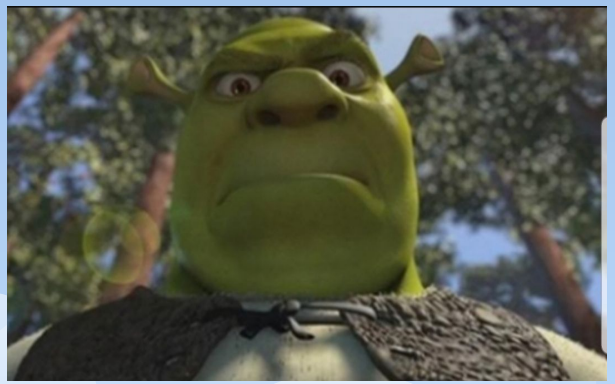
Streamlining the Screenplay Buying Process

By Michael Orlando

Our Problem

Hollywood receives about 50,000 thousand screenplays per year. But continues to make shitty movies. How can we improve this?



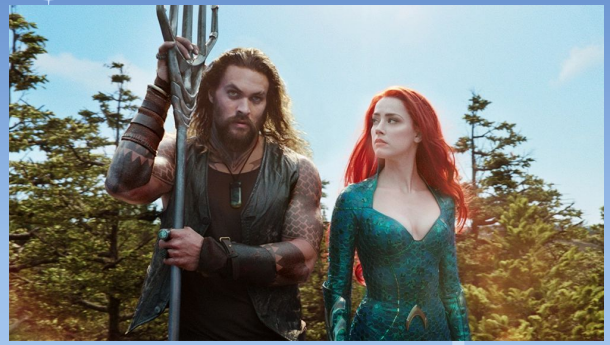
Business Understanding

One important element to movies is **GENRE.**

Genre brings in the target audience.

For example, if Disney was marketing a new Animated Fantasy movie, then the audience would be pretty upset if the film played like a Crime Horror movie.

Business Objective



To build a multi-label classifier, using Natural Language Processing, to classify the genres of a given screenplay.

The Machine -- Simplified

Screenplay

INT. HOUSE - NIGHT

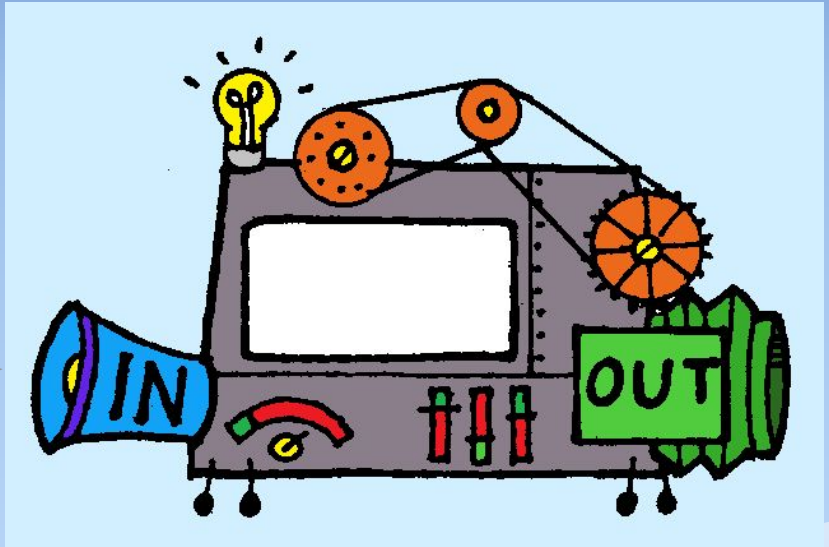
The Phone rings. Jason answers it.

JASON
(into phone)
Hello?

SALLY (V.O.)
Jason, I'm in trouble. I ran out of gas. Can you bring some?

JASON
(listening)
Yes, I can be there soon.

Jason hangs up and runs out the door.



Genre Classification

['Action', 'Comedy']

Data Collection

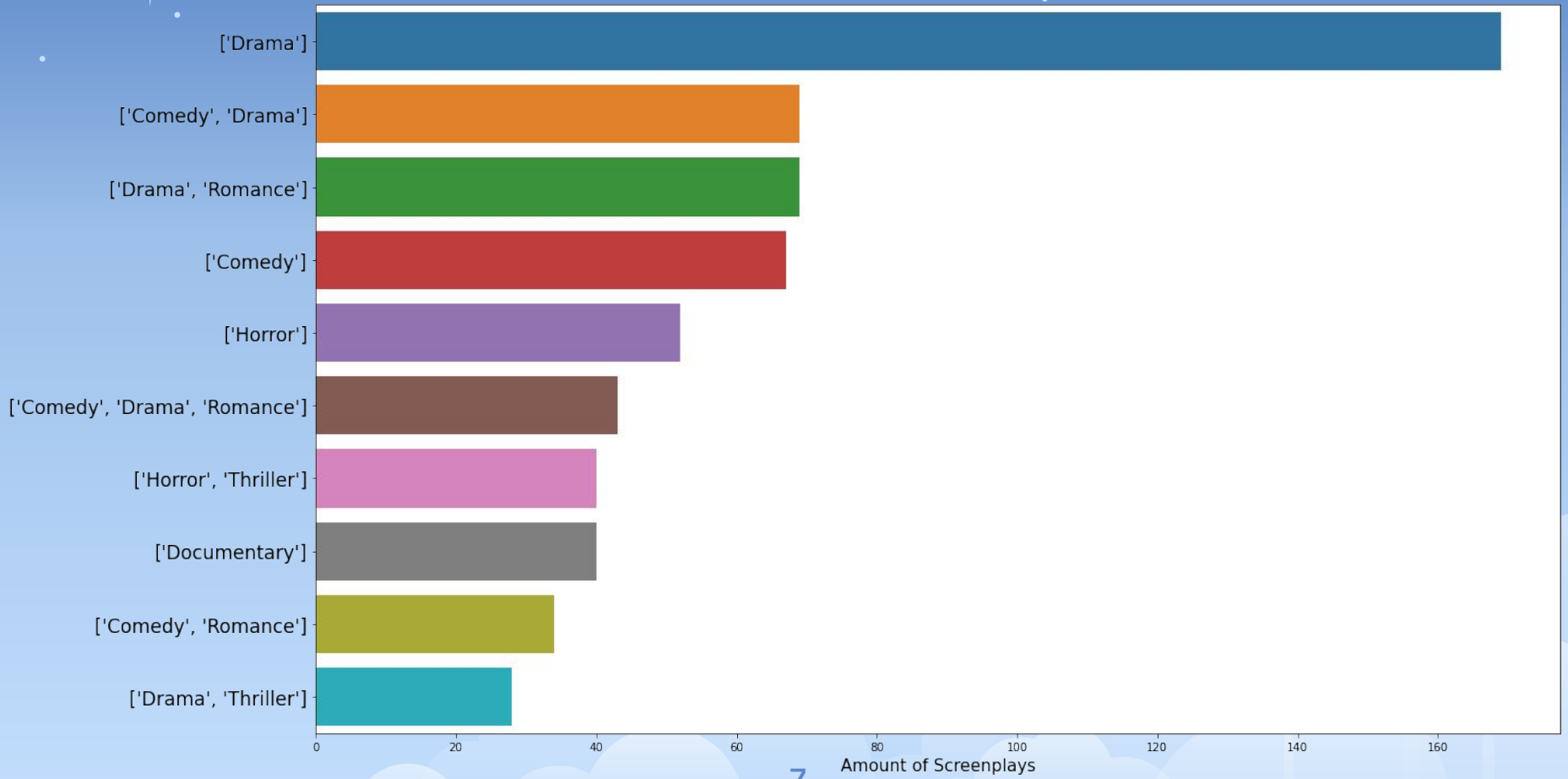
Screenplays:

The scripts in pdf format were scrapped from The Script Savant using Beautifulsoup

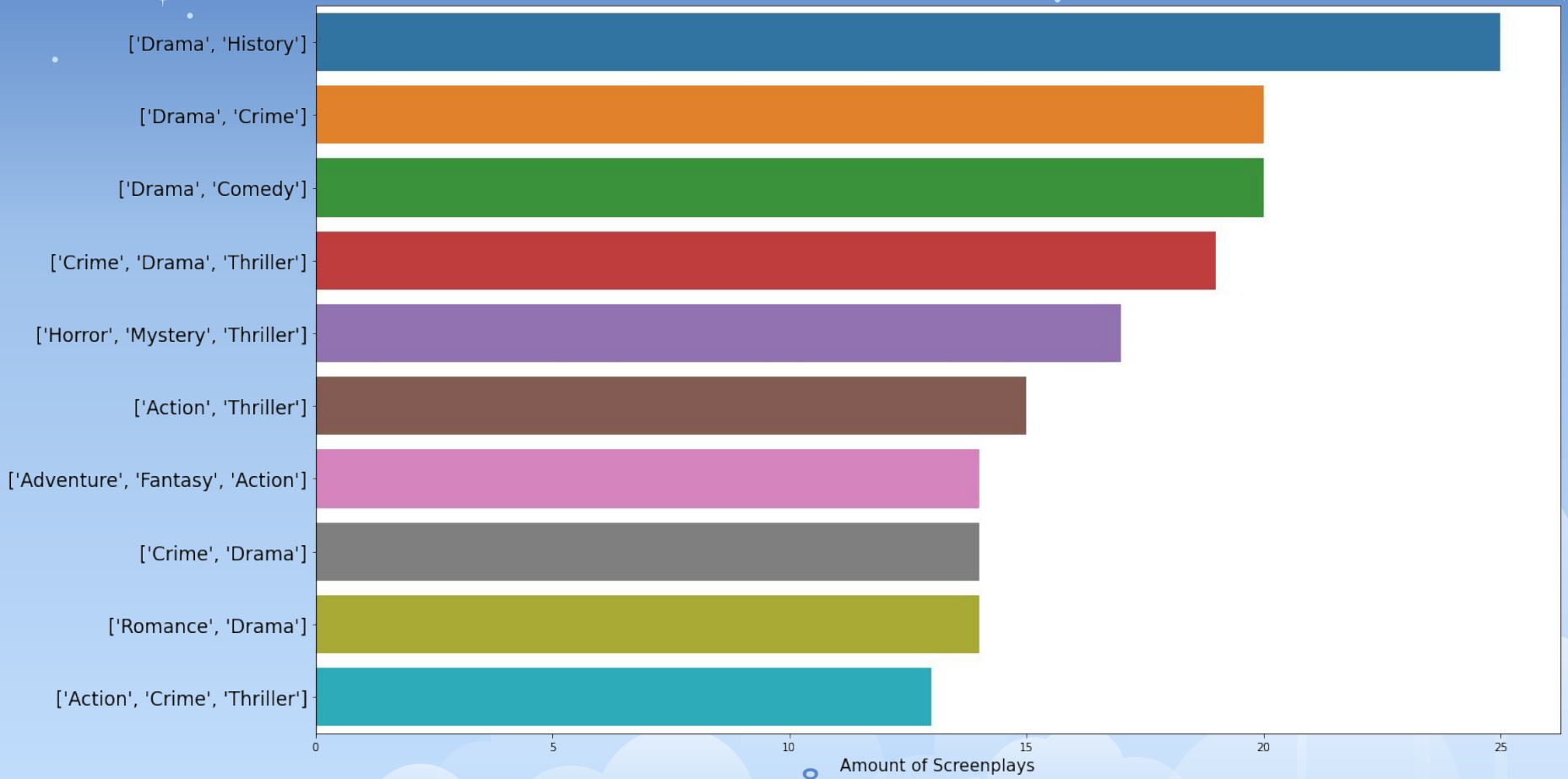
Genres:

The screenplays were labelled with genres using the Movie Database API

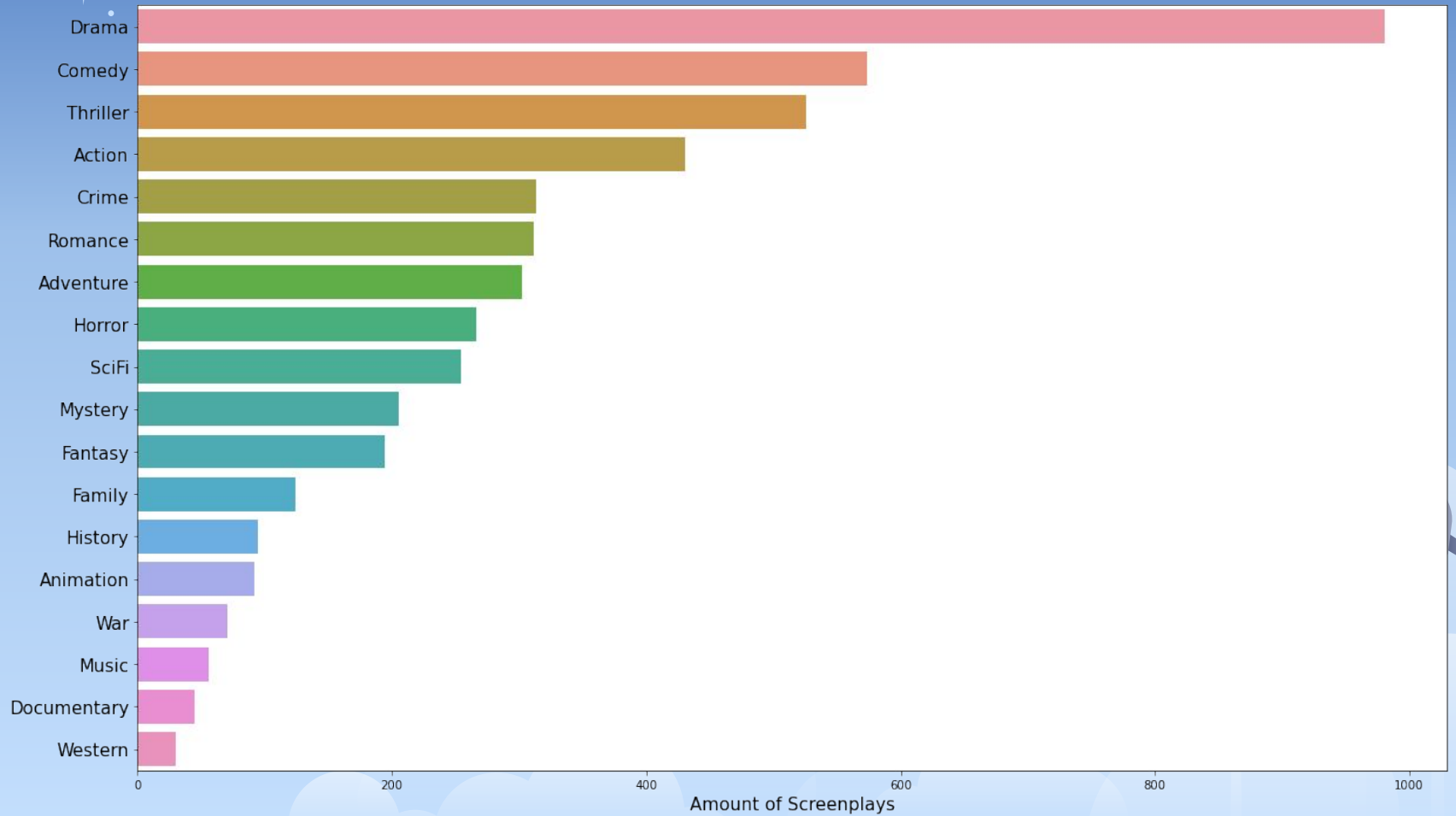
Genre Combination (1-10)



