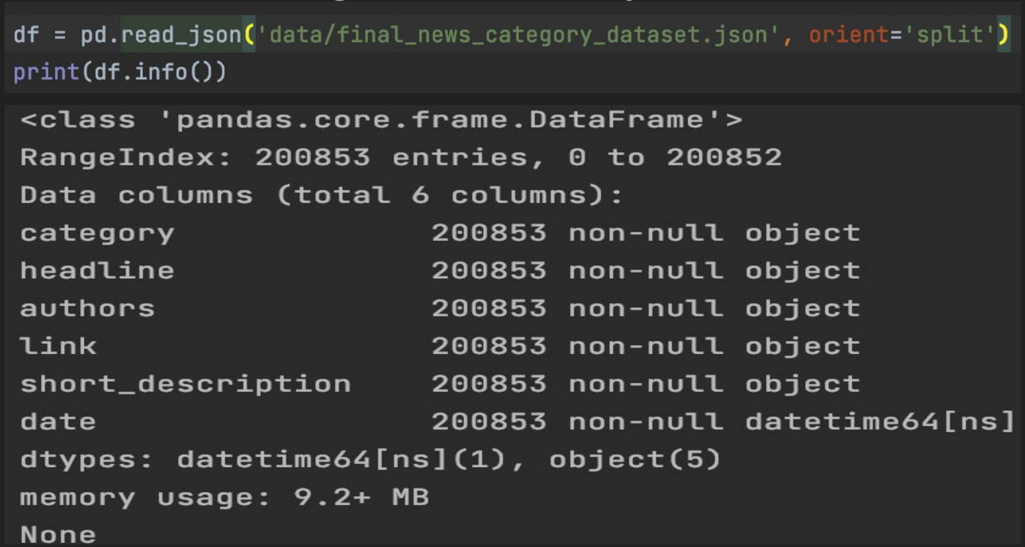
EDA

About the data file

- Initially is a JSON-file (hacked, contents of the file had to be manually reformatted so that python libraries (like json, pandas) could process the proper JSON-file)
- Further, in order to avoid rewriting some existing methods for data analysis, the reformatted .JSON file is read via pandas

Initial info about the data

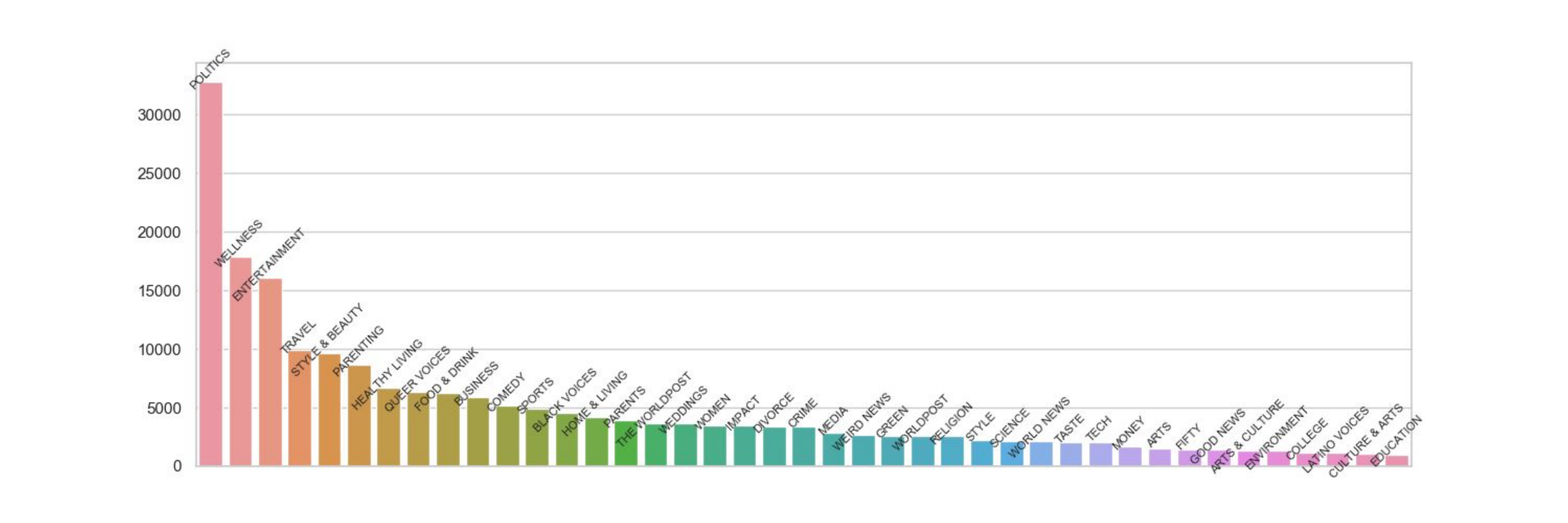
- Data has 6 columns: category, headline, authors, link, short_description, date
- Total number of news (length of the dataframe) is 200853 (news from 28/01/2012 to 26/05/2018
- There are no missing values in any of the column vectors



