MOVIES

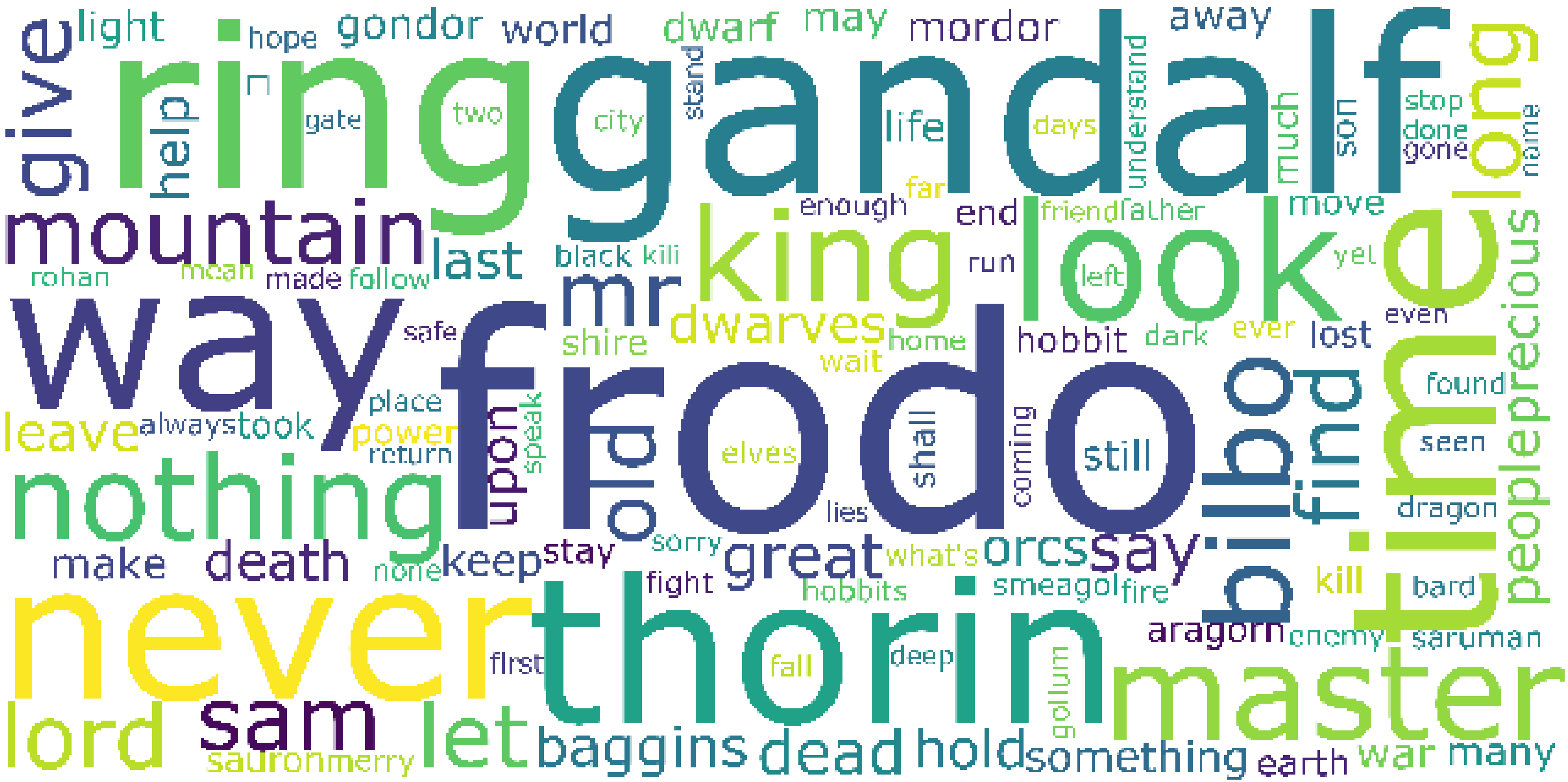
Analyzing subtitles

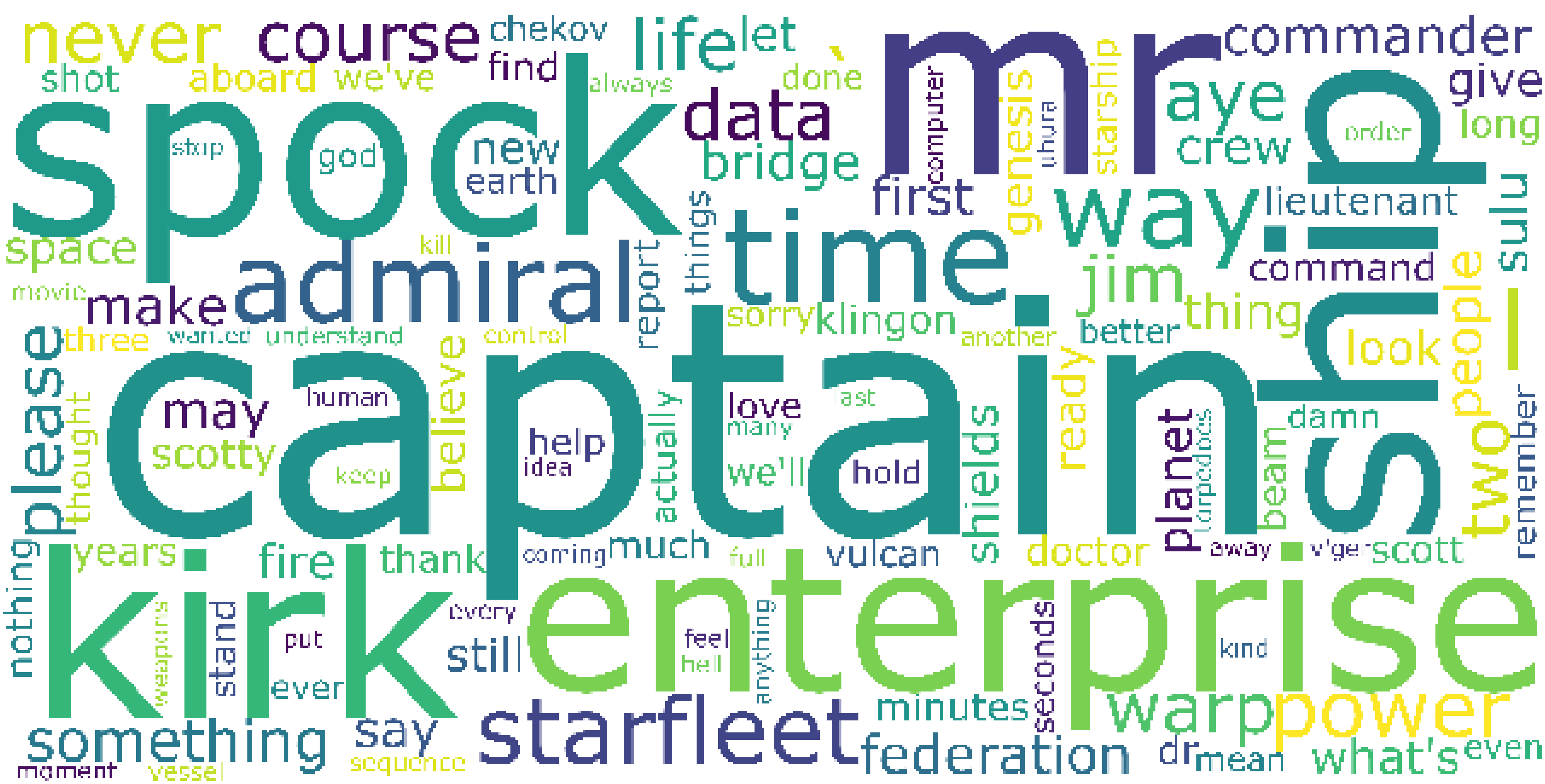
Samer Ghosn

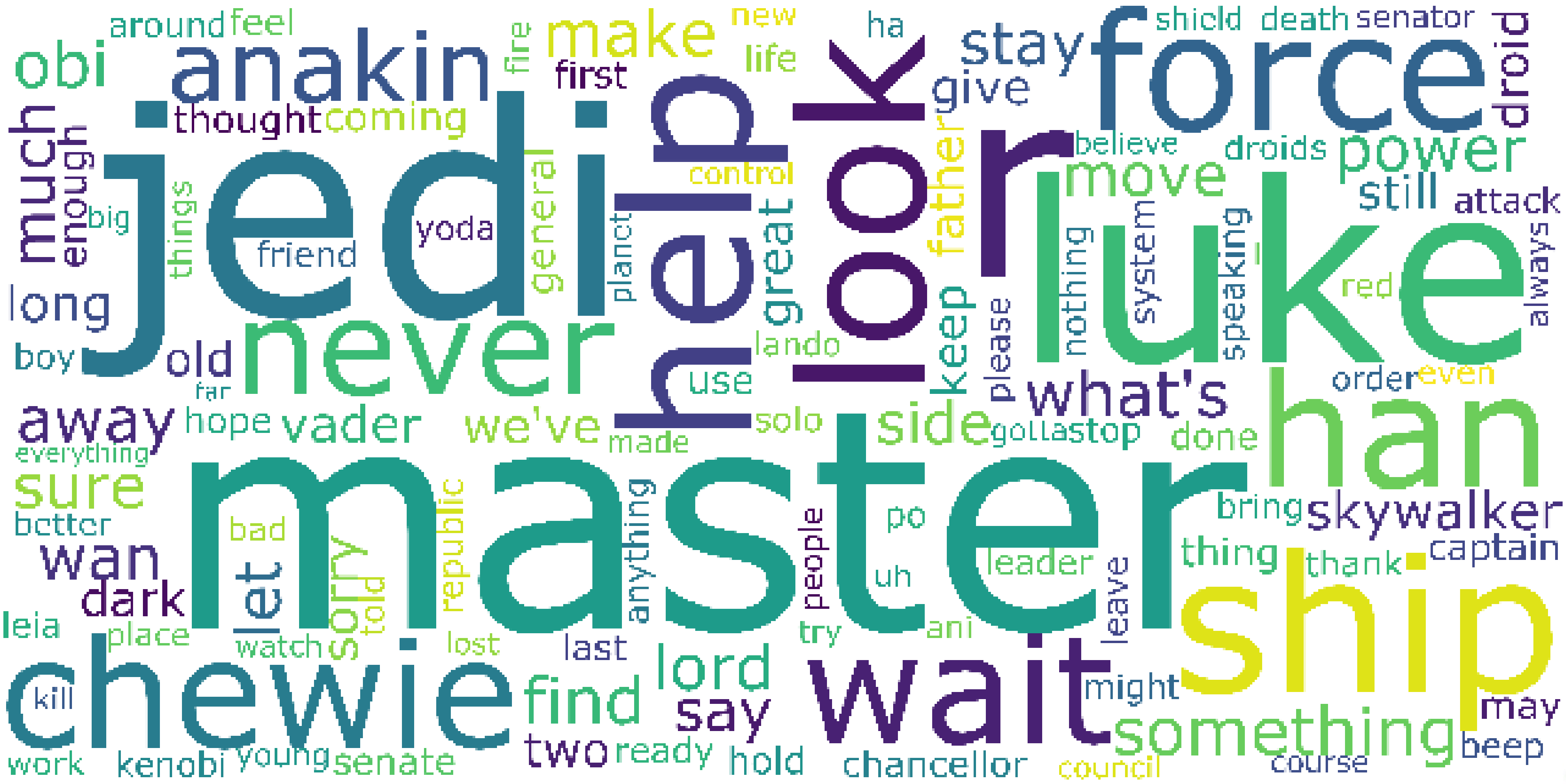