Genre Combination (11-20)



Target Distribution by Single Genre



Overall

- 2000 unique screenplays
- 18 different genres
- 600 unique combinations of genres
- Drama the most frequent in combinations
- Western the least frequent in combinations

Feature Engineering

The feature created is matrix of words (count or frequency)

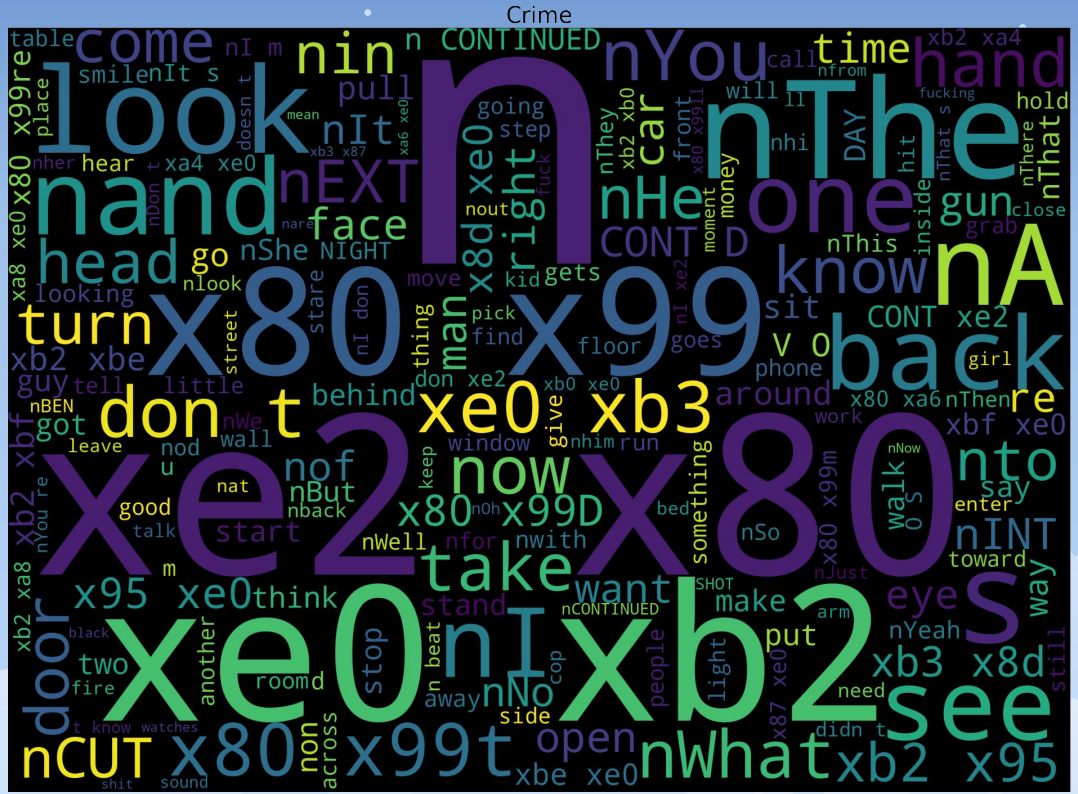
I improved the matrix of words by testing different preprocessing techniques & eliminating common words.

Results were illustrated using word clouds

For example....

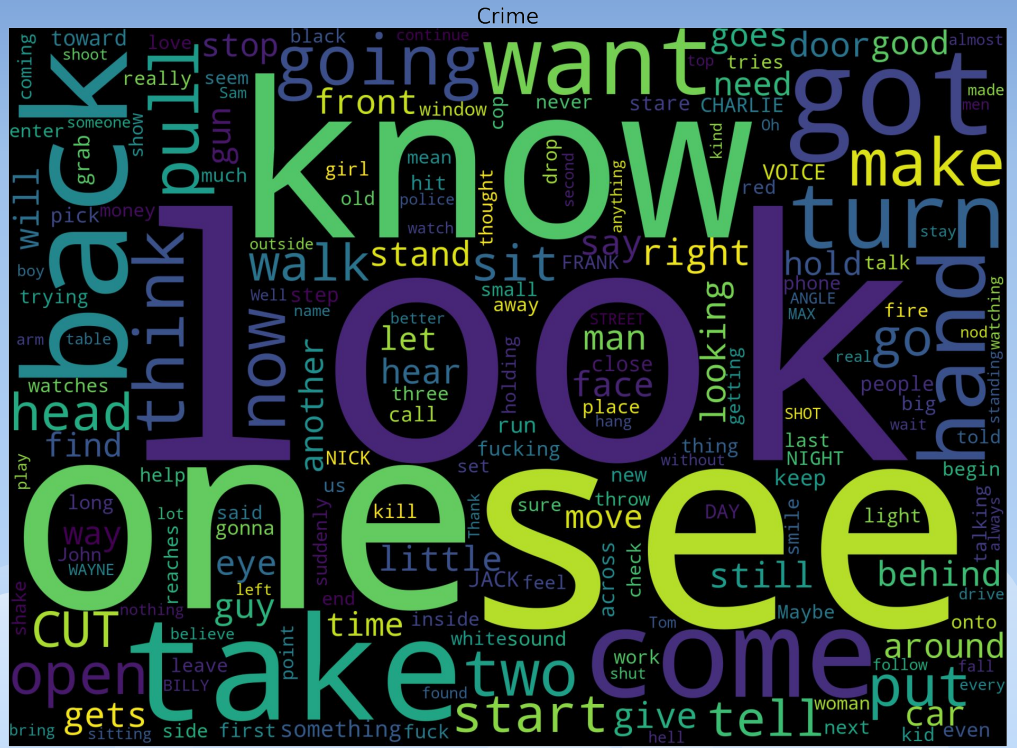
	Jumps	The	brown	dog	fox	lazy	over	quick	the
Doc1	0	1	1	0	1	0	0	1	0
Doc2	1	0	0	1	0	1	1	0	1