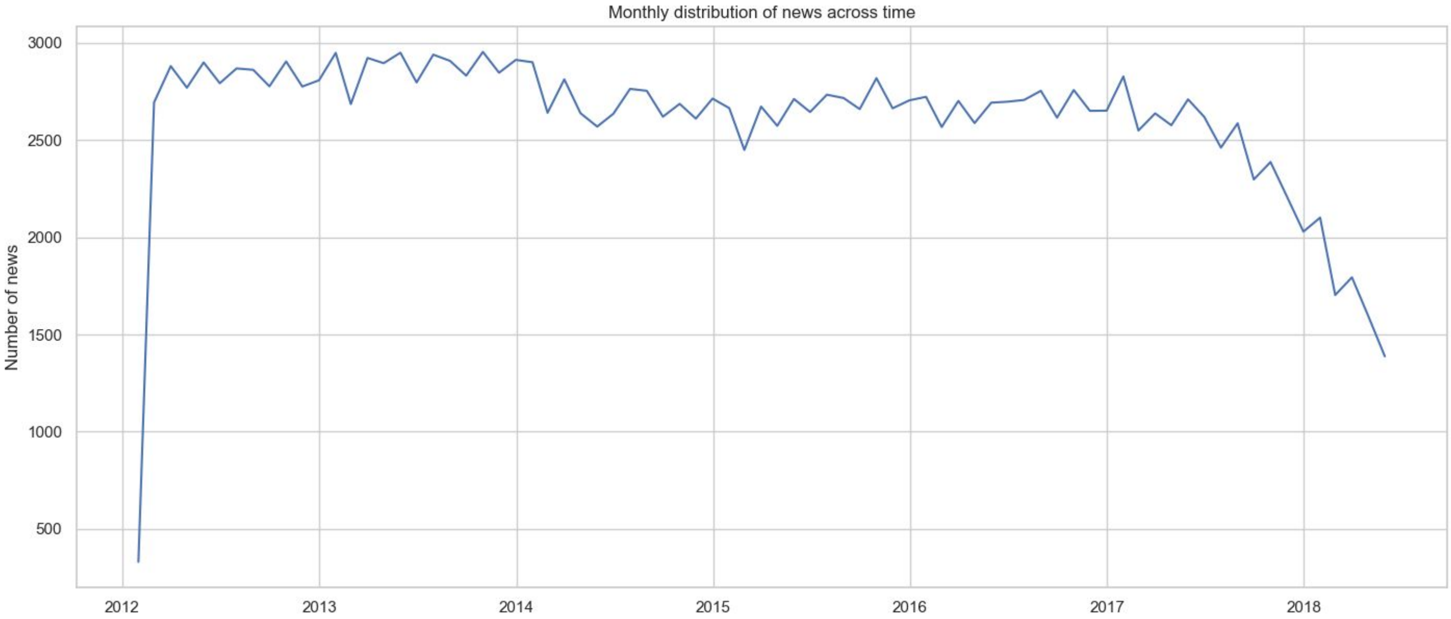
```
df = pd.read_json('data/final_news_category_dataset.json', orient='split')
print(df.info())
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 200853 entries, 0 to 200852
Data columns (total 6 columns):
category             200853 non-null object
headline             200853 non-null object
authors              200853 non-null object
link                 200853 non-null object
short_description    200853 non-null object
date                 200853 non-null datetime64[ns]
dtypes: datetime64[ns](1), object(5)
memory usage: 9.2+ MB
None
```

