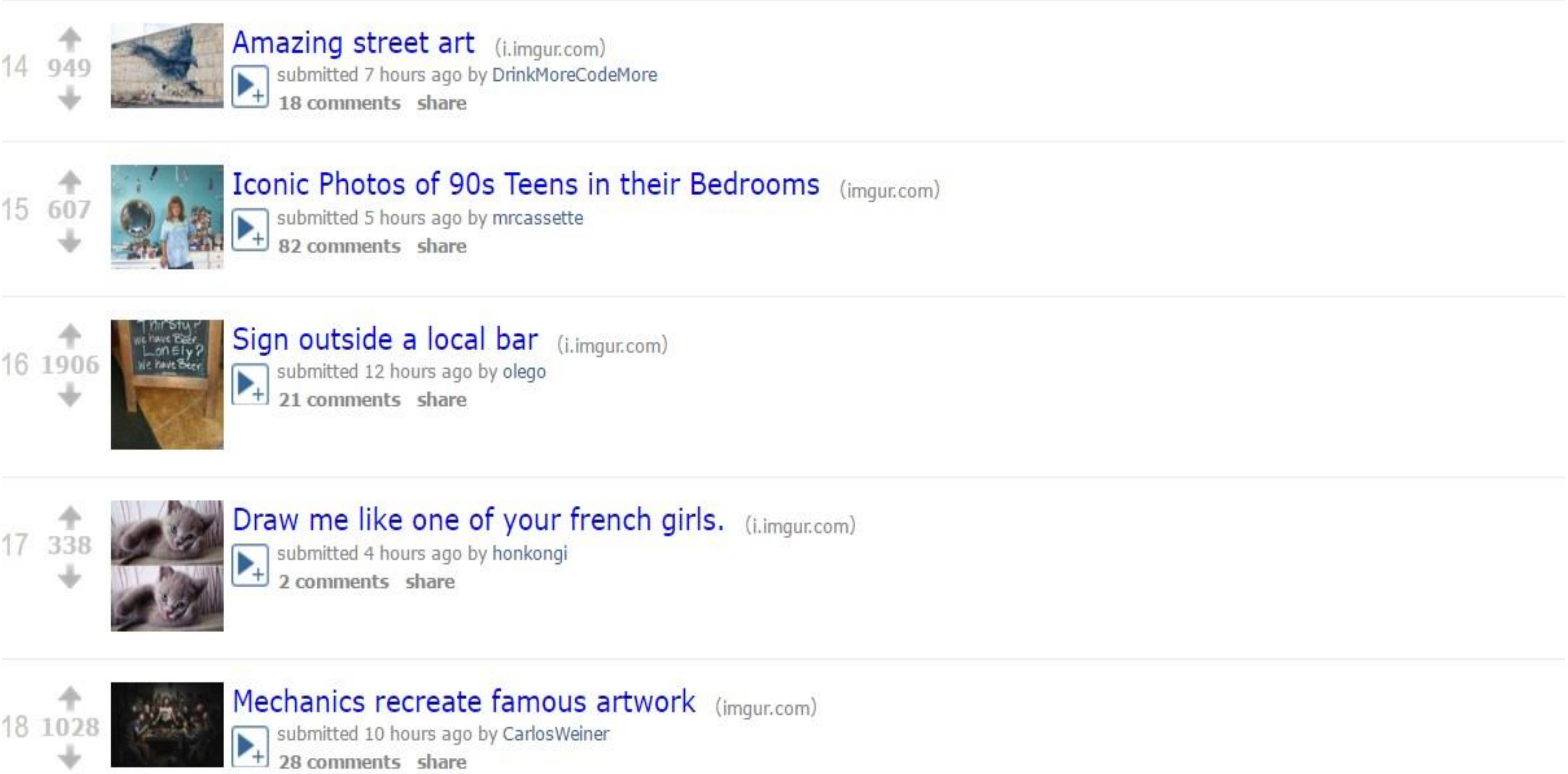
MDS and Clustering of Reddit Subreddit Posts

Dakota Cintron

Reddit & Subreddits

- Reddit is an entertainment, social news networking service, and news website

- Subreddits are essentially forums on different topics



14 949 Amazing street art (i.imgur.com)
submitted 7 hours ago by DrinkMoreCodeMore
18 comments share

15 607 Iconic Photos of 90s Teens in their Bedrooms (imgur.com)
submitted 5 hours ago by mrcassette
82 comments share

16 1906 Sign outside a local bar (i.imgur.com)
submitted 12 hours ago by olego
21 comments share

17 338 Draw me like one of your french girls. (i.imgur.com)
submitted 4 hours ago by honkongi
2 comments share

18 1028 Mechanics recreate famous artwork (imgur.com)
submitted 10 hours ago by CarlosWeiner
28 comments share

Goals

- Explore the (dis)-similarity of 14 highly subscribed to subreddits:
    - AskReddit
    - Funny
    - TodayILearned (TIL)
    - Pics
    - Science
    - Worldnews
    - IAmA
    - Videos
    - Gaming
    - Movies
    - Aww
    - News
    - AskScience
    - Television