MOVIE SCRIPTS

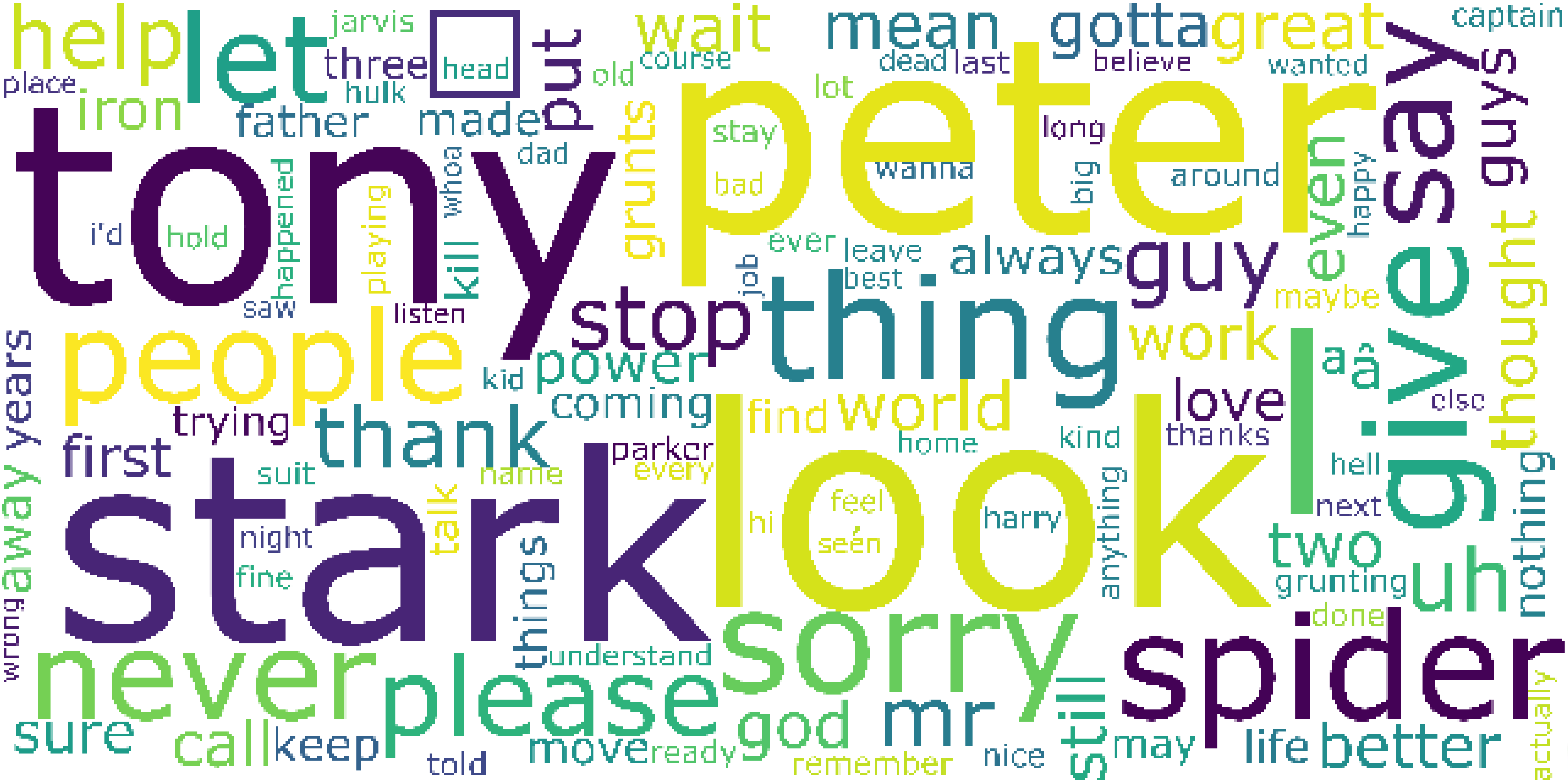
- Analyzing the text in a movie and classify it accordingly.
- Movies scripts contain words that reveal their genre.
- Analyzing the frequency of these words in the subtitle files.
- If the words “Space”, “Ship” appear frequently then the movie is “Sci-fi”

