
end
hydroxychloroquine
understand
follow
absolute
kill
citizen
national
war
boil
security
point
prevent
lead
dems
control
base
look
god
morning
worry
staffer
simple
administration
agree
numb
community
fully
bernie
funny
imagine
even
republicans
democracy
hate
group
frog
reporter
respect
deliberately
uphold
donaldtrump
governor
leader
bias
article
folk
sick
announce
jordan
violence
check
arrest
force
temporarily
include
august
express
shut
comfort