Before Cleaning
Text: Crime
Genre

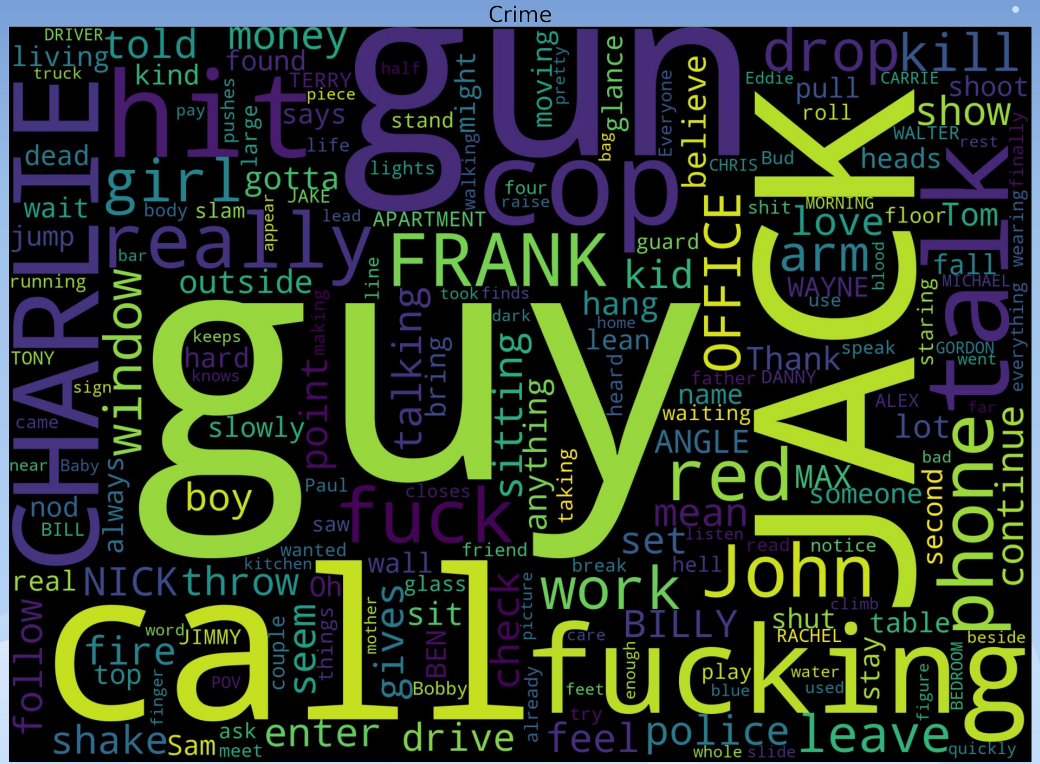


After Cleaning-- Crime Genre

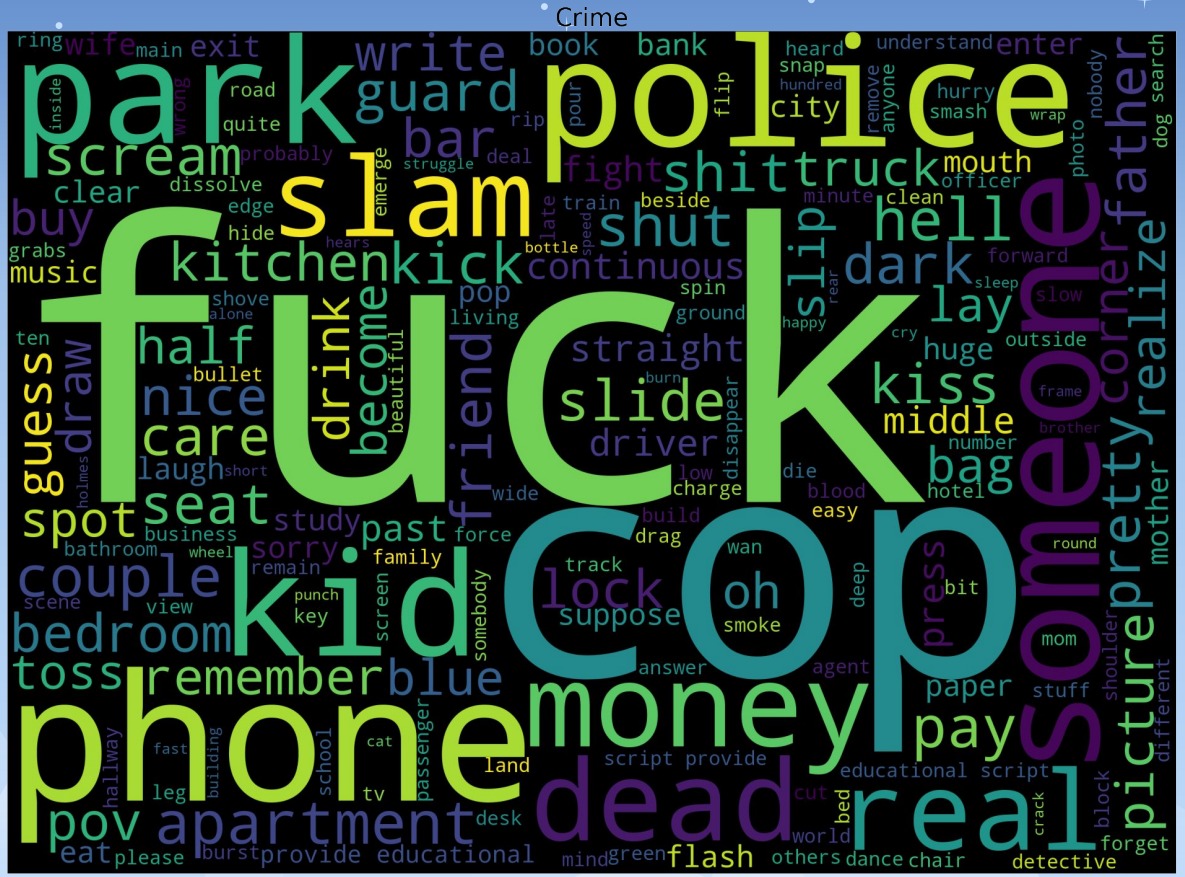
Much better, but still basic



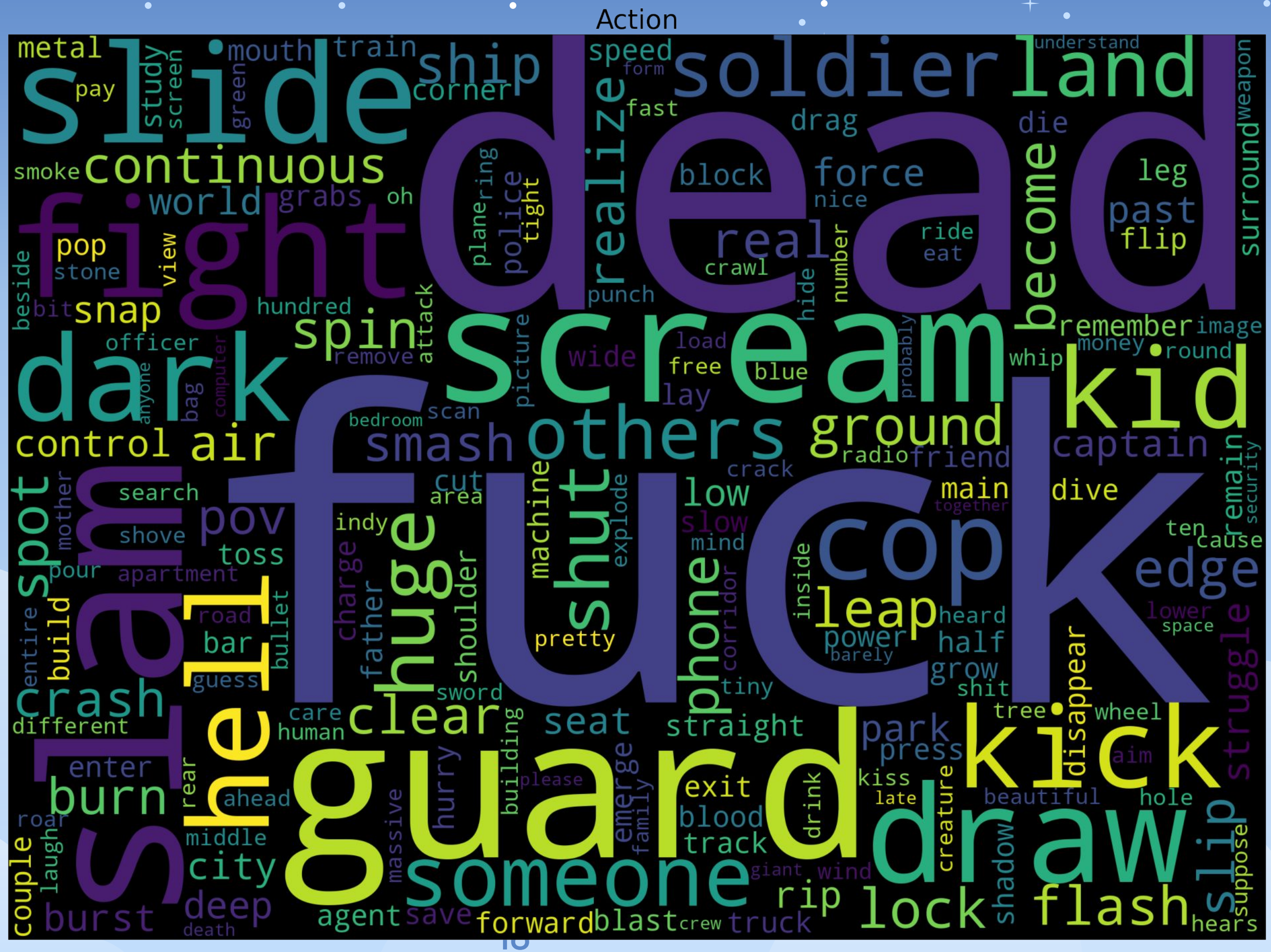
After Cleaning, Removal of Stop Words-- Crime Genre



After cleaning, lemmatizing, and removal of stop words: Crime Genre

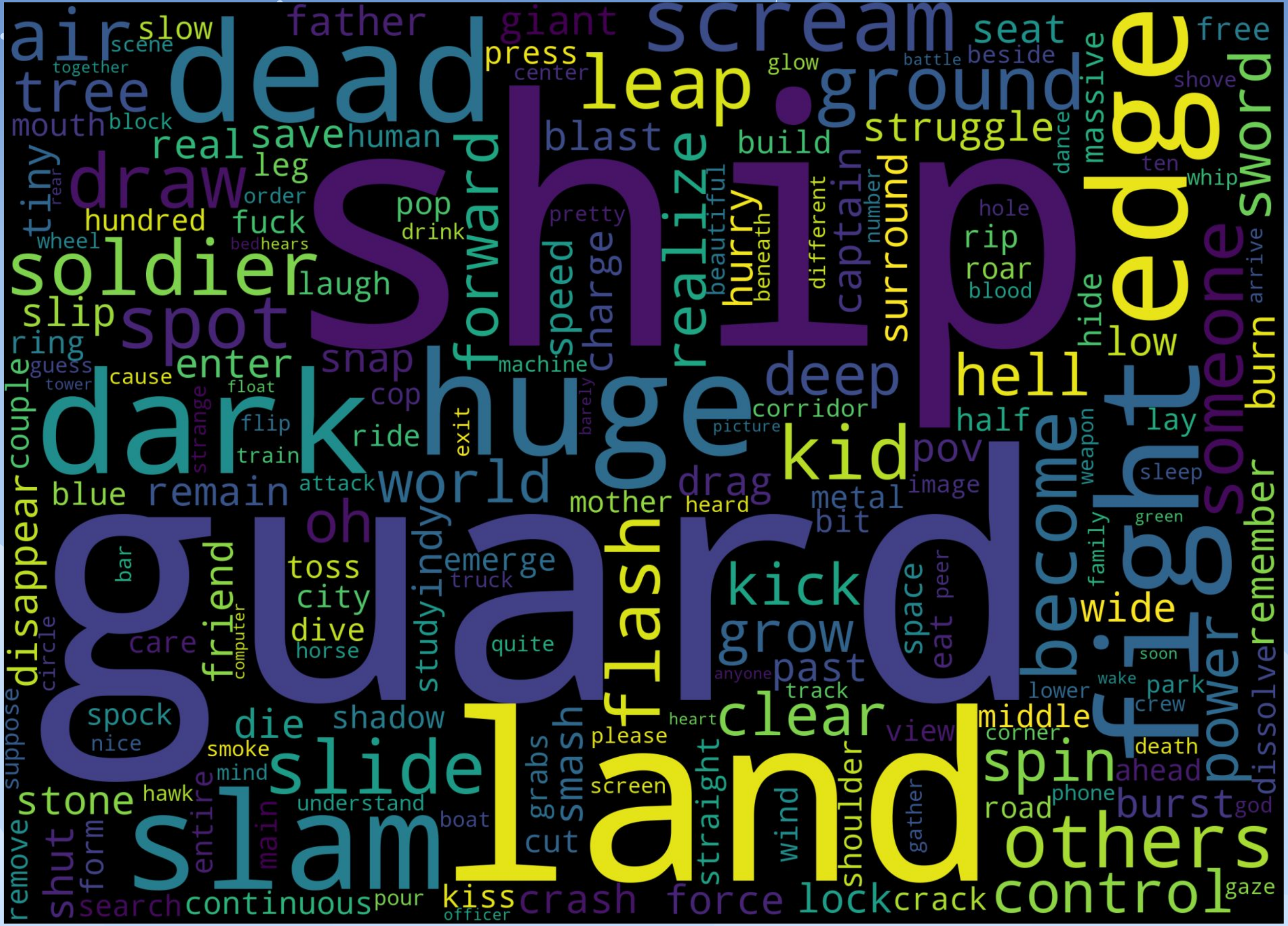


Action

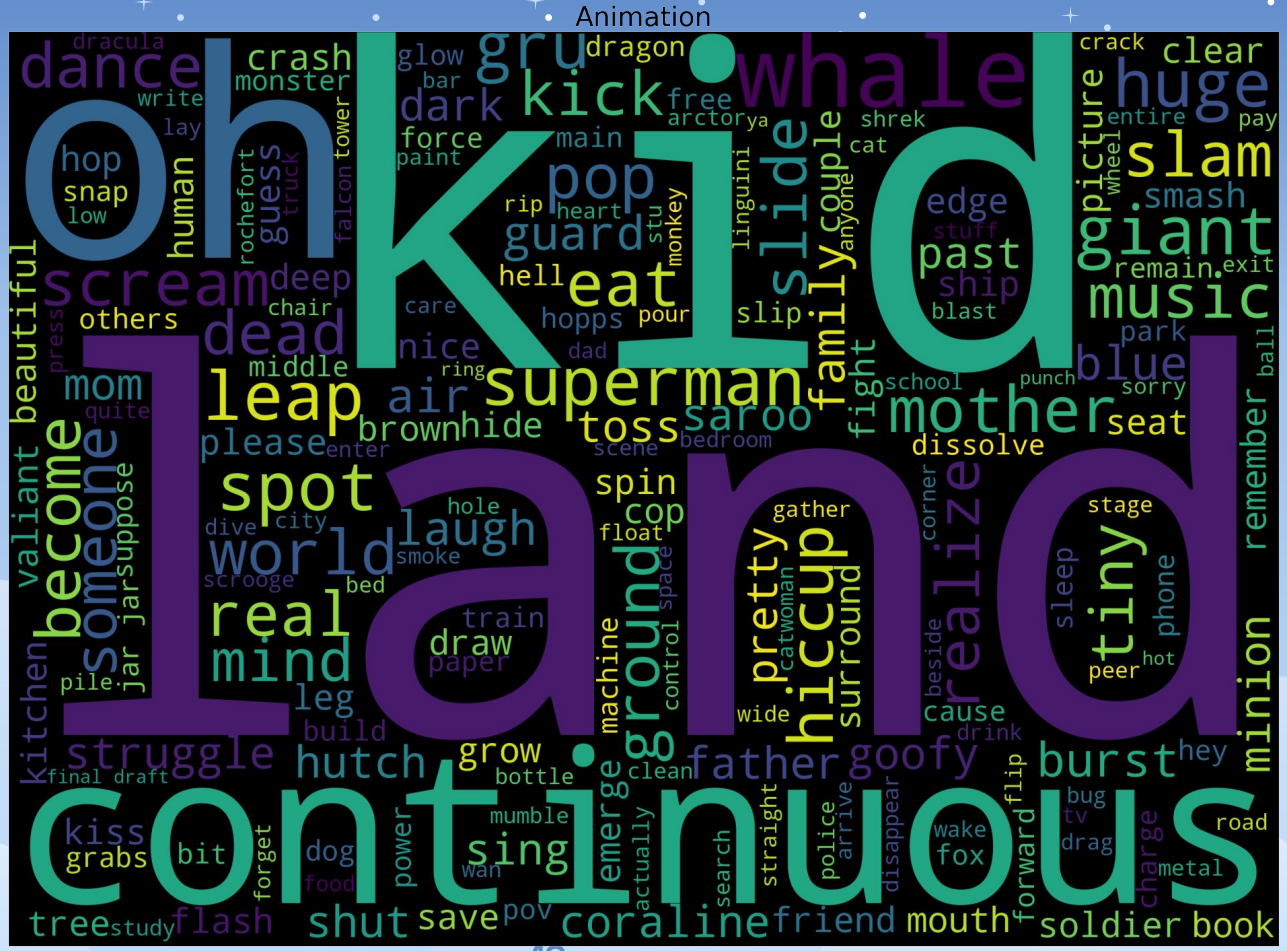

Action
dead
fuck
scream
guard
slide
fight
slam
dark
others
kick
draw
kid
soldier
land
hell
someone
huge
cop
continuous
ground
captain
shut
phone
realize
become
clear
leap
lock
flash
slip
spot
crash
burn
metal
pay
study
screen
green
mouth
train
ship
corner
speed
form
understand
weapon
surround
fast
drag
die
smoke
world
grabs
oh
ring
plane
police
tight
block
force
nice
leg
past
flip
pop
stone
view
real
crawl
ride
eat
number
hide
beside
bit
snap
hundred
attack
punch
spin
remember
image
officer
remove
picture
load
wide
free
blue
probably
whip
money
round
computer
anyone
bag
scan
lay
bedroom
smash
control
air
radio
friend
crack
remain
security
search
mother
cut
area
machine
low
main
together
dive
pov
shove
indy
explode
slow
mind
ten
cause
toss
pour
apartment
charge
edge
corridor
inside
heard
lower
space
road
father
shoulder
pretty
power
half
bar
bullet
barely
entire
build
guess
grow
struggle
disappear
sword
tiny
shit
different
care
human
building
seat
straight
tree
wheel
enter
park
press
aim
rear
hurry
please
exit
kiss
ahead
emerge
family
drink
late
beautiful
hole
creature
roar
middle
laugh
blood
massive
track
giant
wind
city
shadow
suppose
couple
rip
deep
burst
agent
save
forward
blast
crew
truck
death
hears