Methods

- Employ **python** and **scrapy** framework to obtain as many post titles as possible from each subreddit
  - Uses **“spiders”** , “**selectors**, & “**regex”** to navigate webpage HTML
- Employ **text analysis** methods in R to clean the data
  - Identify the number of unique words
  - Identify frequently used words
  - Make word clouds
  - Use r packages **tm** and **wordcloud**
- Use **number of words shared** between subreddits as measure of similarity
  - The greater the # of shared words, the more similar the subreddits
- Employ MDS (SMACOF in R) & Clustering Methods

Text Objects in R

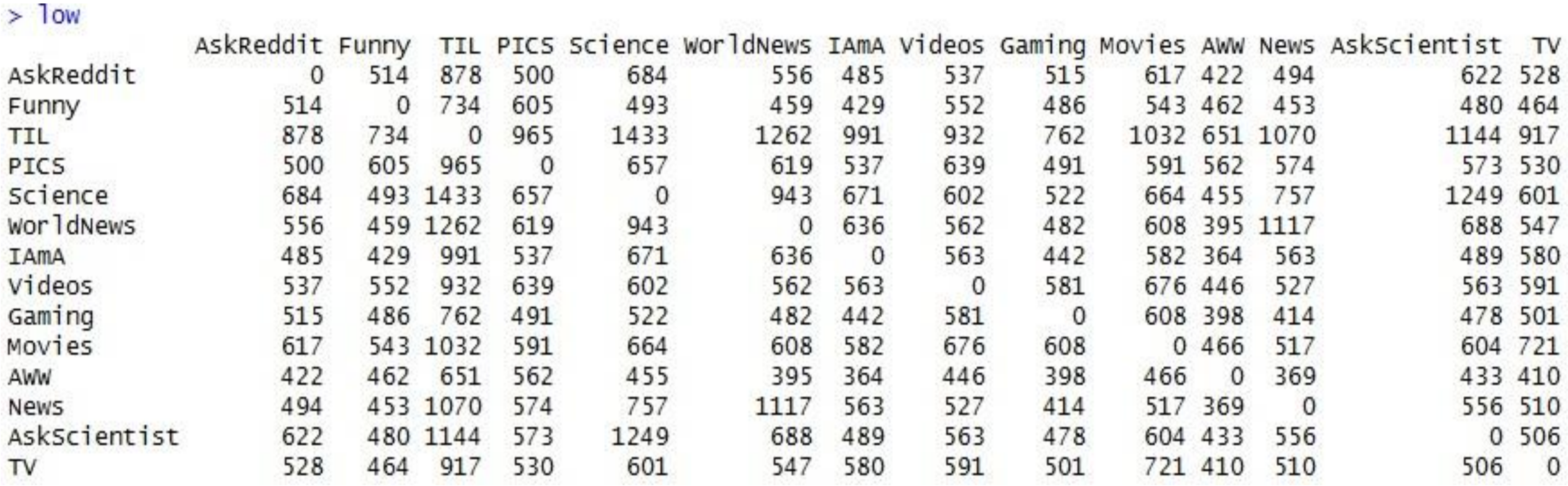
```
> review_text.asi [1] "How do fruits \"know\" that animals spread their seeds?,RichardsIsGod How
were the very first computer languages/operating systems coded?,HungoverHero777 Why do some
people's bruises readily appear while others hardly show?,FarFromAmusing What are the
biological/pharmacological mechanics involved that cause tinnitus (ototoxicity) to occur after
taking antibiotics i.e. vancomycin/cephalexin?,xRAPIERx How do nuclear power plants draw and
convert energy for our use?,Dazd95 If there is no \"absolute\" velocity, how can we determine
centripetal force?,Huugnuut Is the set of all countably infinite sets countably
infinite?,NOTaCreativeUsername When looking at pictures of Earth taken from space, why do you
see only black and no stars?,romfus87 Is there any particular reason that some planets rotate
very fast on their axis, while others rotate slowly?,ed123dead Why does excessive wing angle of
attack (AoA) cause roll?,accounttoberacist Do light waves cast a shadow?,GOD_DAM_IT Is there a
liquid with the same density of our atmosphere at 1 atm? If so, would a glass full of it not
experience refraction?,ColonolCool How do microbes in the human body survive our immune
systems?,ChainedBroletariat Can you charge your phone from a plant? Is this legit?,EyePad Why
does the carboniferous period have a mean surface temp equal or lower than today, when it had
2x the atmospheric CO2?,Snaz5 Why does light change direction when it refracts?,pynoobpy
AskScience AMA Series: I'm Alexis Kaushansky, a Principal Investigator at the Center for
Infectious Disease Research in Seattle, WA. I research malaria and the interactions between
host and pathogens. Iâ€™m excited to talk to you about it. AMA!,CIDResearch Why do people
change the tone of their voice depending on who their talking to?,Mellow-as-fuk Why is the
counterweight arm of a trebuchet always shorter than the sling/launching
```

```
> head(frequency.tvs, n=20)

   show   season   new    series    game   thrones   discussion    shows   netflix
    64      51      43      37       30       29          27          27       22

    one    anyone    thread   best    dead    like   star   television  time
    19       17         17     16      16      16     16        16       16

    will   prince
    16       15
```