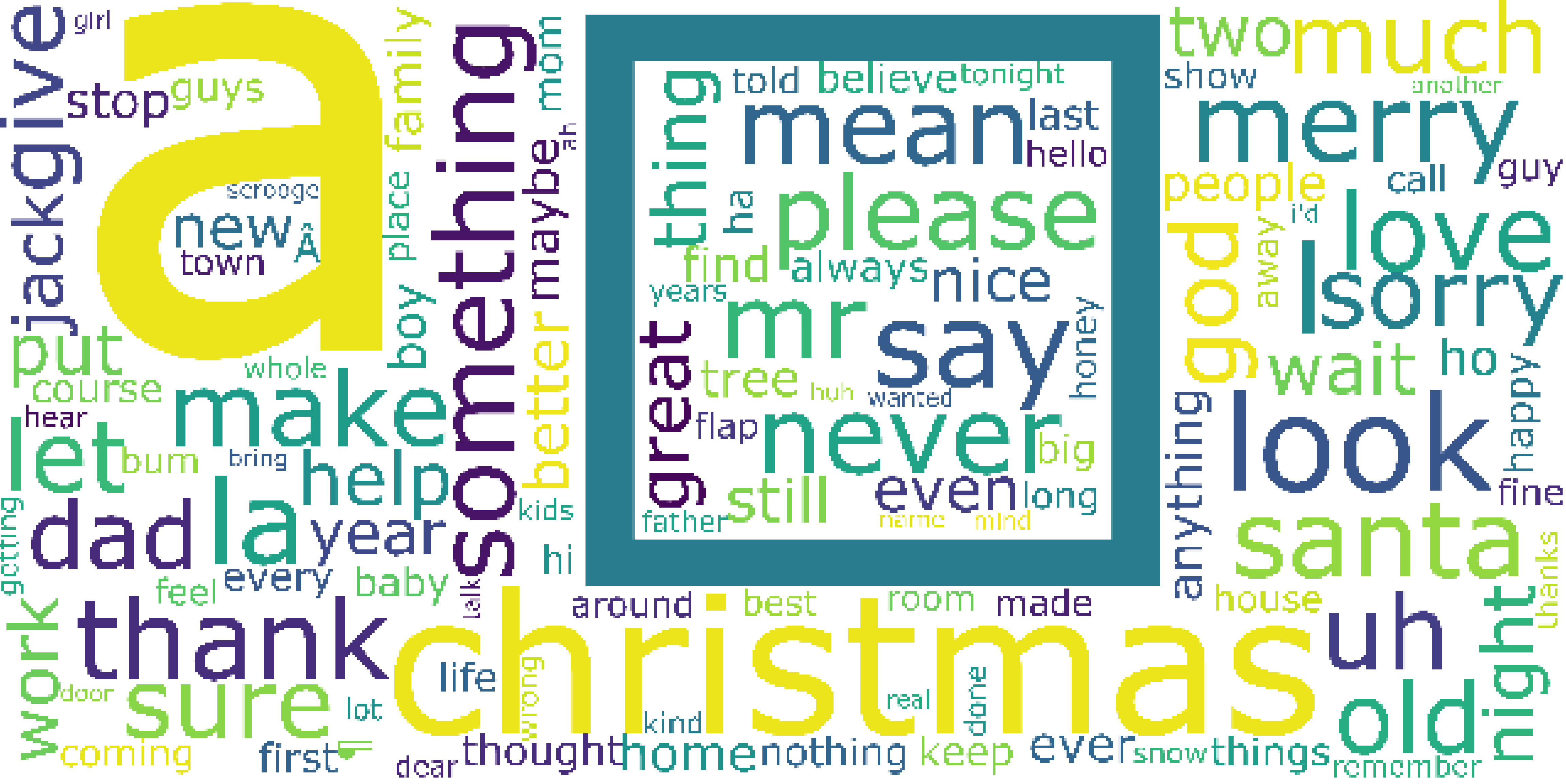
Let’s Analyze some movie scripts

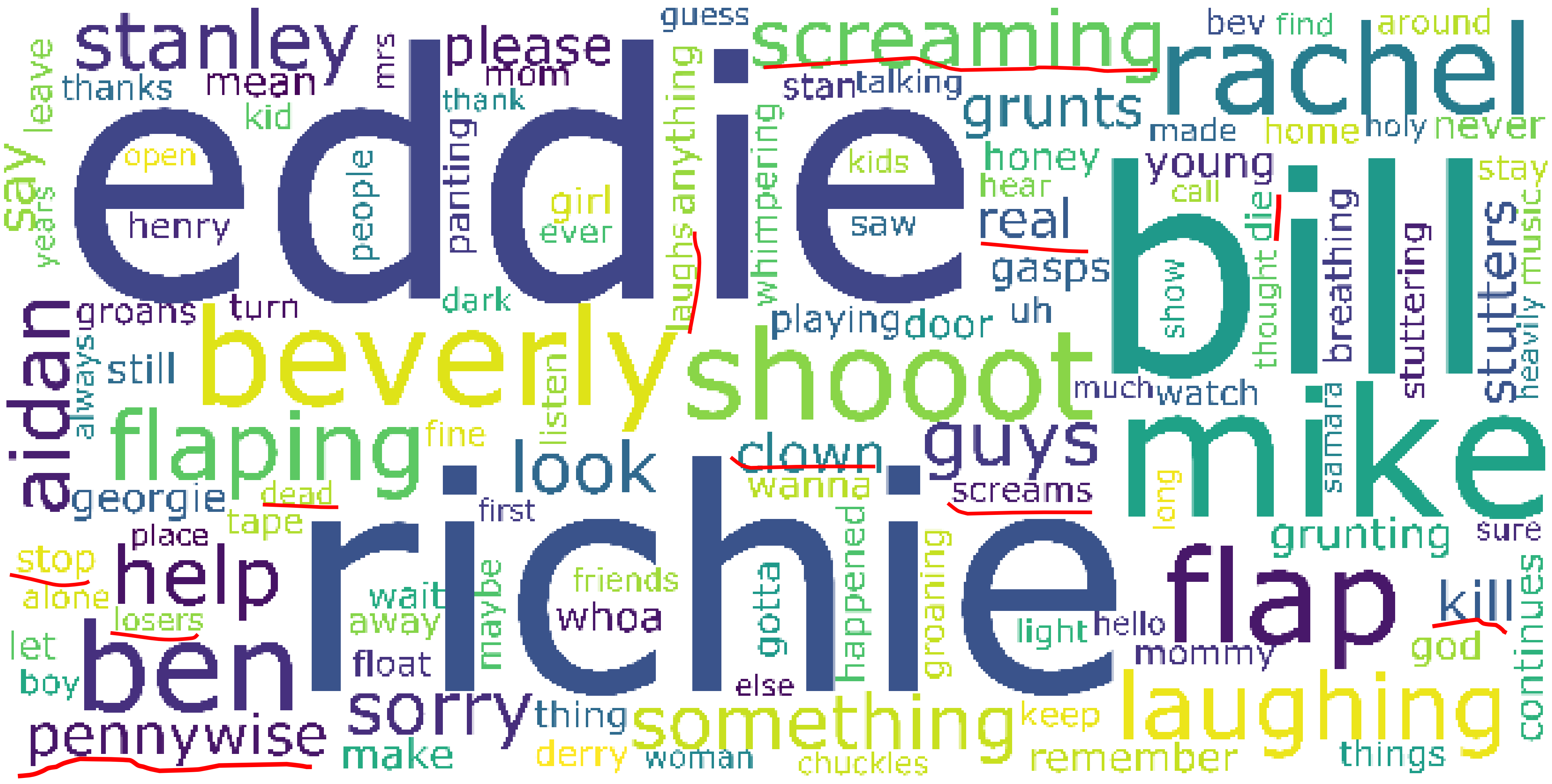

light
hope
gondor
world
dwarf
may
mordor
away
give
ring
gandalf
long
help
n
gate
two
city
stand
life
days
understand
much
son
stop
done
gone
name
mountain
enough
far
end
friend
father
move
last
black
kili
king
look
rohan
mean
made
follow
run
left
yet
mr
way
time
precious
safe
dwarves
shire
home
hobbit
dark
ever
lost
even
wait
frodo
bilbo
find
found
people
place
leave
always
took
power
return
upon
speak
old
elves
shall
coming
still
seen
nothing
lies
dragon
make
death
sorry
great
what's
orcs
say
none
keep
stay
fight
hobbits
smeagol
fire
kill
bard
never
thorin
aragorn
enemy
saruman
first
fall
deep
gollum
master
lord
sam
let
baggins
dead
hold
something
earth
war
many
sauron
merry


captain
spock
kirk
enterprise
mr
ship
time
admiral
starfleet
way
never
course
chekov
life
let
commander
shot
aboard
we've
find
always
done
give
data
aye
long
stop
god
new
earth
bridge
computer
genesis
uhura
starship
crew
order
first
lieutenant
sulu
space
kill
things
jim
command
movie
make
report
sorry
klingon
another
better
thing
three
wanted
understand
control
people
look
please
thought
may
human
believe
actually
love
last
many
ready
shields
planet
torpedoes
damn
beam
two
remember
scotty
help
idea
keep
we'll
hold
years
fire
thank
coming
much
full
vulcan
doctor
away
v'ger
scott
nothing
weapons
stand
every
put
feel
hell
anything
seconds
kind
still
ever
something
say
minutes
warp
power
federation
dr
mean
what's
even
moment
vessel
sequence


around
feel
anakin
make
new
ha
stay
shield
death
senator
obi
fire
first
life
give
force
droid
much
enough
thought
coming
jedi
help
look
r
father
believe
droids
power
big
things
yoda
control
move
still
attack
friend
general
planet
great
nothing
system
speaking
red
long
luke
always
boy
old
never
keep
please
lando
order
even
far
use
what's
away
hope
vader
we've
made
solo
side
gotta
stop
done
han
everything
sure
master
bring
skywalker
better
bad
republic
anything
people
po
thing
thank
captain
wan
sorry
told
let
leader
dark
uh
leave
ship
leia
place
watch
lost
try
ani
last
lord
chewie
kill
find
say
wait
might
may
something
two
ready
hold
chancellor
beep
work
kenobi
young
senate