Adventure

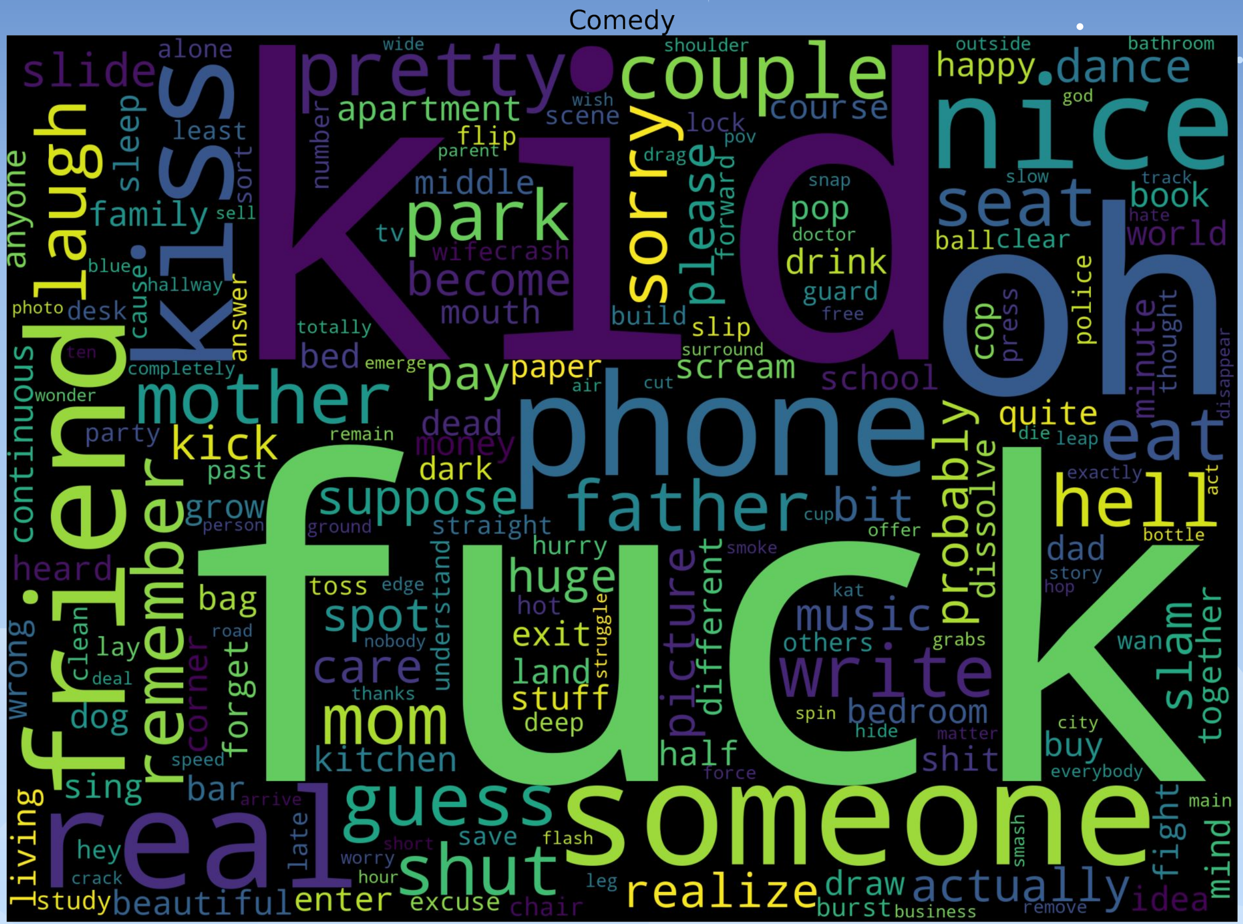
Adventure



Animation

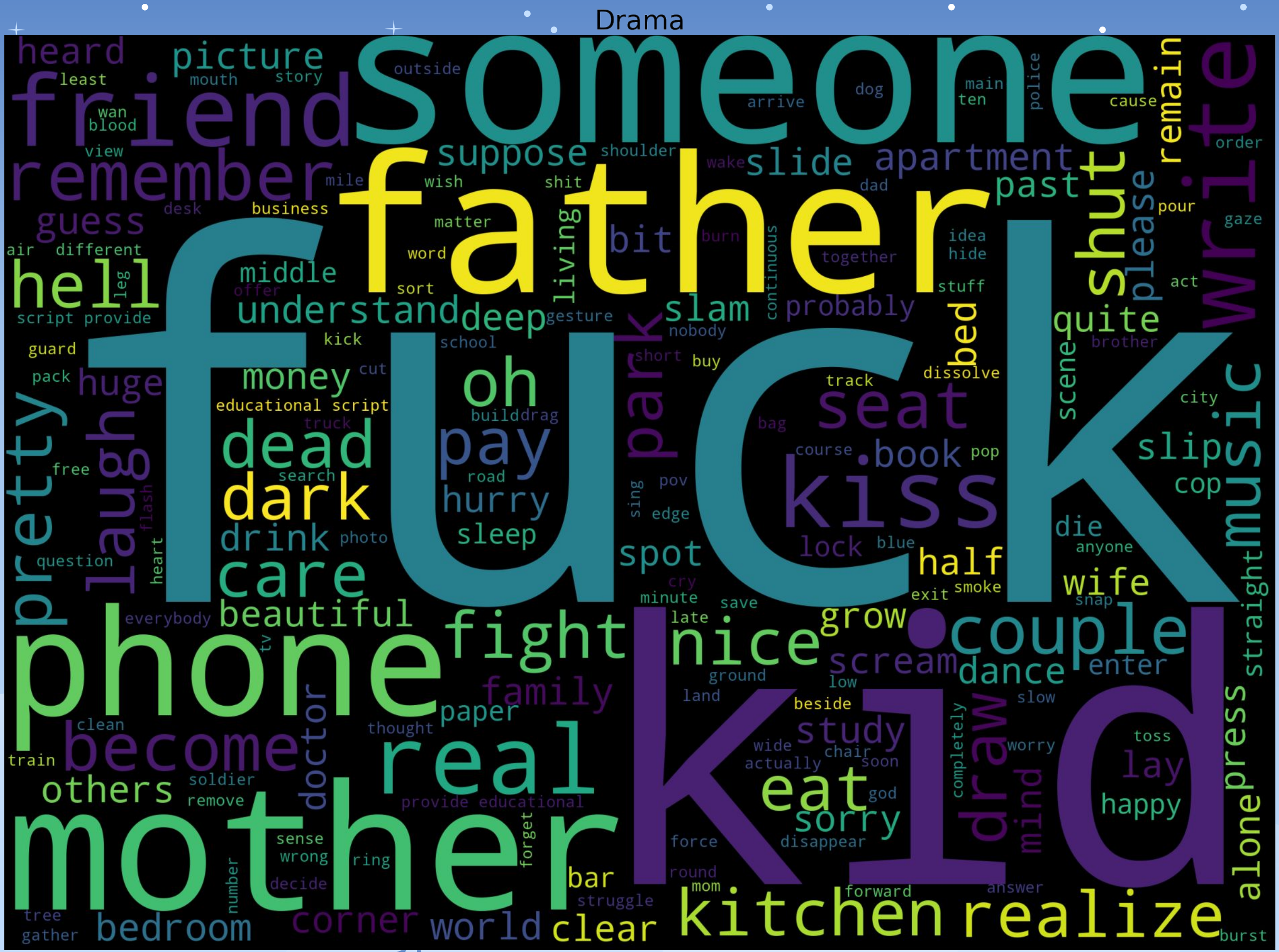

Animation
oh
kid
Land
continuous
dance
crash
monster
dragon
gru
whale
huge
clear
kick
dark
free
slide
slam
pop
guard
eat
past
giant
music
scream
dead
leap
superman
mother
blue
become
someone
spot
world
real
mind
laugh
hiccup
pretty
surround
ground
realize
tiny
minion
struggle
hutch
grow
father
goofy
burst
kiss
sing
emerge
coraline
shut
save
friend
mouth
soldier
book
tree
flash
human
guess
hop
snap
smash
picture
family
couple
edge
ship
blast
remain
nice
toss
saroo
brownhide
please
dissolve
seat
spin
cop
float
gather
stage
draw
train
paper
machine
control
space
cause
wide
drink
sleep
phone
peer
hot
beside
corner
dive
city
hole
smoke
scrooge
bed
leg
build
pile
kitchen
jar
suppose
valiant
beautiful
remember
ball
park
school
punch
sorry
fight
ring
middle
mom
quite
others
chair
deep
care
hopps
pour
slip
dad
rip
heart
stu
monkey
linguini
anyone
shrek
cat
entire
pay
wheel
exit
stuff
rochefort
truck
falcon
tower
lay
write
bar
glow
paint
force
main
arctorya
crack
dracula
bottle
clean
mumble
wan
actually
search
straight
police
arrive
disappear
forward
flip
bug
tv
drag
charge
metal
road
wake
fox
forget
grabs
bit
dog
food
power
pov
study
final draft
hey

Comedy

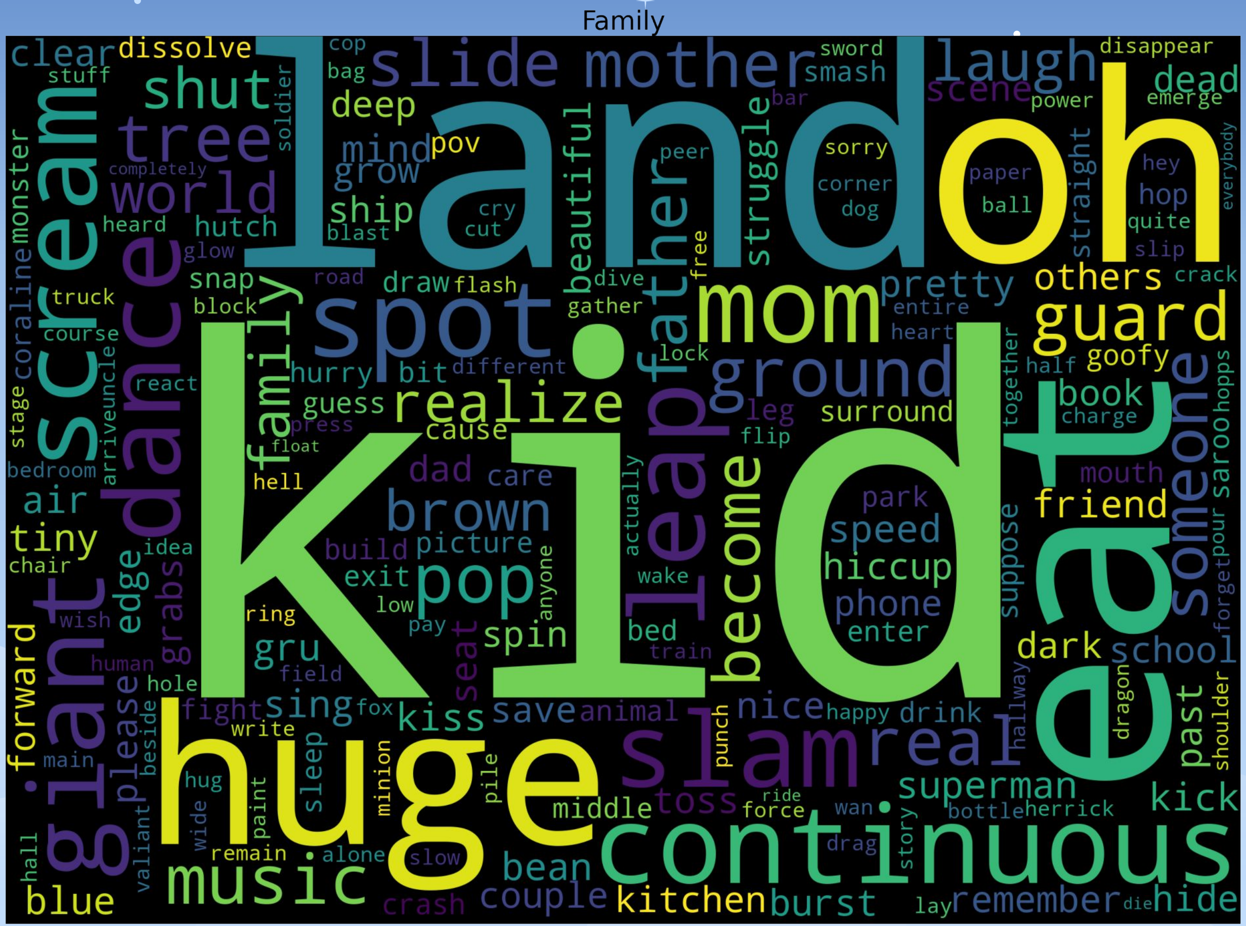

Comedy
kid
fuck
phone
someone
oh
nice
real
friend
kiss
pretty
couple
alone
wide
shoulder
outside
bathroom
slide
happy
dance
apartment
scene
wish
god
course
least
flip
lock
pov
parent
drag
sleep
laugh
sort
number
sorry
middle
snap
slow
track
book
family
sell
park
pop
seat
hate
tv
doctor
ball
clear
world
wifecrash
please
forward
drink
blue
hallway
become
guard
anyone
cause
photo
desk
mouth
build
free
slip
police
totally
answer
surround
cop
press
minute
thought
ten
completely
bed
emerge
pay
paper
scream
school
disappear
wonder
mother
air
cut
quite
continuous
die
leap
eat
party
dead
remain
kick
money
probably
exactly
past
dark
remember
hell
act
suppose
father
grow
bit
cup
person
ground
straight
offer
bottle
dissolve
hurry
smoke
dad
heard
huge
story
toss
edge
kat
hop
bag
spot
hot
music
understand
picture
different
wrong
road
exit
others
grabs
wan
slam
together
lay
clean
nobody
struggle
write
care
land
deal
forget
corner
thanks
stuff
dog
mom
spin
bedroom
city
deep
hide
matter
speed
kitchen
half
shit
buy
sing
bar
force
everybody
arrive
main
guess
fight
living
hey
late
smash
mind
short
save
flash
worry
shut
crack
hour
leg
draw
actually
study
beautiful
enter
excuse
chair
realize
burst
business
remove
idea