Results

Subreddit	Pages Captured	Posts Captured	Top 5 Words	Unique Words
AskReddit	39	907	Whats reddit thing youve ever	2752
Funny	36	828	just like dog dont new	**2422**
TodayILearned	35	802	til one first people can	**6367**
Pics	36	884	found like just one picture	3178
Science	33	756	new study researchers scientists may	5052
Worldnews	36	857	new says china police will	4120
IAmA	29	669	ama request iama ask author	3267
Videos	33	773	video new first ever like	2982
Gaming	34	779	game dark souls games new	2591
Movies	34	789	movie movies film one best	3356
Aww	39	924	little dog just meet cat	2445
News	**23**	**516**	man new police says found	2903
AskScience	**41**	**974**	can like black mass different	4010
Television	26	598	show season new series game	2833

Matrix of Similarities



```
> low
             AskReddit Funny  TIL PICS Science WorldNews IAmA Videos Gaming Movies AWW News AskScientist  TV
AskReddit            0   514  878  500     684       556  485    537    515    617 422  494          622 528
Funny              514     0  734  605     493       459  429    552    486    543 462  453          480 464
TIL                878   734    0  965    1433      1262  991    932    762   1032 651 1070         1144 917
PICS               500   605  965    0     657       619  537    639    491    591 562  574          573 530
Science            684   493 1433  657       0       943  671    602    522    664 455  757         1249 601
WorldNews          556   459 1262  619     943         0  636    562    482    608 395 1117          688 547
IAmA               485   429  991  537     671       636    0    563    442    582 364  563          489 580
Videos             537   552  932  639     602       562  563      0    581    676 446  527          563 591
Gaming             515   486  762  491     522       482  442    581      0    608 398  414          478 501
Movies             617   543 1032  591     664       608  582    676    608      0 466  517          604 721
AWW                422   462  651  562     455       395  364    446    398    466   0  369          433 410
News               494   453 1070  574     757      1117  563    527    414    517 369    0          556 510
AskScientist       622   480 1144  573    1249       688  489    563    478    604 433  556            0 506
TV                 528   464  917  530     601       547  580    591    501    721 410  510          506   0
```

Notes on Similarities

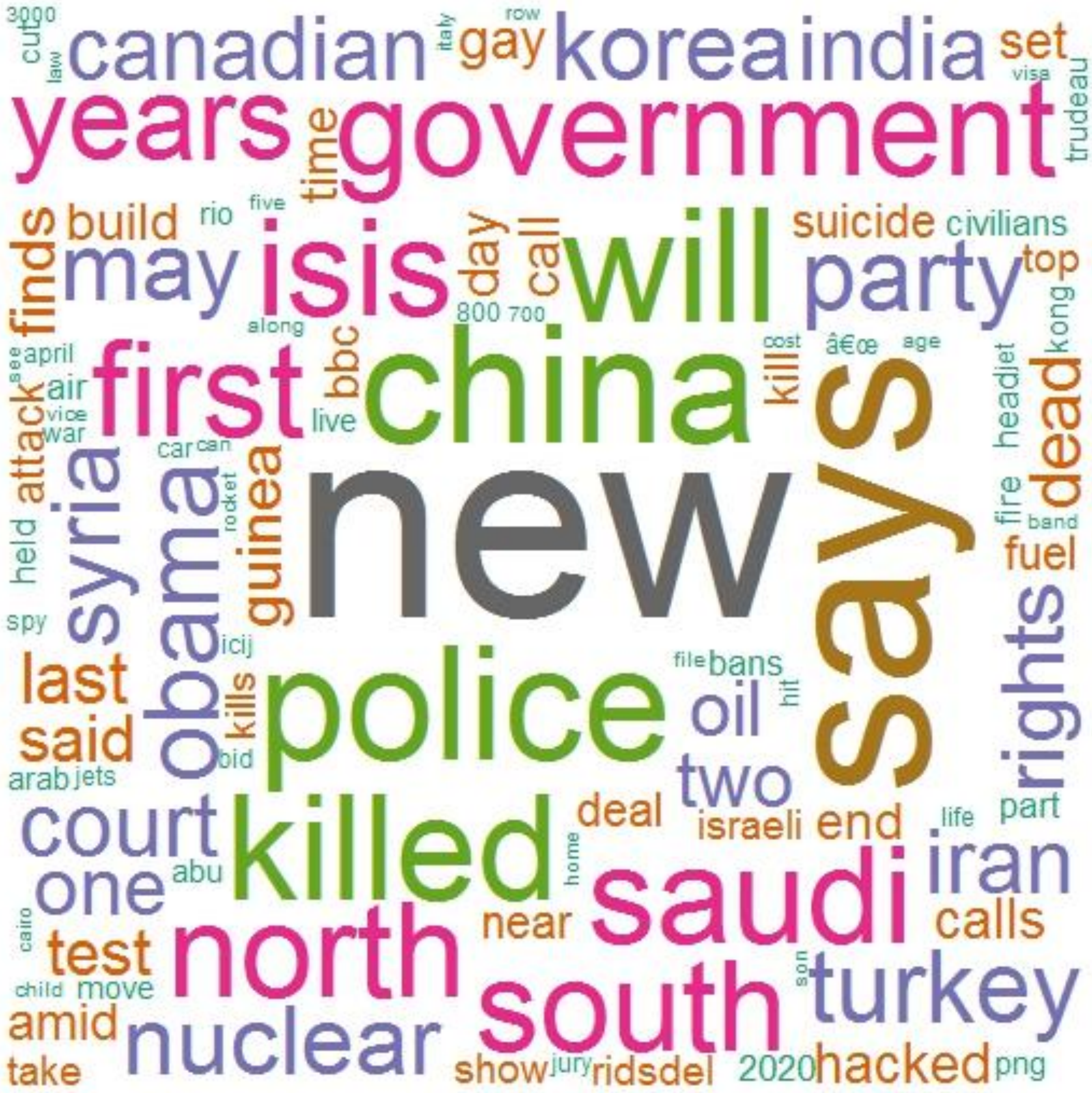
- TIL & Science most similar
  - 1143 matched words