council
course


help
let
jarvis
three
head
hulk
place
iron
father
made
put
old
wait
course
mean
dead
last
gotta
great
believe
captain
wanted
say
guys
tony
peter
lot
stay
long
dad
whoa
happened
i'd
hold
playing
grunts
bad
wanna
big
around
even
happy
saw
listen
kill
job
ever
leave
best
always
guy
maybe
thought
people
stop
thing
work
power
kid
coming
trying
thank
love
a
â
else
first
find
world
home
kind
thanks
give
parker
suit
name
every
years
feel
harry
hell
next
away
night
talk
hi
seen
look
anything
two
grunting
nothing
stark
fine
things
uh
wrong
done
understand
sorry
spider
never
please
god
mr
still
sure
call
keep
move
ready
remember
nice
may
life
better
told
actually


give
jack
p
girl
a
stop
guys
scrooge
new
Â
town
put
course
whole
make
let
hear
bum
bring
help
la
dad
year
family
place
boy
something
mom
maybe
ah
better
kids
hi
getting
every
feel
baby
talk
work
thank
sure
life
door
lot
wrong
coming
first
dear
thought
home
christmas
around
best
room
made
thing
told
believe
tonight
mean
last
hello
ha
please
find
always
nice
years
mr
say
honey
great
tree
huh
wanted
flap
never
big
still
even
long
father
name
mind
kind
real
done
nothing
keep
ever
snow
things
remember
two
much
show
another
merry
call
guy
people
i'd
god
away
love
l
sorry
wait
ho
happy
look
anything
fine
santa
house
uh
thanks
night
old


eddie
richie
bill
mike
beverly
shooot
rachel
screaming
flap
laughing
something
flaping
ben
stanley
aidan
guys
help
sorry
look
pennywise
please
mean
mom
thanks
leave
kid
say
years
open
henry
people
panting
girl
ever
anything
laughs
mrs
guess
stan
talking
grunts
whimpering
kids
saw
bev
find
around
made
home
holy
never
honey
hear
young
call
real
gasps
stay
music
heavily
stutters
stuttering
breathing
die
thought
show
uh
playing
door
dark
turn
groans
always
still
listen
fine
much
watch
samara
long
clown
wanna
screams
grunting
sure
georgie
dead
tape
first
place
stop
alone
losers
let
boy
wait
away
friends
whoa
maybe
float
gotta
happened
groaning
light
hello
mommy
else
kill
god
continues
thing
make
derry
woman
chuckles
keep
remember
things

More Examples

STARWARS MOVIES

- may the Force be with you.
- The Empire is still out there!
- Commander Skywalker
- Our plan, captain?
- Luke Skywalker, Jedi knight.