Documentary

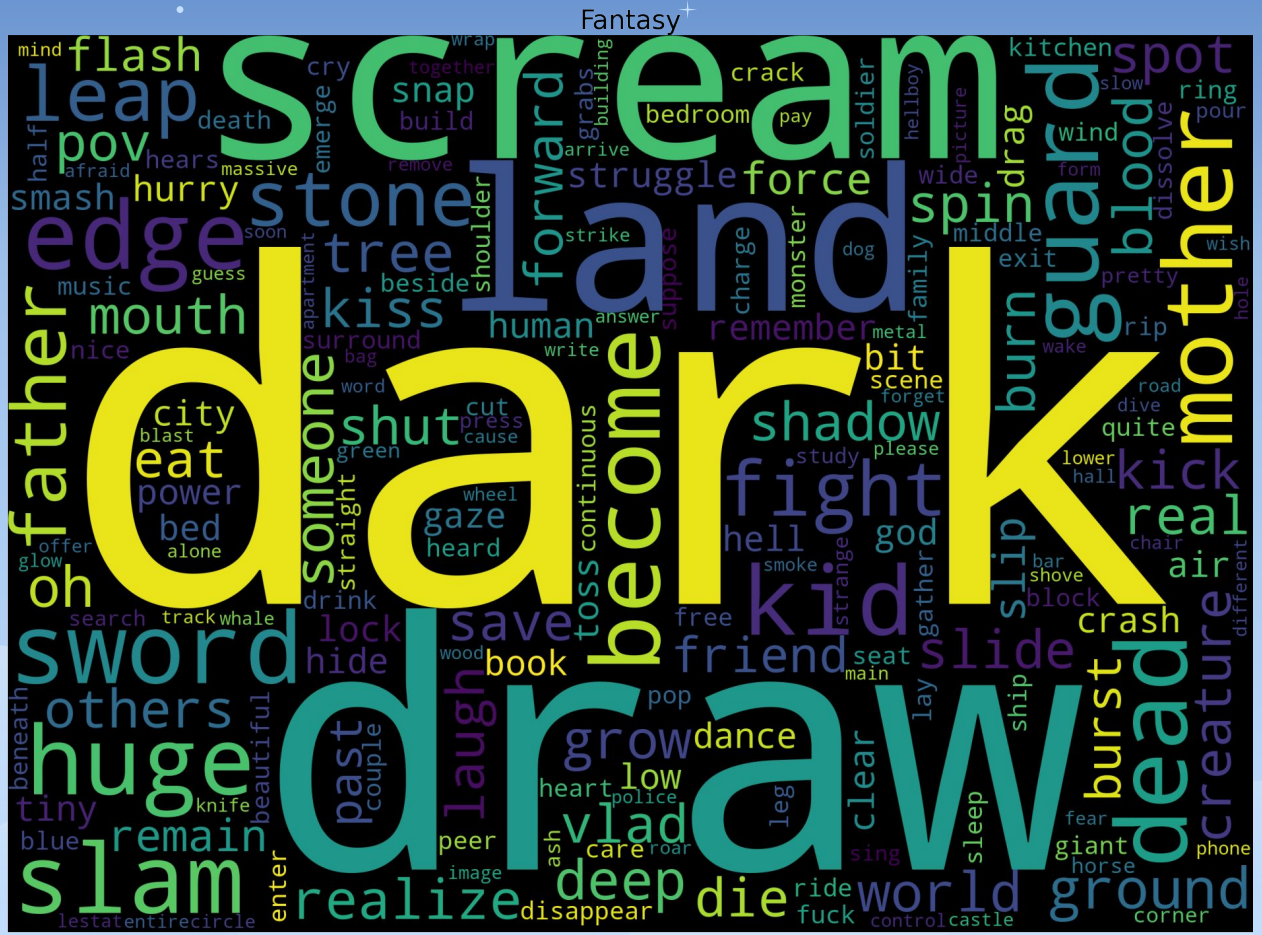
Drama


Drama
fuck
kid
someone
father
mother
phone
friend
remember
write
real
kiss
kitchen
realize
couple
nice
fight
become
dead
dark
pay
oh
seat
music
shut
please
hell
laugh
pretty
care
drink
money
understand
deep
slam
probably
book
park
eat
draw
study
sorry
scream
dance
enter
slip
cop
wife
die
half
quite
past
apartment
slide
suppose
picture
heard
guess
middle
hurry
sleep
spot
lock
grow
family
paper
others
bedroom
world
clear
corner
doctor
beautiful
bit
huge
bar
mind
lay
happy
alone
press
straight
remain
bed
scene
living
heard
least
mouth
story
outside
arrive
dog
main
ten
police
cause
order
wan
blood
view
shoulder
wake
mile
wish
shit
dad
desk
business
matter
burn
together
pour
gaze
air
different
word
idea
hide
leg
offer
sort
continuous
stuff
act
script
provide
gesture
kick
school
nobody
brother
guard
short
buy
track
dissolve
pack
cut
city
educational script
truck
builddrag
bag
course
pop
free
search
road
pov
sing
edge
flash
photo
blue
anyone
question
heart
cry
minute
save
exit
smoke
snap
everybody
tv
late
ground
low
slow
land
beside
clean
thought
wide
chair
worry
toss
train
actually
soon
completely
soldier
remove
provide educational
god
sense
wrong
ring
forget
force
disappear
decide
round
mom
forward
answer
number
struggle
tree
gather
burst

Family

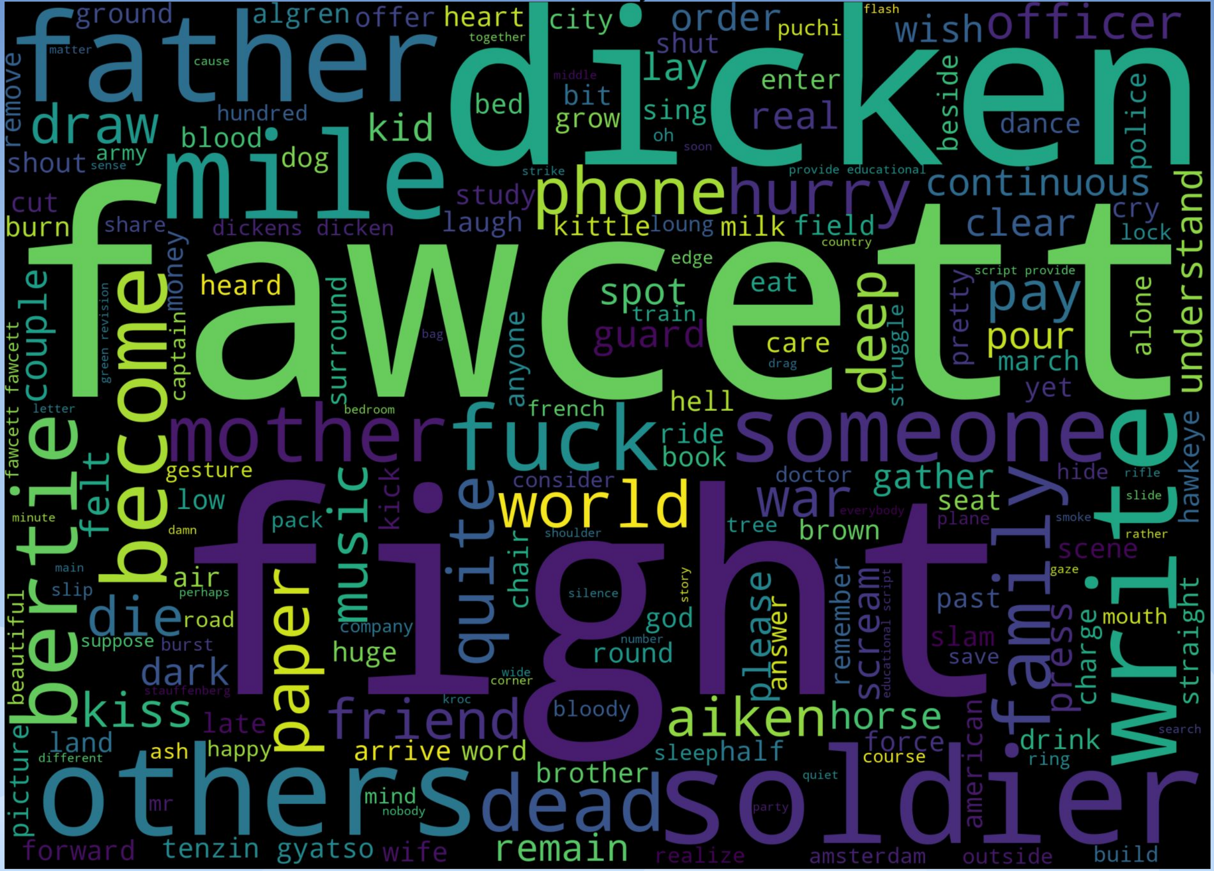

Family
clear
dissolve
stuff
cop
bag
slide
mother
sword
smash
laugh
disappear
dead
emerge
power
scene
bar
shut
deep
tree
soldier
completely
world
mind
pov
grow
ship
cry
cut
blast
heard
hutch
glow
monster
scream
land
oh
peer
sorry
corner
dog
paper
ball
hey
hop
quite
slip
straight
everybody
beautiful
struggle
father
free
snap
block
road
draw
flash
dive
gather
mom
pretty
entire
heart
others
crack
guard
truck
course
coraline
dance
family
spot
lock
different
bit
hurry
ground
half
goofy
book
charge
together
hopps
react
stage
arrive
uncle
guess
press
realize
cause
float
leg
flip
surround
someone
saroo
mouth
bedroom
hell
dad
care
kid
leap
become
park
friend
air
brown
actually
speed
tiny
idea
build
picture
hiccup
suppose
chair
exit
pop
anyone
wake
phone
pour
forget
edge
grabs
wish
ring
low
pay
enter
spin
bed
dark
school
human
gru
field
seat
train
eat
hole
please
forward
fight
sing
fox
kiss
save
animal
nice
happy
drink
hallway
dragon
past
shoulder
write
punch
main
beside
huge
slam
real
sleep
minion
pile
middle
toss
ride
force
wan
superman
bottle
herrick
kick
hug
giant
valiant
hall
wide
paint
remain
alone
slow
bean
continuous
drag
story
music
crash
couple
kitchen
burst
lay
remember
die
hide
blue

Fantasy

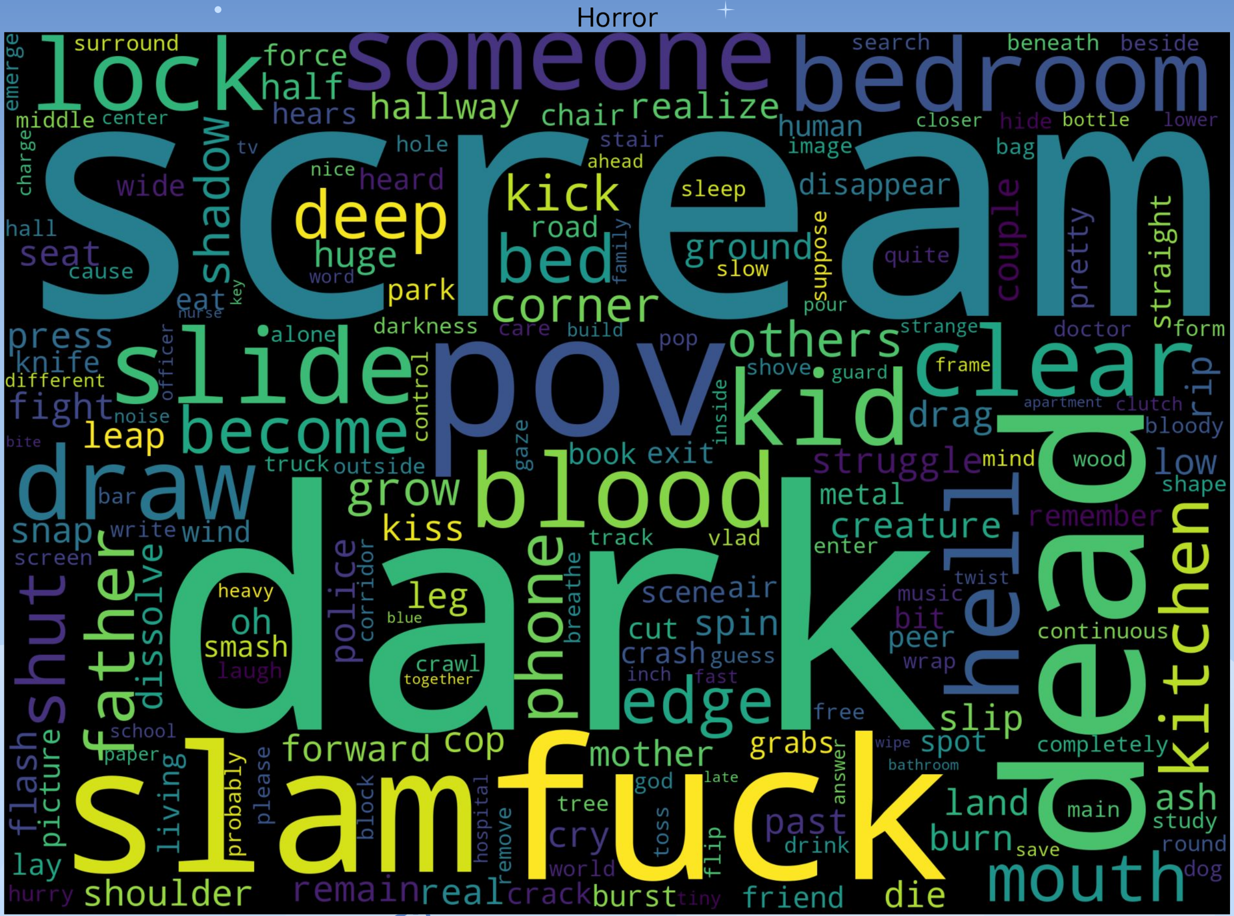

Fantasy
scream
land
dark
draw
mother
guard
dead
kid
sword
huge
slam
fight
become
realize
world
ground
stone
edge
father
mouth
kiss
shadow
eat
power
bed
city
shut
someone
oh
others
past
laugh
grow
dance
vlad
deep
die
remain
tiny
blue
lock
save
friend
slide
crash
burst
creature
real
kick
air
spot
blood
burn
spin
force
struggle
forward
tree
leap
flash
pov
smash
hurry
snap
crack
bedroom
soldier
kitchen
ring
wind
drag
remember
human
bit
scene
rip
quite
lower
hell
god
slip
gaze
heard
toss
continuous
straight
wheel
alone
drink
search
track
whale
hide
book
seat
main
pop
lay
ship
clear
low
heart
police
leg
peer
care
roar
sing
sleep
fear
giant
horse
phone
ride
fuck
control
castle
corner
disappear
enter
image
ash
beneath
beautiful
knife
couple
lestat
entirecircle
mind
half
afraid
hears
massive
emerge
cry
together
wrap
build
death
grabs
building
arrive
remove
shoulder
strike
soon
guess
music
nice
apartment
surround
bag
word
cut
press
cause
green
blast
offer
glow
beside
answer
write
metal
wide
middle
exit
family
monster
charge
suppose
dog
wish
hole
pretty
wake
road
dive
forget
study
please
lower
hall
chair
different
smoke
bar
shove
block
strange
gather
free
wood
dissolve
form
pour
hellboy
picture
slow
pay