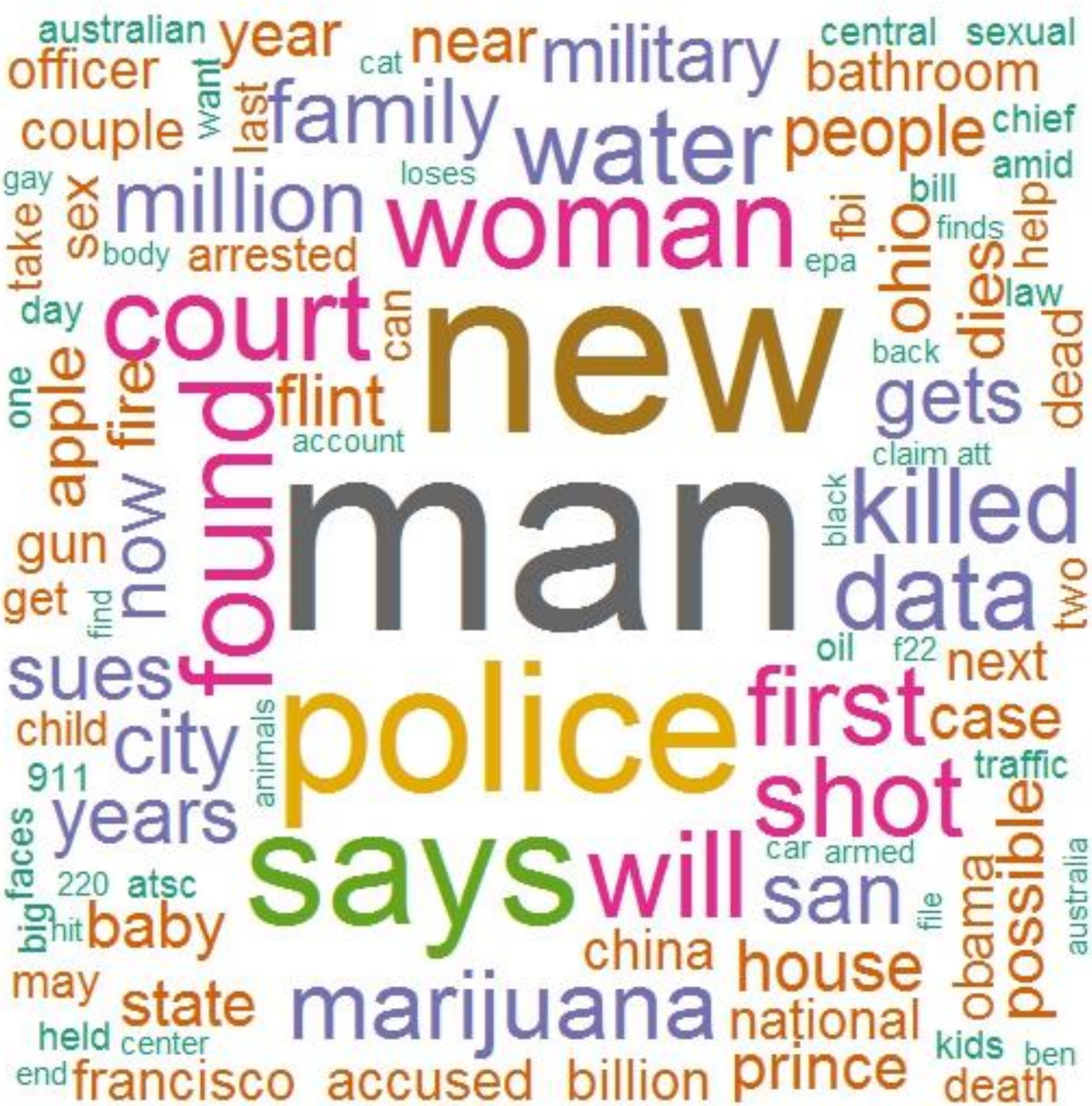
- AWW and IAmA least similar
  - 364 matched words