MARVEL MOVIES

- Can we hold them? - They're the **Avengers**!!
- I'm **Captain America**.
- who is also known as **Iron Man**
- the **strong**est substance in the **universe**..
- **Xandarian** outposts throughout the **galaxy**

LORD OF THE RINGS MOVIES

- What did you tell him about <u>Frodo</u> and the <u>Ring</u>?
- An <u>Elf</u> will go <u>underground</u> where a <u>Dwarf</u> dare not?
- But we swore to serve the <u>master</u> of the <u>precious</u>.
- They think we have the <u>Ring</u>.

HORROR MOVIES

- I'm not afraid of you!
- Run, run, run.
- Someone save me.
- Please, I need your help.
- You seem sad.

MORE ABOUT THE PROCESS

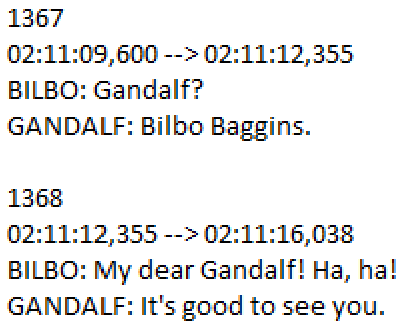
Analyzing subtitles

LEARNING

- Several English subtitles are downloaded from the internet (public use).
- Subtitles from several movie genres are gathered in several folders.
- Cleaning was done on 2 phases, phase 1 is removing punctuation, and unneeded characters. Phase 2 is removing stopwords.
- Flatten subtitle file and label it. Each subtitle on a line and add genre next to it.
- Process each genre separately and get an idea on most used words.

CLEANING DATA

- Example of a subtitle before cleaning:

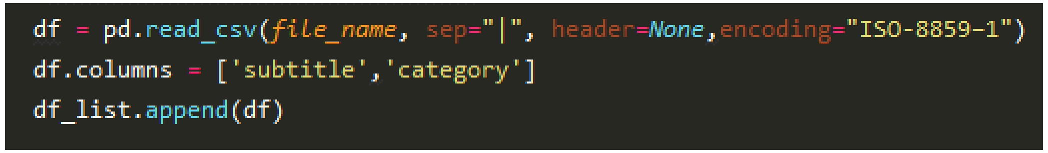


```
1367
02:11:09,600 --> 02:11:12,355
BILBO: Gandalf?
GANDALF: Bilbo Baggins.

1368
02:11:12,355 --> 02:11:16,038
BILBO: My dear Gandalf! Ha, ha!
GANDALF: It's good to see you.
```

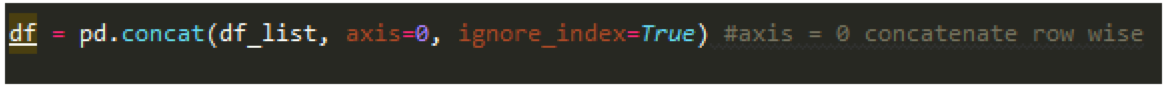
- Remove Numbers, columns and dots.
- Remove Stopwords. Stopwords are most commonly used words (such as “the”, “a”, “an”, “in”).
- Write a python program to clean data, remove new lines and label it.

PROCESSING DATA

- Code is written in python.
- NLTK library is used (Natural Language Toolkit)

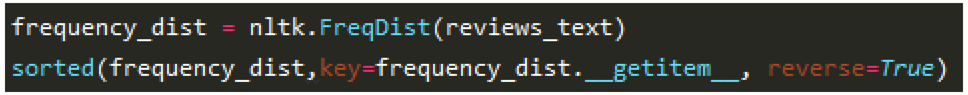


```
df = pd.read_csv(file_name, sep="|", header=None,encoding="ISO-8859-1")
df.columns = ['subtitle','category']
df_list.append(df)
```



```
df = pd.concat(df_list, axis=0, ignore_index=True) #axis = 0 concatenate row wise
```

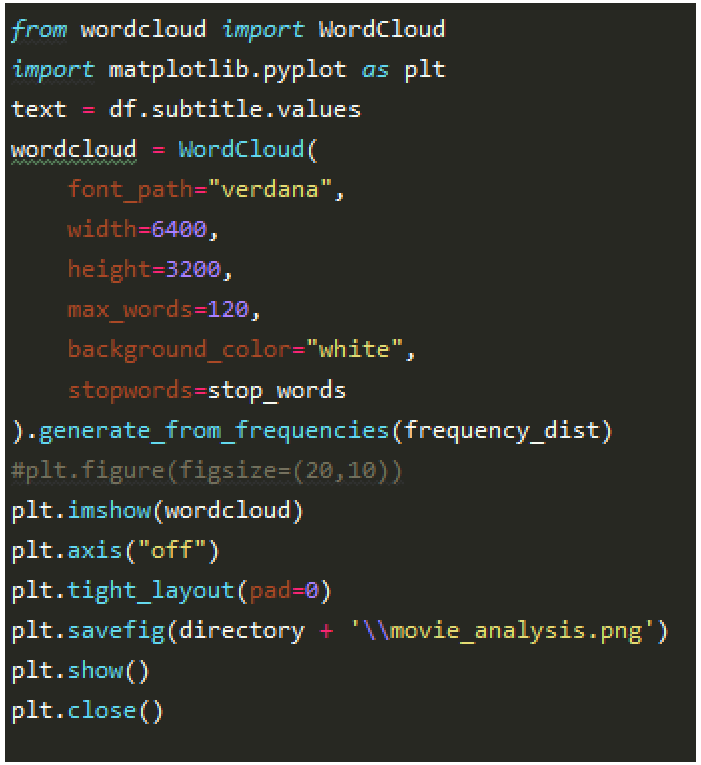
- TweetTokenizer is used instead of word_tokenize to split data.
- FreqDist is used to get the frequency of the words.



```
frequency_dist = nltk.FreqDist(reviews_text)
sorted(frequency_dist,key=frequency_dist.__getitem__, reverse=True)
```