History

History



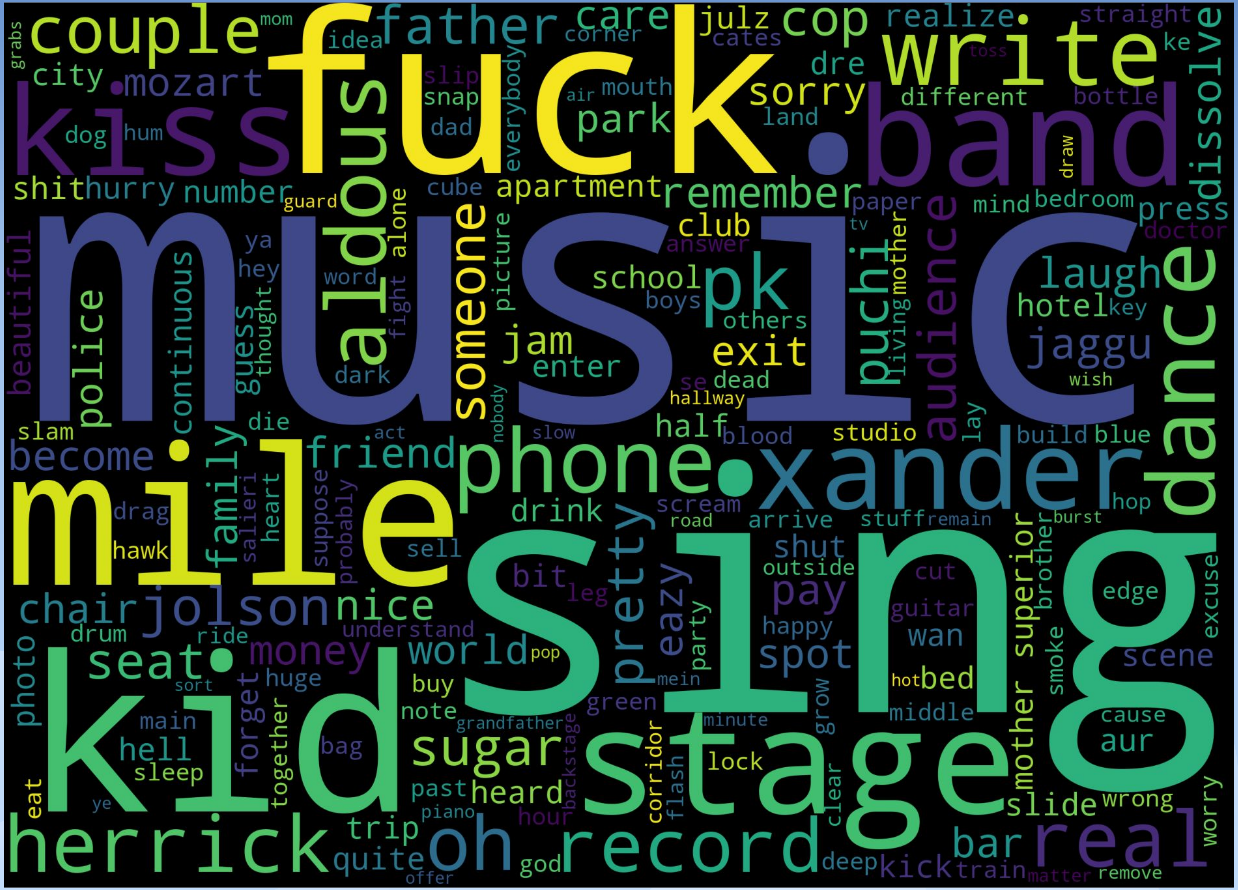
Horror



Horror
emerge
surround
lock
force
half
someone
search
bedroom
beneath
beside
middle
center
hears
hallway
chair
realize
human
closer
hide
bottle
lower
charge
scream
shadow
tv
hole
stair
image
bag
nice
ahead
wide
heard
deep
kick
sleep
disappear
couple
pretty
straight
hall
road
family
suppose
seat
huge
bed
ground
quite
cause
word
slow
park
key
eat
nurse
corner
pour
press
alone
darkness
care
build
pop
strange
doctor
form
knife
officer
slide
control
pov
others
clear
different
shove
guard
frame
rip
fight
inside
kid
apartment
clutch
noise
become
drag
bloody
bite
leap
gaze
book
exit
truck
outside
struggle
mind
wood
low
draw
blood
held
dead
shape
bar
grow
metal
write
kiss
remember
snap
wind
creature
track
vlad
enter
screen
police
corridor
phone
breathe
twist
heavy
leg
scene
air
music
shut
father
dissolve
blue
bit
oh
dark
cut
spin
continuous
peer
smash
crash
guess
kitchen
laugh
crawl
wrap
together
inch
fast
edge
free
slip
school
flash
forward
cop
grabs
wipe
spot
completely
paper
mother
answer
bathroom
picture
living
probably
please
slam
block
hospital
fuck
god
late
tree
land
ash
main
study
past
cry
remove
toss
flip
drink
burn
save
round
lay
world
mouth
dog
hurry
shoulder
remain
real
crack
burst
tiny
friend
die

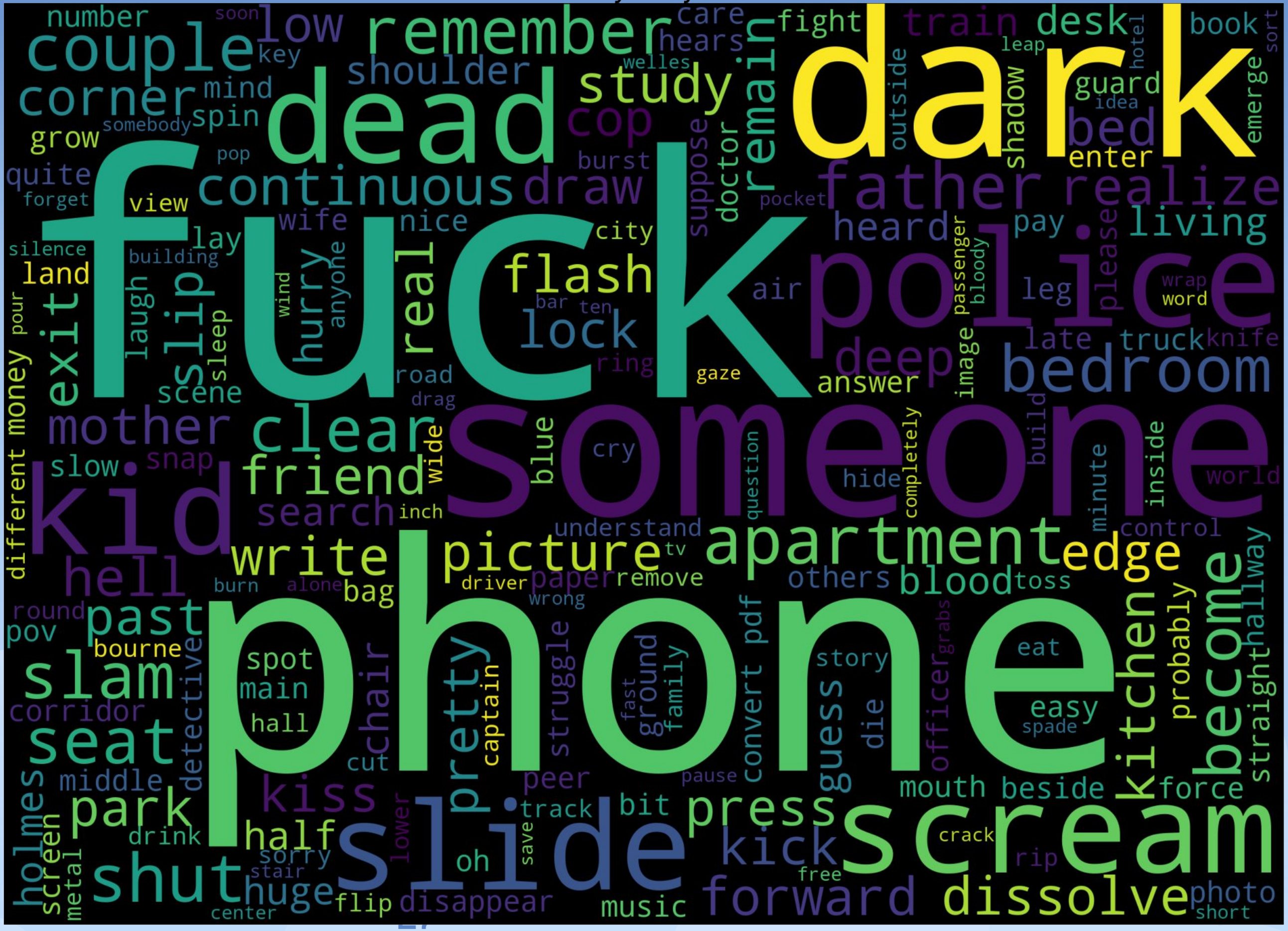
Music

Music

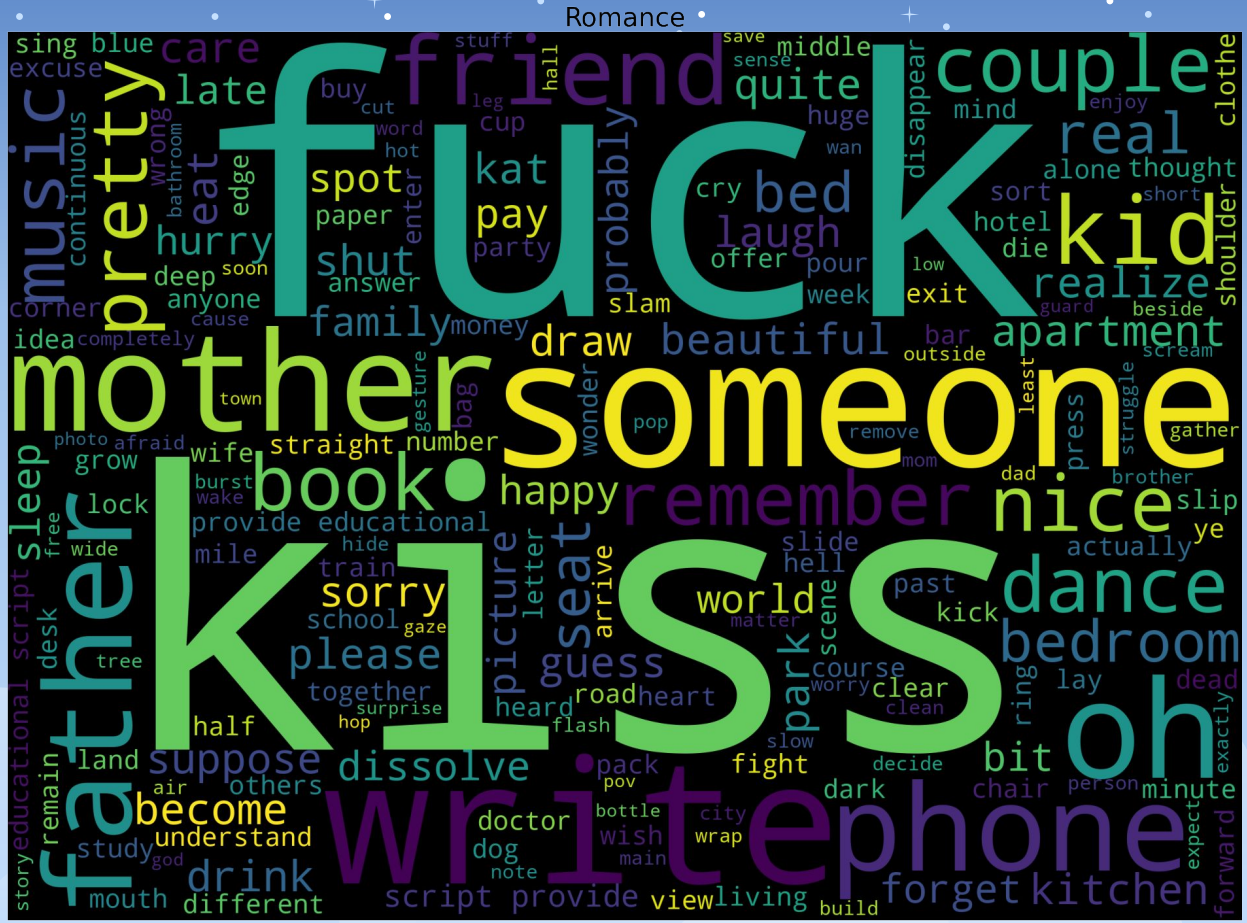

music
sing
fuck
kid
mile
stage
band
kiss
xander
phone
dance
herrick
record
oh
real
write
couple
father
aldous
pk
puchi
audience
someone
sugar
jolson
seat
family
friend
pretty
dissolve
cop
care
julz
realize
straight
mom
idea
corner
cates
ke
toss
grabs
city
mozart
slip
snap
air
mouth
sorry
dre
different
bottle
dog
hum
dad
everybody
park
land
draw
shit
hurry
number
guard
cube
apartment
remember
paper
mind
bedroom
press
alone
club
tv
mother
doctor
ya
answer
beautiful
hey
word
picture
school
laugh
continuous
guess
thought
fight
boys
others
hotel
key
police
jam
exit
living
jaggu
dark
enter
se
dead
wish
hallway
nobody
slam
die
act
slow
half
blood
studio
lay
build
blue
become
drink
scream
hop
drag
salieri
heart
suppose
probably
road
arrive
stuff
remain
burst
superior
brother
hawk
sell
shut
outside
cut
edge
eazy
bit
leg
pay
guitar
excuse
chair
nice
understand
happy
wan
drum
ride
party
photo
money
world
pop
spot
scene
sort
huge
buy
mein
hot
bed
smoke
forget
main
note
grandfather
green
minute
grow
middle
cause
hell
bag
aur
sleep
together
lock
backstage
corridor
flash
clear
mother
eat
ye
past
heard
slide
wrong
worry
piano
hour
trip
quite
god
offer
deep
kick
train
bar
matter
remove