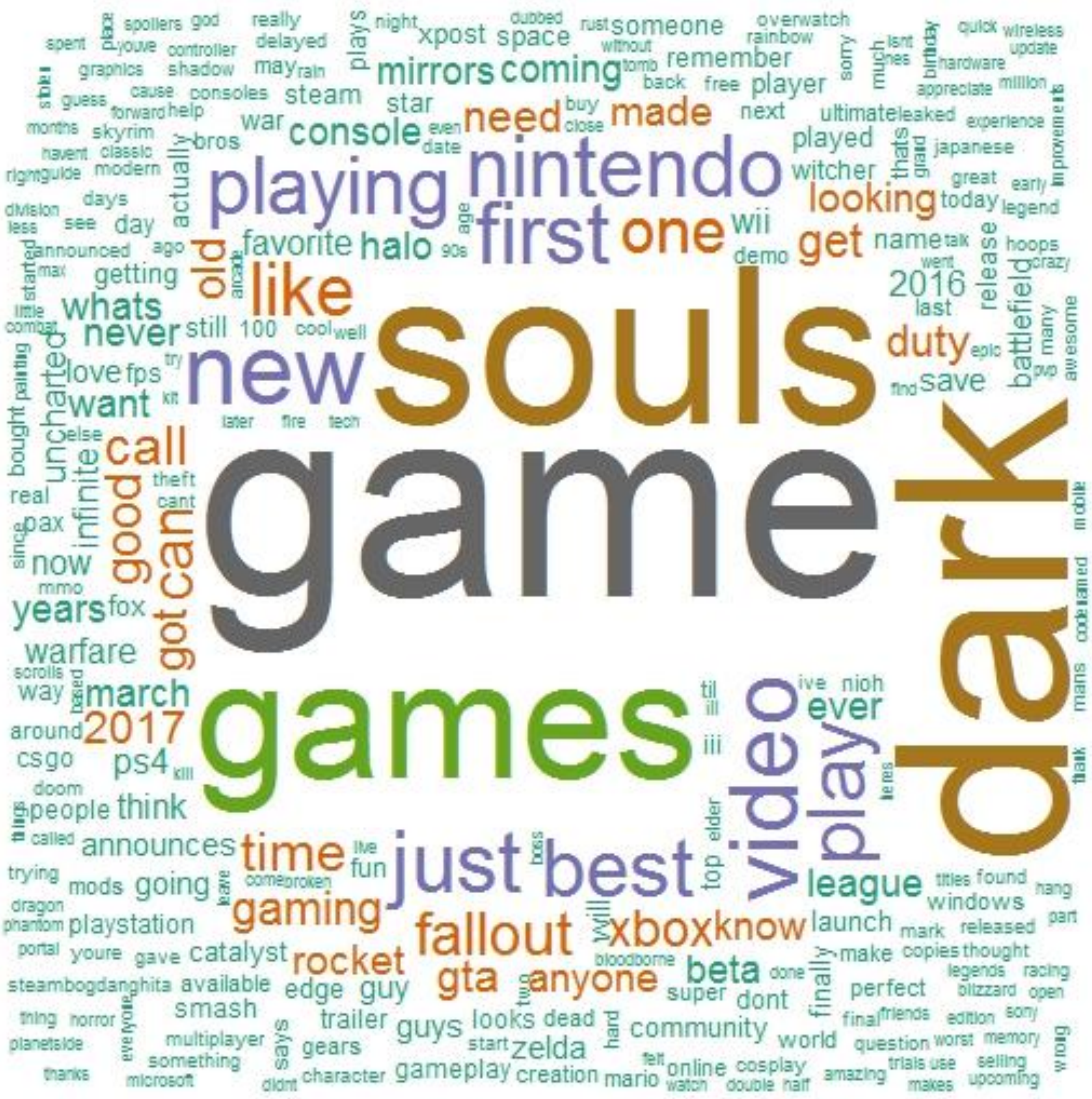
- TIL very similar to many of other subreddits (assuming this is driven by # of unique words in TIL)