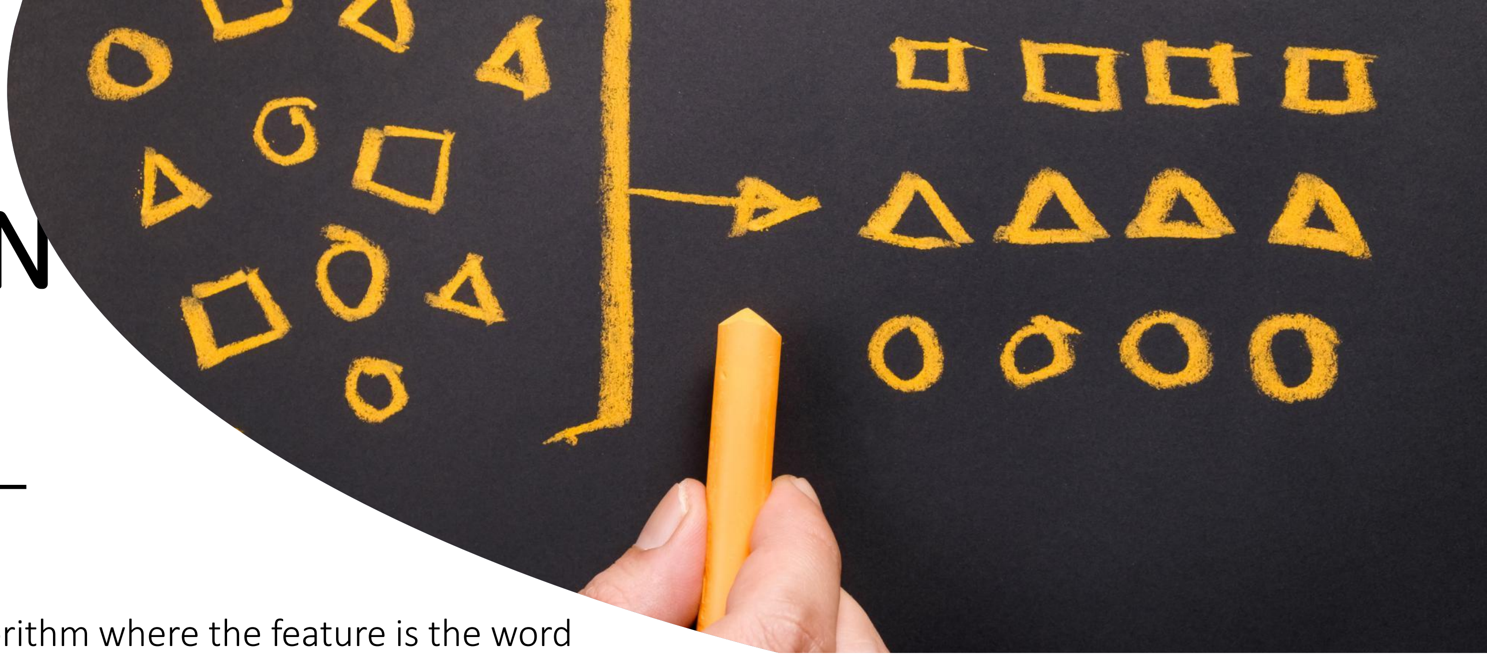
PROCESSING DATA

- wordcloud

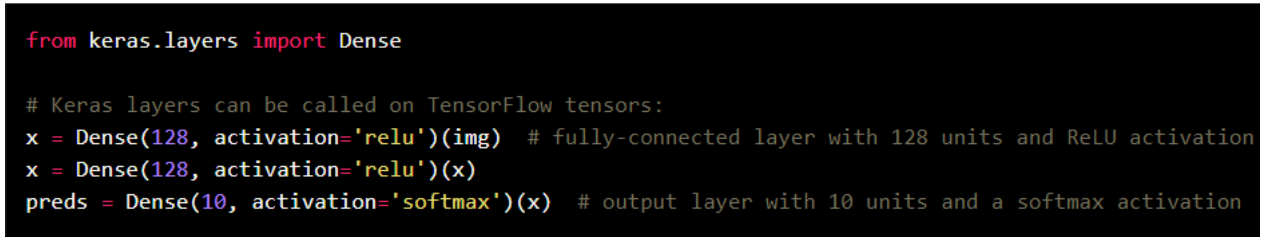


```
from wordcloud import WordCloud
import matplotlib.pyplot as plt
text = df.subtitle.values
wordcloud = WordCloud(
    font_path="verdana",
    width=6400,
    height=3200,
    max_words=120,
    background_color="white",
    stopwords=stop_words
).generate_from_frequencies(frequency_dist)
#plt.figure(figsize=(20,10))
plt.imshow(wordcloud)
plt.axis("off")
plt.tight_layout(pad=0)
plt.savefig(directory + '\\movie_analysis.png')
plt.show()
plt.close()
```

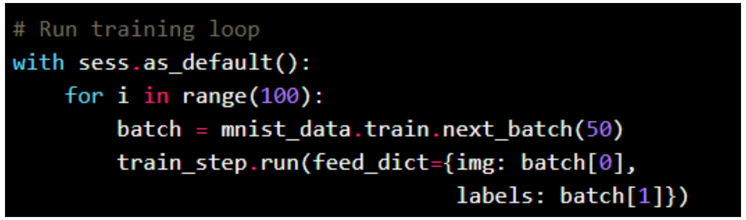
DATA CLASSIFICATION



- 2 Approaches to classify data:
  - Consider all columns as features and apply Knn algorithm where the feature is the word and the value is number of occurrences.
  - Using Deep learning to classify the movies. Tensor flow keras library