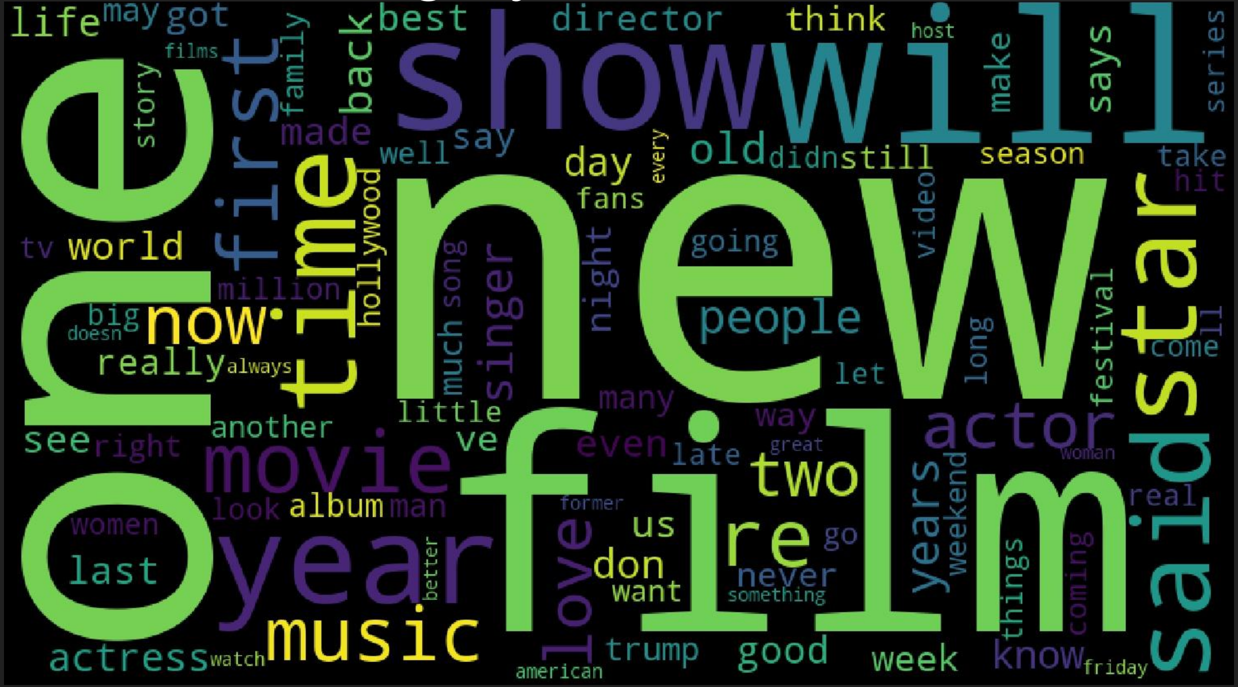
Distribution of news across categories



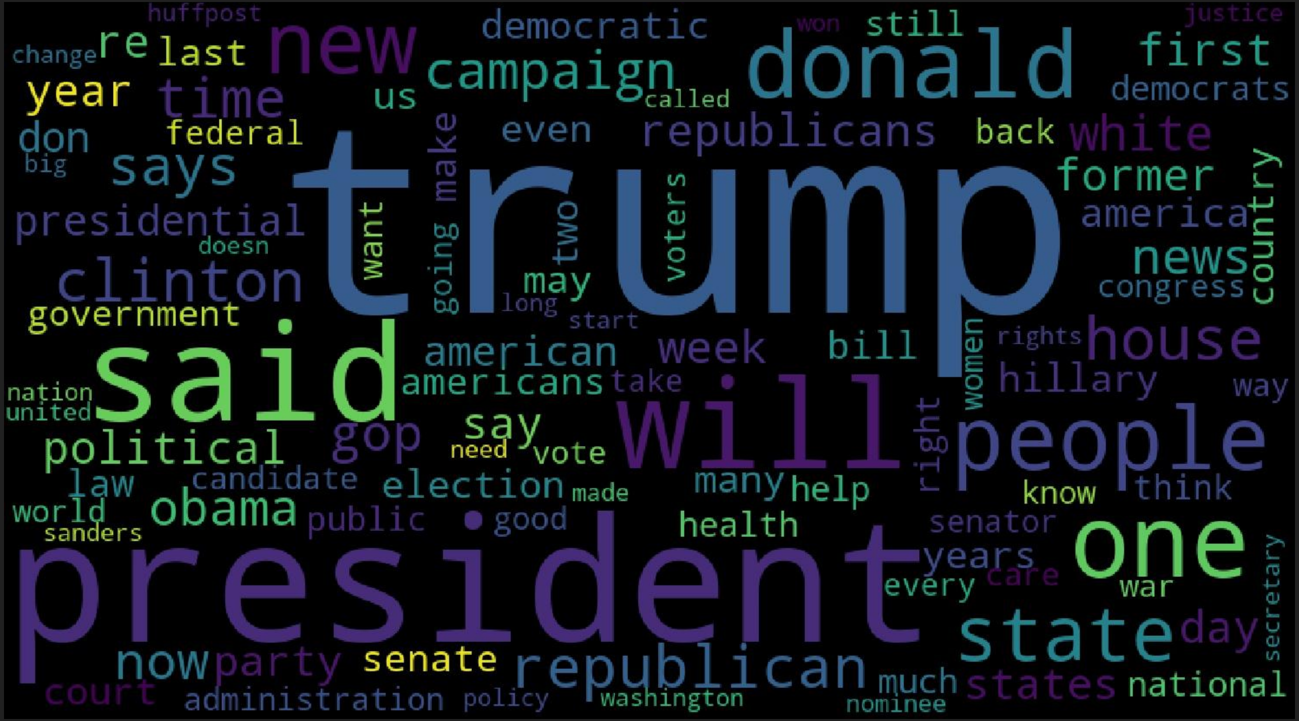
Distribution of news through time



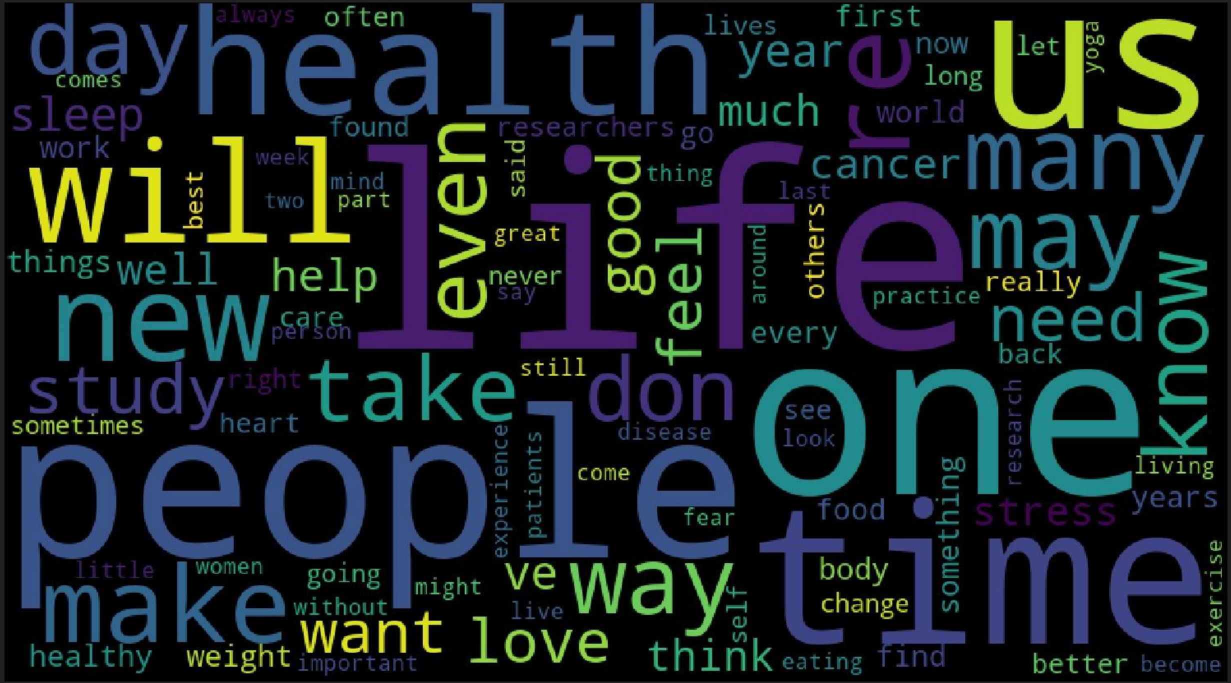
Wordclouds, consisting of 100 most common words, (for the three most frequent categories), starting for Entertainment category of news



For Politics category



For Wellness category of news


day
always
health
often
lives
first
us
comes
year
re
now
let
yoga
long
sleep
found
researchers
much
go
world
work
week
will
life
even
said
thing
cancer
many
mind
two
part
best
good
last
may
great
things
well
help
never
feel
around
others
really
practice
new
say
know
care
person
need
every
back
study
right
take
still
don
one
sometimes
heart
disease
see
look
people
experience
patients
come
research
living
stress
years
food
fear
something
little
women
going
might
ve
way
time
body
change
without
live
self
exercise
want
love
healthy
weight
important
think
eating
find
better
become