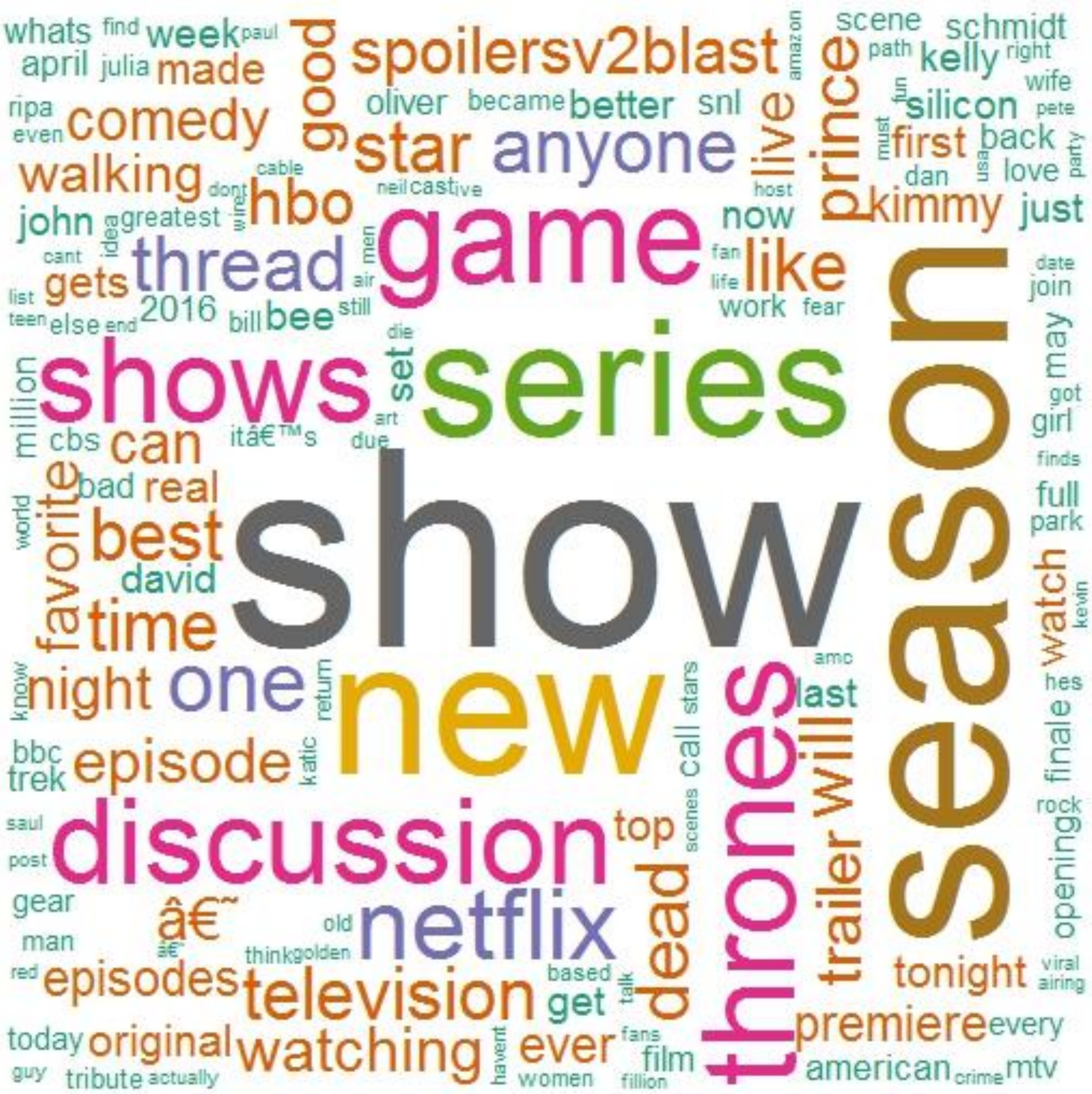
Word Cloud Examples

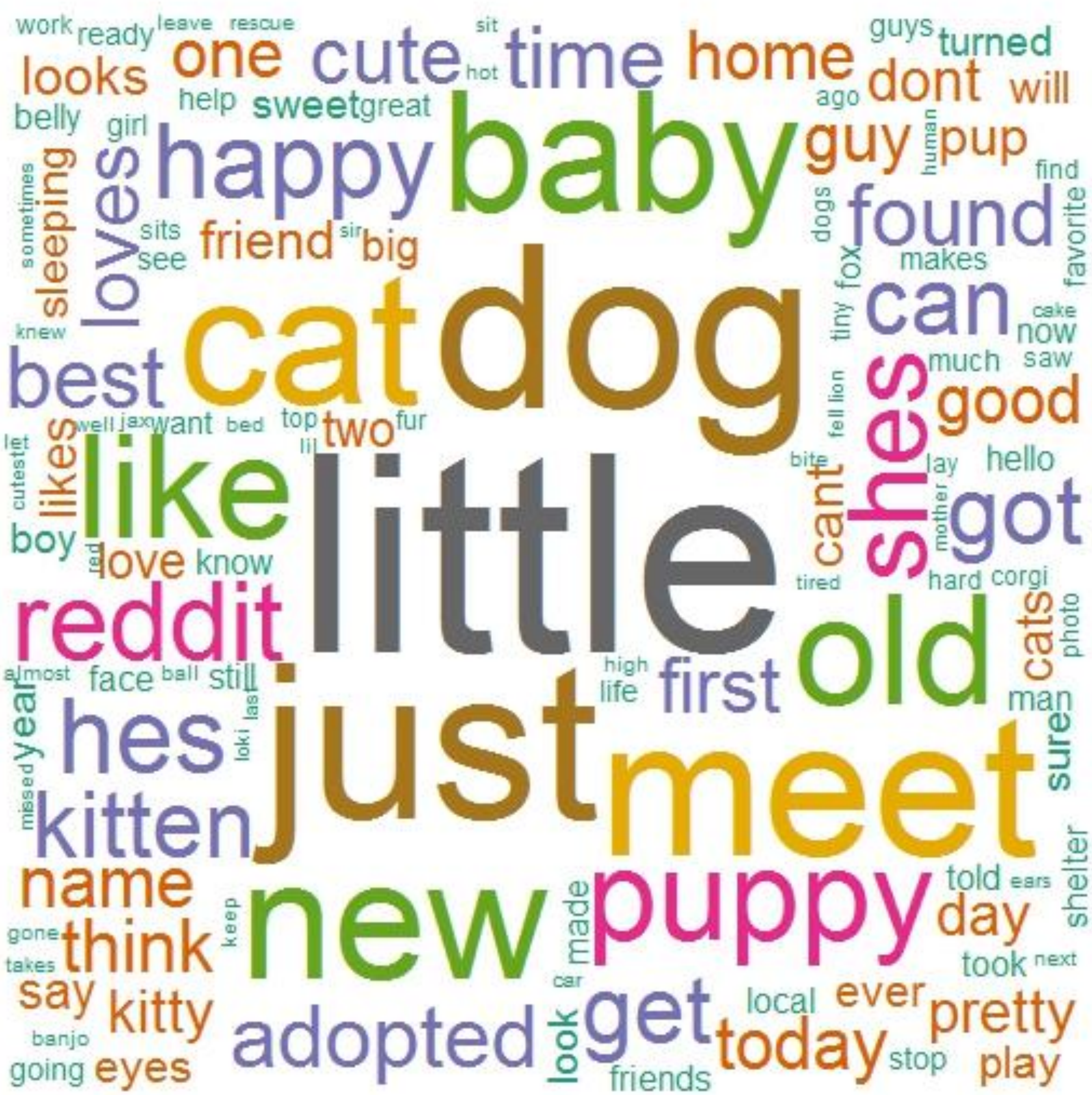
til
one
called
first
people
years
can
also
million
two
used
world
named
american
time
made
due
day
man
new
president
became
japanese
united
game
government
played
north
house
school
black
world
video
chinese
japan
human
john
living
created
considered
english
nuclear
military
british
without
series
released
company
character
war
city
since
make
different
island
back
national
german
killed
name
found
record
responsible
public
family
women
based
days
run
three
system
person
many
every
still
original


3000
cut
law
canadian
italy
gay
row
korea
india
set
visa
trudeau
years
government
time
build
rio
five
finds
may
isis
day
call
will
suicide
civilians
party
top
800
700
along
april
see
air
attack
vice
war
first
bbc
live
china
kill
cost
â€œ
age
headjet
kong
dead
car
can
syria
obama
rocket
guinea
new
says
fire
band
held
fuel
spy
icij
file
bans
hit
rights
last
kills
police
oil
said
bid
arab
jets
two
court
deal
israeli
end
life
part
killed
home
one
abu
saudi
iran
cairo
test
north
near
calls
child
move
south
son
turkey
amid
nuclear
take
show
jury
ridsdel
2020
hacked
png