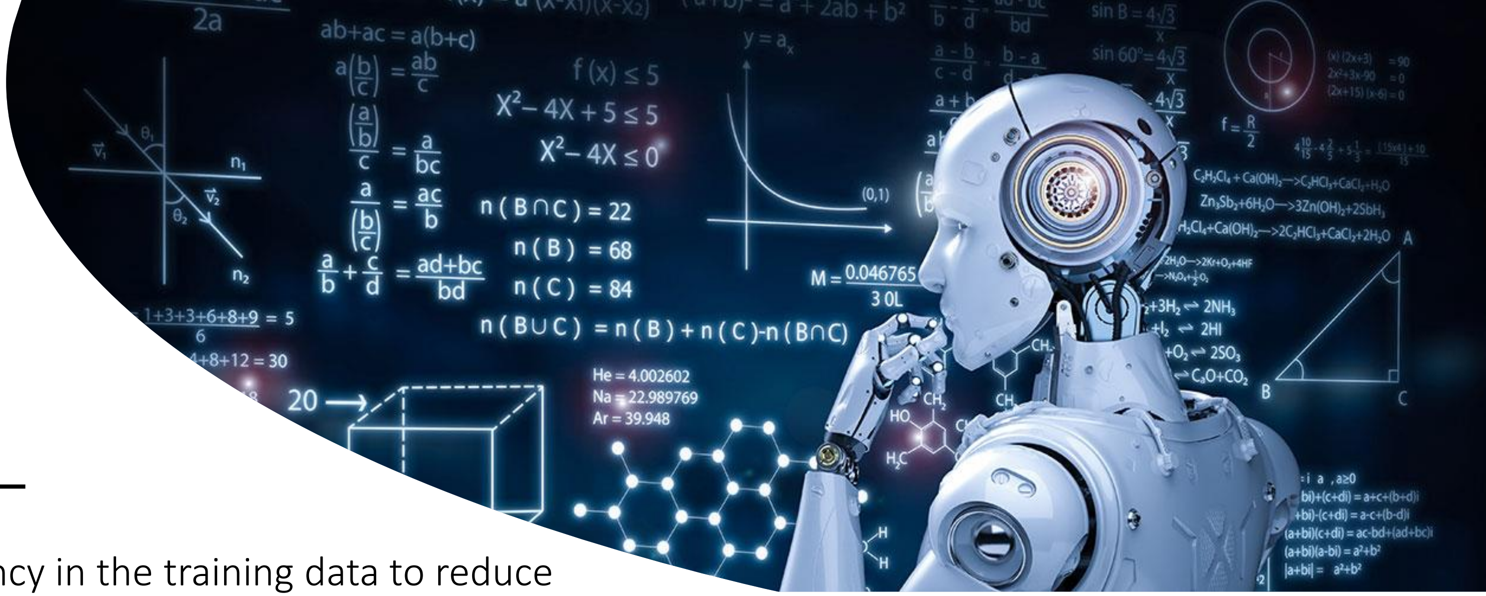
```
from keras.layers import Dense

# Keras layers can be called on TensorFlow tensors:
x = Dense(128, activation='relu')(img)  # fully-connected layer with 128 units and ReLU activation
x = Dense(128, activation='relu')(x)
preds = Dense(10, activation='softmax')(x)  # output layer with 10 units and a softmax activation
```



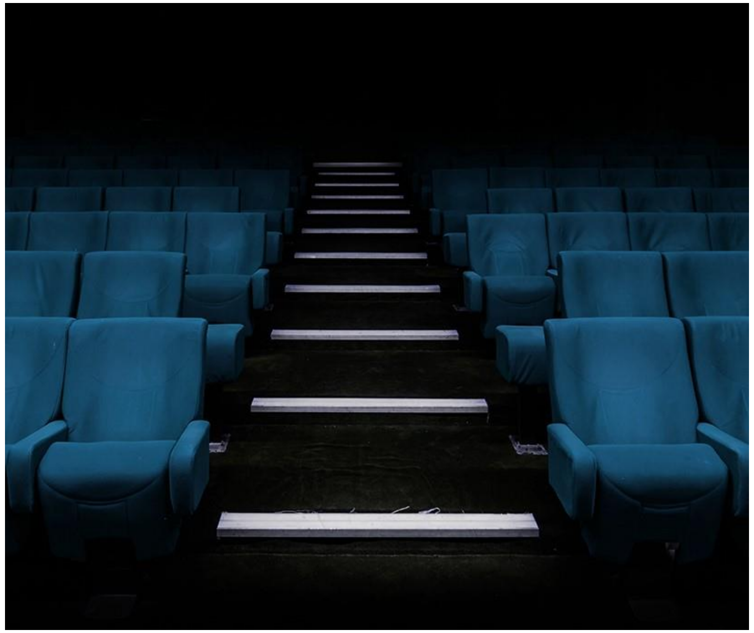
```
# Run training loop
with sess.as_default():
    for i in range(100):
        batch = mnist_data.train.next_batch(50)
        train_step.run(feed_dict={img: batch[0],
                                  labels: batch[1]})
```

- https://blog.keras.io/keras-as-a-simplified-interface-to-tensorflow-tutorial.html

IMPROVEMENTS



- Remove words that appear at the same frequency in the training data to reduce features.
- Apply Dimensionality reduction on the dataset, for example:
  - “Happy”, “Joy”
  - “Run”, “Sprint”, “Rush”, “Running”, “rushing”, “rushed”
  - “afraid”, “fear”, “scared”
- Ignore words that appear frequently in some movies, for example:
  The name “Rachel” in a horror movie has no importance. However, the name “Stark” as tony stark is important toward classifying a movie as a Marvel.
- Some action verbs are important not to ignore, like “run”, “go”, “move”. If they appear frequently, this means that it’s an action movie.
- Capture special characters ♪
- Maybe, Convert features to numbers.



END OR QUESTIONS