Mystery

Mystery

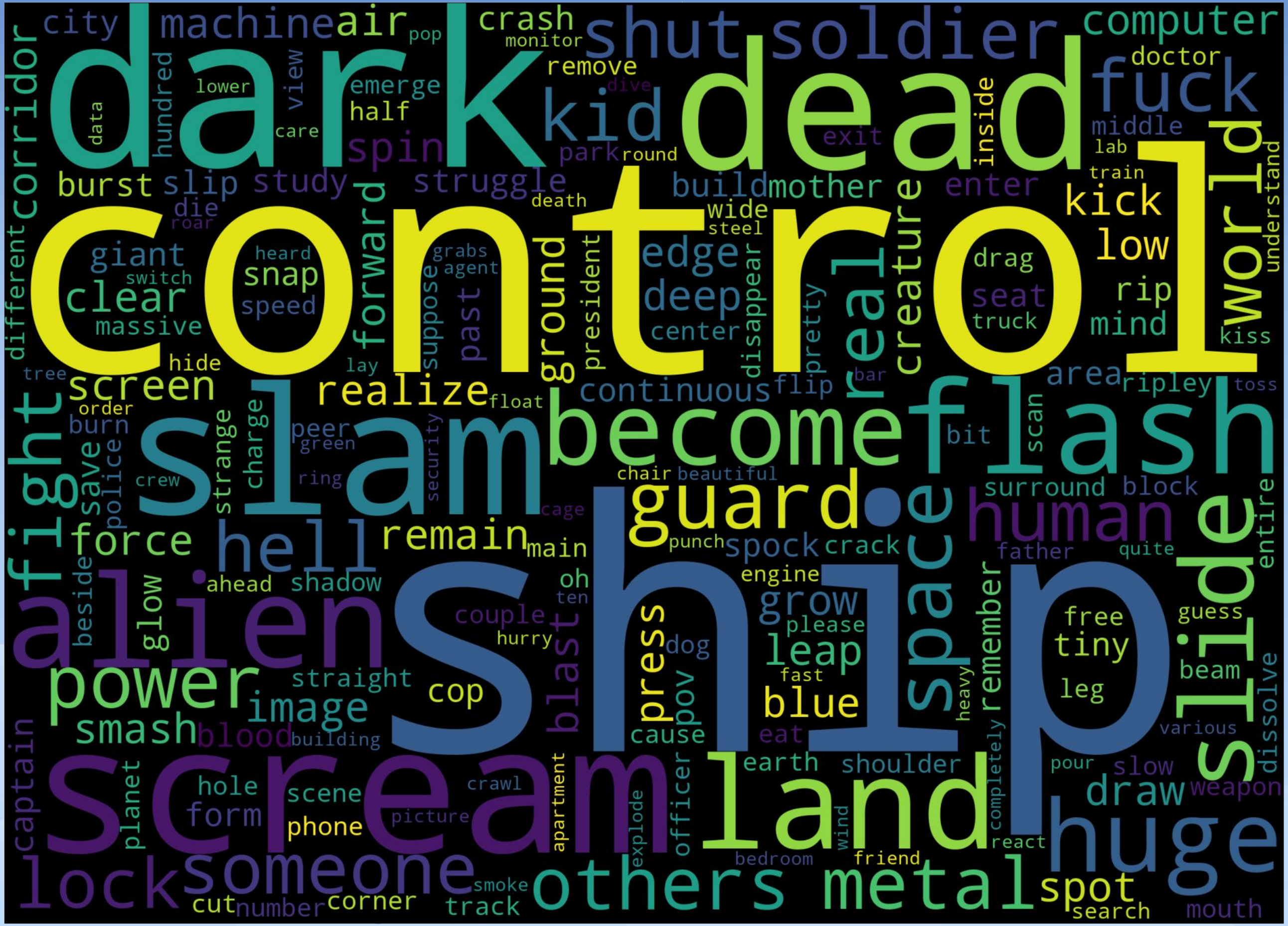


Romance


Romance
fuck
kiss
mother
someone
father
write
phone
book
remember
nice
dance
bedroom
oh
couple
friend
kid
real
realize
apartment
beautiful
draw
music
pretty
hurry
spot
shut
kat
pay
bed
laugh
family
sorry
please
guess
suppose
dissolve
become
understand
drink
forget
kitchen
living
script
provide
different
mouth
study
doctor
dog
wish
bit
fight
dark
chair
minute
person
dead
lay
clear
worry
course
road
heart
heard
together
surprise
half
land
others
air
pack
decide
pov
bottle
city
wrap
main
note
view
build
forward
expect
exactly
story
educational
remain
desk
tree
sleep
lock
grow
wife
straight
number
happy
provide
educational
mile
train
hide
wide
free
picture
letter
seat
arrive
park
scene
world
school
gaze
slide
hell
past
kick
matter
actually
ye
slip
brother
dad
mom
remove
pop
gather
press
struggle
scream
outside
least
bar
beside
guard
shoulder
clothe
enjoy
mind
huge
cup
quite
middle
sense
save
stuff
hall
ieg
cut
buy
word
hot
wan
disappear
probably
party
offer
pour
low
exit
week
slam
hotel
die
sort
short
alone
thought
cry
enter
paper
answer
deep
soon
anyone
cause
corner
idea
completely
money
draw
bag
gesture
wonder
town
burst
wake
afraid
photo
eat
edge
bathroom
wrong
continuous
excuse
sing
blue
care
late

SciFi

SciFi



city
machine
air
pop
crash
monitor
shut
soldier
computer
dark
dead
doctor
fuck
corridor
data
hundred
lower
view
emerge
half
care
remove
dive
kid
exit
inside
middle
lab
spin
park
round
train
world
understand
burst
slip
study
struggle
death
build
mother
enter
kick
control
die
roar
forward
wide
steel
low
giant
heard
snap
grabs
agent
edge
drag
switch
clear
speed
suppose
past
ground
president
disappear
deep
real
creature
seat
rip
different
massive
pretty
center
truck
mind
kiss
hide
lay
tree
bar
area
ripley
toss
screen
realize
float
continuous
flip
order
fight
burn
become
flash
slam
peer
green
security
scan
bit
save
police
crew
strange
charge
ring
chair
beautiful
surround
block
cage
guard
human
entire
force
hell
remain
main
punch
spock
crack
father
quite
beside
ahead
shadow
oh
engine
glow
alien
ten
couple
grow
space
remember
ship
slide
free
guess
hurry
please
dog
leap
tiny
power
straight
blast
press
fast
pov
heavy
beam
leg
image
cop
blue
dissolve
smash
blood
building
cause
eat
various
captain
scream
earth
shoulder
pour
slow
hole
crawl
apartment
draw
weapon
planet
form
scene
picture
explode
officer
land
completely
phone
wind
huge
react
someone
bedroom
friend
lock
smoke
others
metal
spot
cut
number
corner
track
search
mouth

Thriller

Thriller

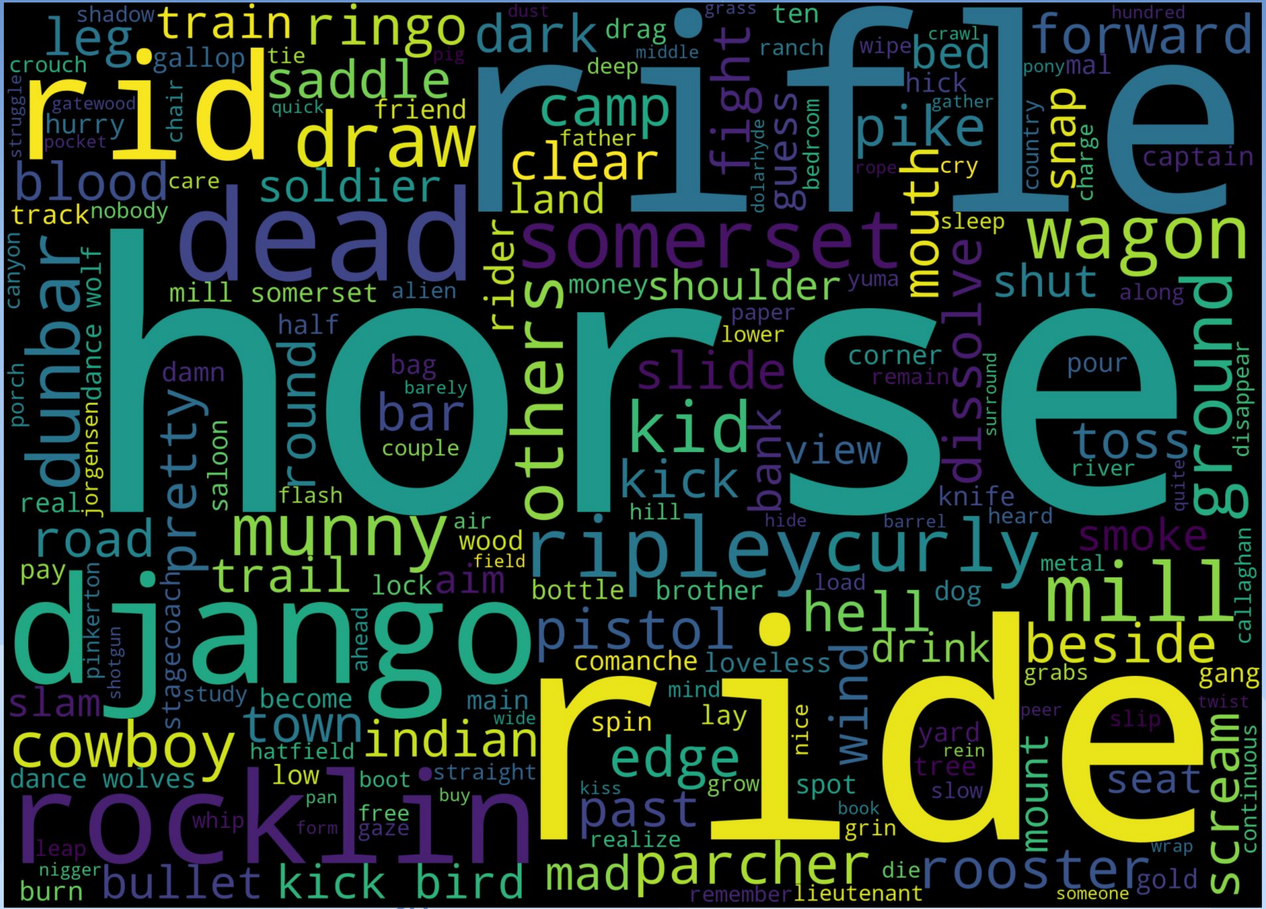
War

War


war
company
horse
guard
bomb
drink
mother
hundred
field
world
barrack
letter
scream
ride
radio
order
iraqi
round
military
ahead
friend
outside
load
hears
captain
father
view
land
realize
szpilman
colonel
music
brother
beside
crawl
snap
forward
bit
hide
phone
tight
pretty
bullet
silence
position
desert
kittle
become
drag
draw
leg
past
blood
fuck
educational
script
rid
postman
family
prisoner
corner
ship
bloody
word
peer
tree
ground
hurry
plane
deep
area
slam
sanborn
disappear
algren
trench
fight
blue
city
hub
wood
march
hawkeye
loung
save
dog
reverend
slow
provide
pack
suppose
weapon
wound
officer
hill
study
dig
build
pour
cry
picture
agu
care
air
sleep
grow
sense
village
american
green
sergeant
provide
educational
machine
write
french
remain
flash
kiss
half
train
script
schofield
chief
vlad
soon
couple
shout
hell
die
road
kick
dance
ten
tuvia
tent
river
laugh
burst
low
camp
ape
kid
real
japanese
soldier
nobody
gate
someone
anyone
steiner
straight
eat
hole
truck
press
clear
different
paper
cut
ring
enter
edge
prew
spot
clean
offer
huge
smoke
quite
smash
struggle
bed
search
sword
lay
heavy
shoulder
force
stauffenberg
groucho
dark
slip
army
dissolve
shut
charge
dead
others
rifle
gather
mouth
seat
wide
burn
lieutenant
enemy
battle
understand
remove
attack
heard
remember

Western

Western



The Machine - Extended Version

Input:
Screenplay Text

Text Preprocessing
- Lemmatization
- Removal of stopwords

Feature Extraction
- Count Vectorization
- TFDI

MultiLabel Classifier
- OnevsRestClassifier
- Chain Classifier

Output:

[0,1,0,1,0,0,0...]

Output

- An array of 18 elements
- Each position of the array is represented by a unique genre
- Elements are either 1 or 0 (True or False)

Example:

Let's say I input a horror, action screenplay into the machine.

The predicted output could be:

[0, 0, 0, 1, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0]

Action **Thriller** **Horror**

Metric to Measure Success:

Hamming Loss: **The fraction of the wrong labels to the number of labels**

Example:

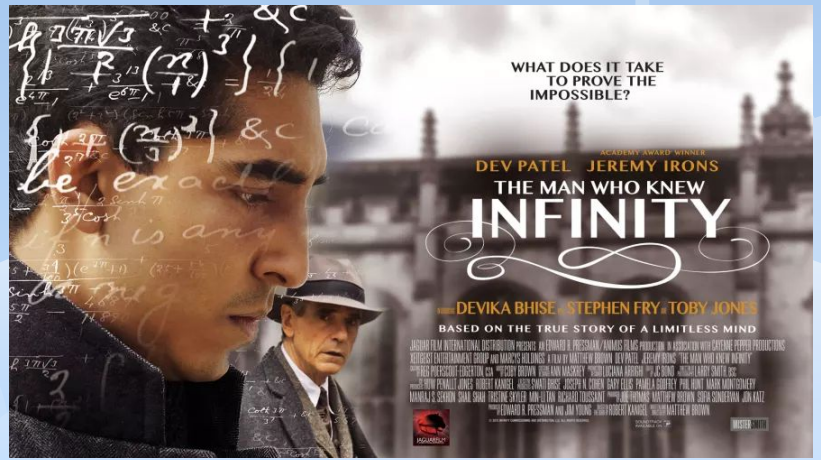
Predicted: [0, 0, 0, 1, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0]

Actual: [0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0]

Hamming Loss = 0.055

Goal of the Machine:

To minimize hamming loss; the closer to zero the better

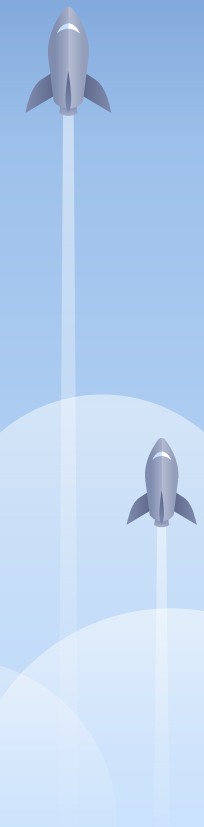


Classification Models Tested:

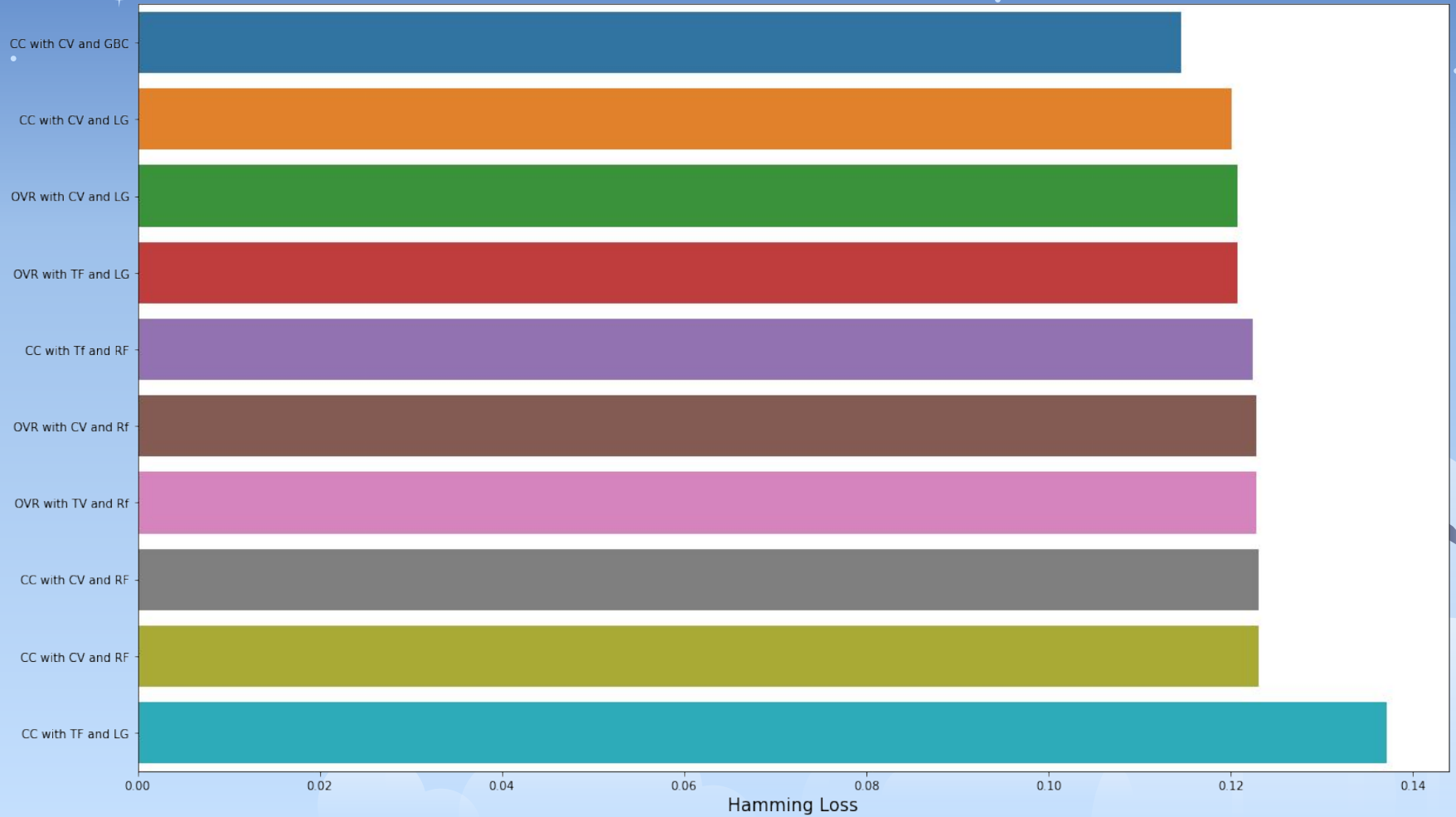
- **Classifier Chain**
- **OneVsRestClassifier**

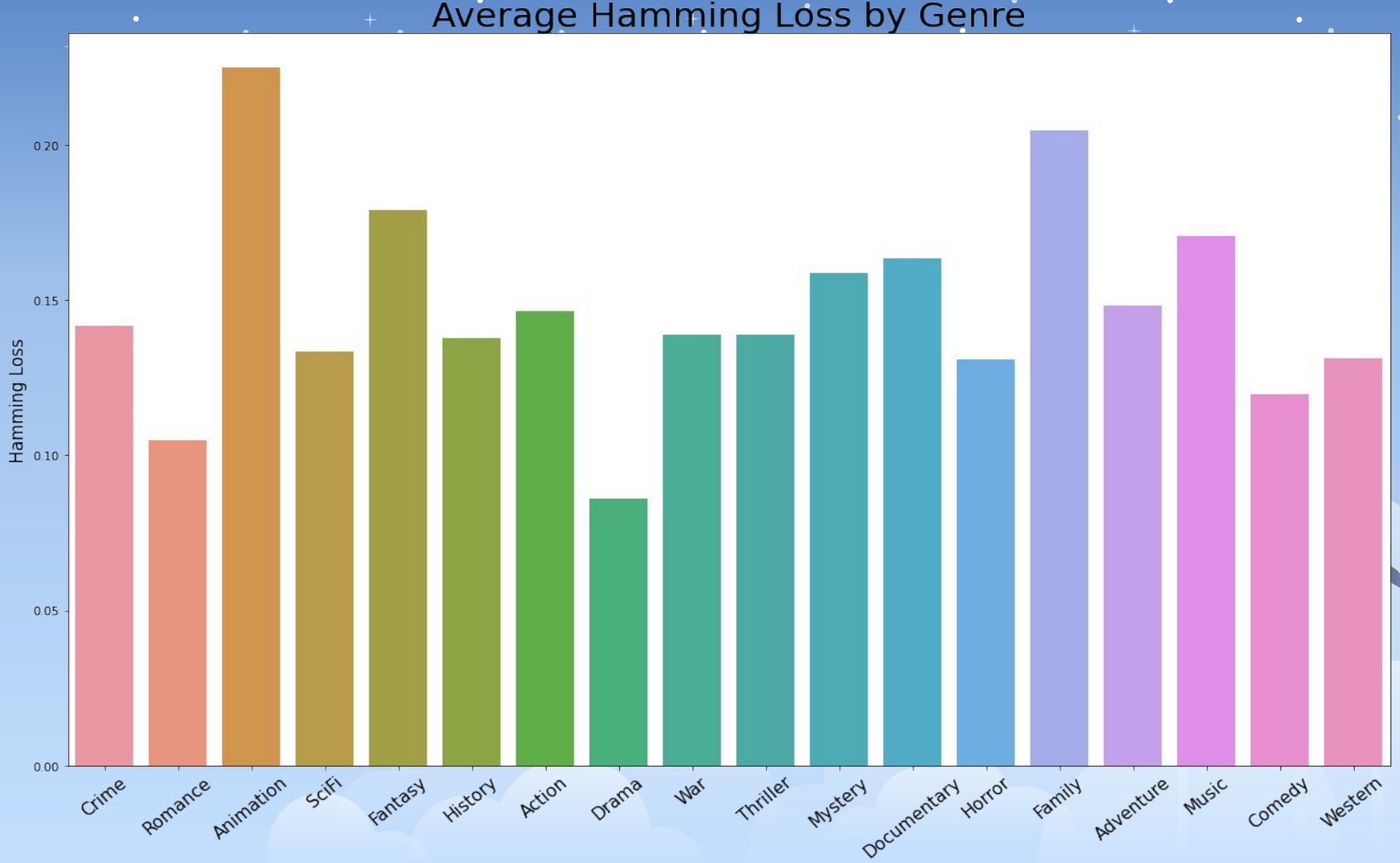
Both classification models takes in a classifier as a base estimator (Logistic Regression, RandomForestClassifier, etc) and arranges classifiers into a chain.

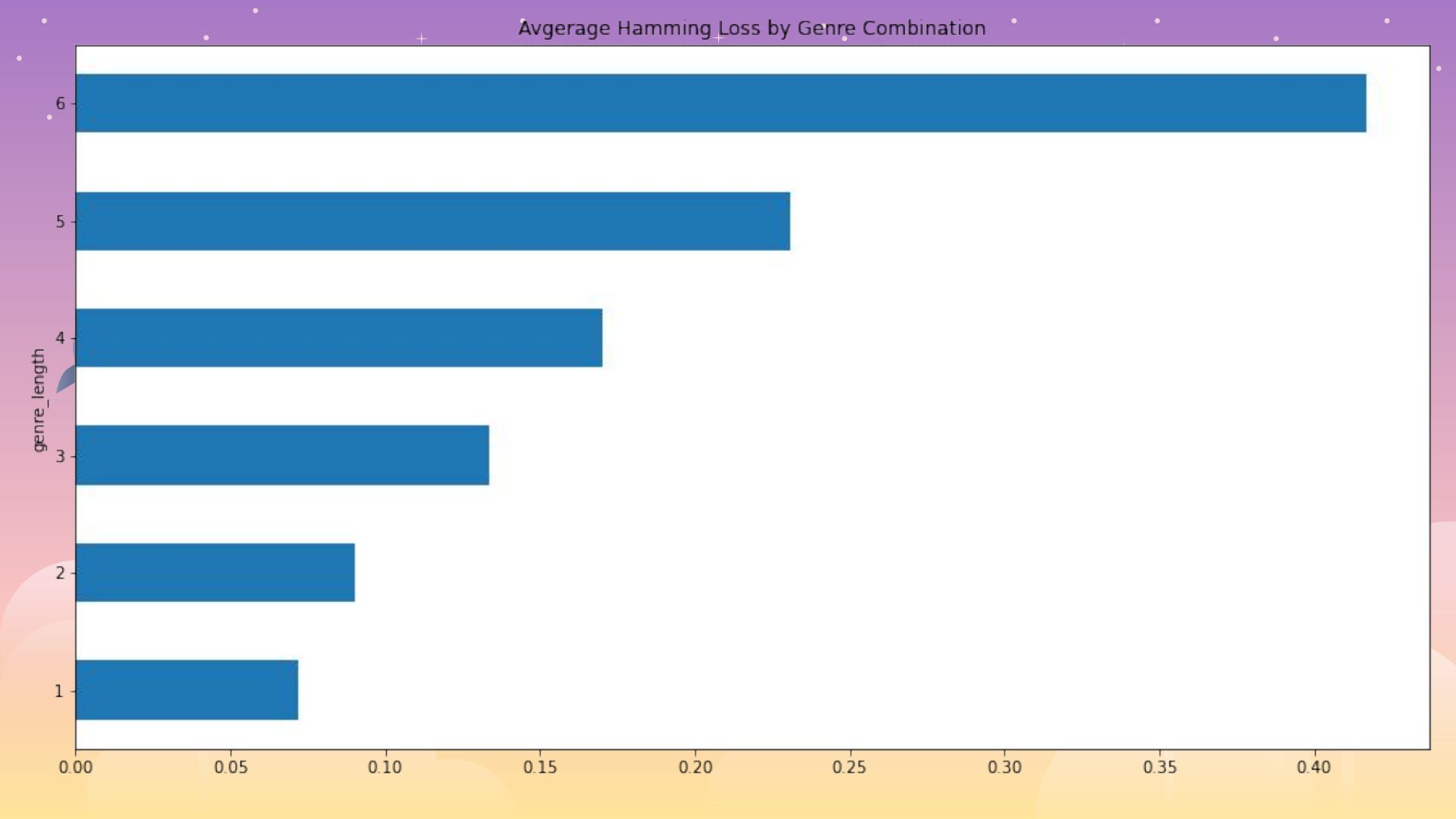
Predictions from previous classifiers are used as features for the next classifier



Average Hamming Loss between Model Types




Average Hamming Loss by Genre
Hamming Loss
0.00
0.05
0.10
0.15
0.20
Crime
Romance
Animation
SciFi
Fantasy
History
Action
Drama
War
Thriller
Mystery
Documentary
Horror
Family
Adventure
Music
Comedy
Western


Avgerage Hamming Loss by Genre Combination
genre_length
6
5
4
3
2
1
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40

Movies from the test set ->

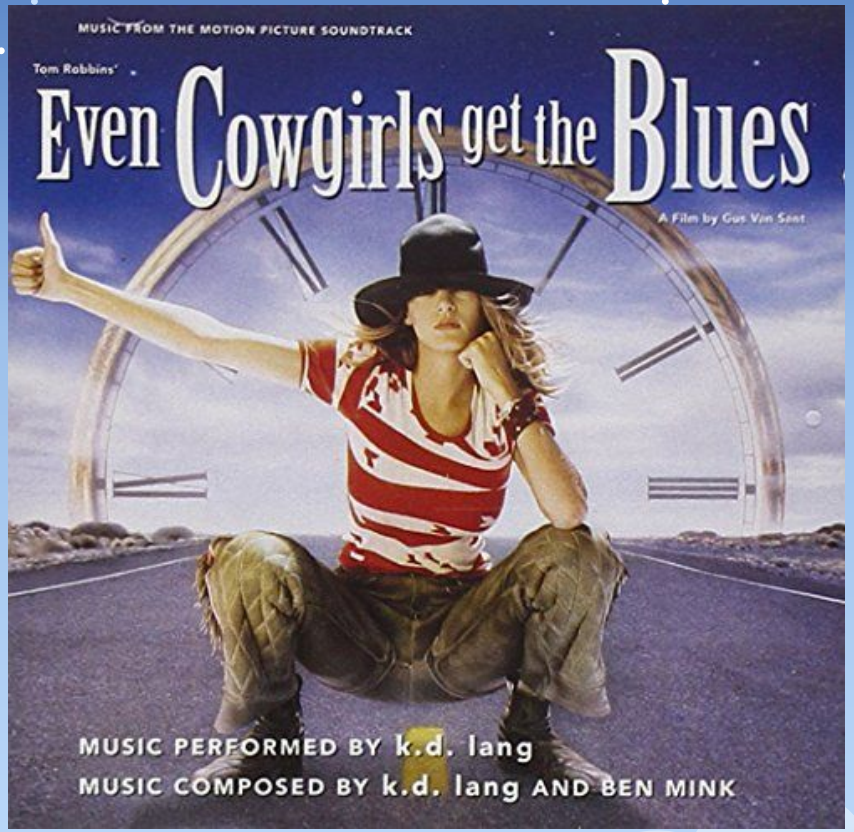
Predicted: SciFi, Action, Adventure

Actual: SciFi, Adventure

Hamming Loss: 0.055

Tomato Rating: 83%

Box Office: 400 million



Predicted: Drama

Actual: Romance, Drama, Comedy, Western

Hamming Loss: 0.167

Tomato Rating: 18%

Box Office: 1.7 million

Predicted: Comedy

Actual: Thriller, Horror

Hamming Loss: 0.167

Tomato Rating: 85%

Box Office: 47.3 million

Suggestions & Takeaways

- Genres can be identified given a screenplay
- Some screenplays that fit the genre do well with ratings while others don’t
  - However, having the right genre doesn’t mean high ratings
- Studios can use a genre classifier to pick screenplays

Improvements & Future

- Use other Natural Language Processing Techniques to create new features
  - Topic modeling, etc
- Classify and predict other film elements
  - Theme, Act Structure, etc
- A.I generated screenplays
  - Screenwriters cost a lot of money

Any Questions?

Thank you!

https://github.com/mikeyo4800

https://www.linkedin.com/in/michael-orlando-2835a3149