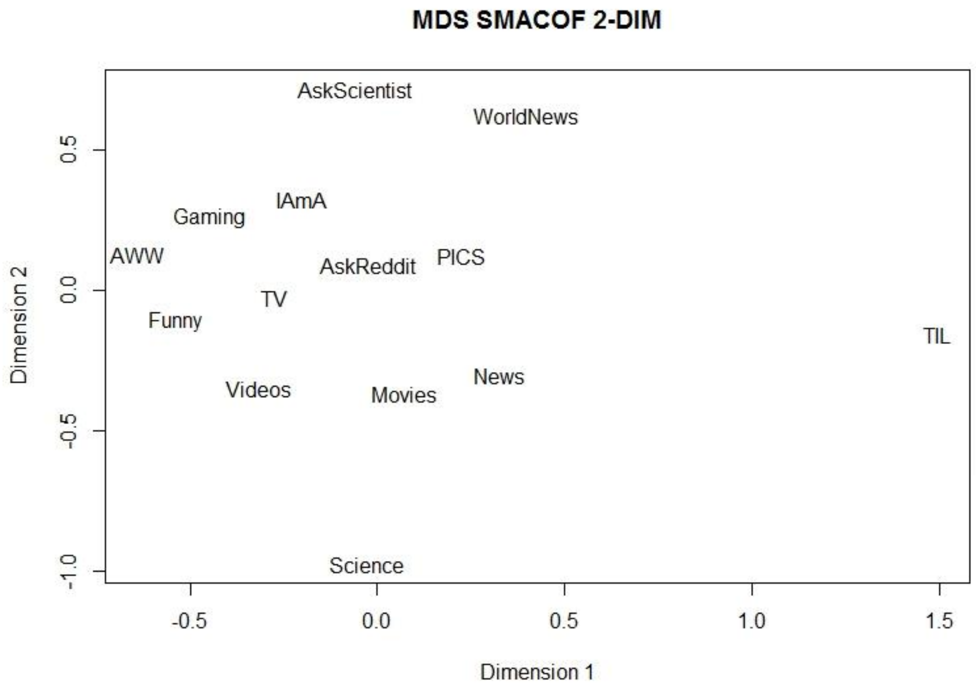

australian
year
near
military
central
sexual
officer
cat
bathroom
want
last
family
couple
water
people
chief
gay
loses
amid
sex
million
bill
woman
fbi
take
finds
help
body
arrested
epa
ohio
dies
law
day
court
can
new
back
dead
one
apple
fire
flint
gets
account
claim
att
black
killed
now
found
man
gun
data
get
find
two
oil
f22
sues
next
first
animals
case
child
city
police
traffic
911
shot
years
possible
armed
car
says
will
australia
faces
220
atsc
obama
san
file
big
hit
baby
china
house
may
state
marijuana
national
held
center
kids
ben
end
francisco
accused
billion
prince
death

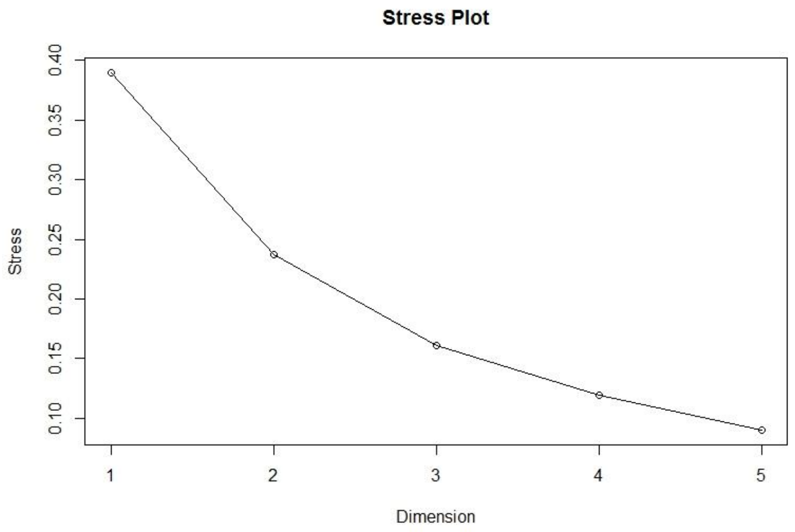

game
souls
dark
games
nintendo
playing
first
new
like
one
video
play
just
best
fallout
xbox
time
gaming
call
good
can
got
need
made
console
mirrors
coming
halo
get
looking
duty
old
rocket
gta
anyone
know
beta
league
2017
march
ever
whats
never
want
love
fps
uncharted
infinite
years
warfare
now
pax
real
favorite
still
100
cool
well
wii
name
2016
last
save
battlefield
release
ps4
think
people
announces
going
mods
playstation
catalyst
available
smash
edge
guy
guys
trailer
looks
dead
zelda
gears
gameplay
creation
community
mario
online
cosplay
launch
windows
make
perfect
finally
super
dont
world
question
top
will
elder
iii
til
ill
ive
nioh
star
war
bros
steam
consoles
space
xpost
someone
overwatch
rainbow
remember
player
played
witcher
ultimate
leaked
next
free
back
buy
close
even
date
age
90s
tomb
without
rust
dubbed
night
plays
really
delayed
may
rain
god
spoilers
place
youve
controller
shadow
spent
graphics
cause
guess
forward
help
months
skyrim
classic
havent
right
guide
modern
days
division
less
see
day
announced
ago
actually
started
max
arcade
getting
little
combat
try
kit
later
fire
tech
else
painting
bought
theft
cant
since
mmo
fox
scrolls
way
based
around
csgo
kill
doom
called
trying
dragon
phantom
portal
steambogdanghita
thing
horror
everyone
planetside
thanks
microsoft
something
multiplayer
says
didnt
character
start
felt
watch
double
half
amazing
trials
use
makes
selling
upcoming
memory
worst
sony
edition
friends
final
legends
blizzard
racing
open
copies
thought
titles
found
hang
part
released
mark
live
boss
leave
come
broken
youre
gave
fun
two
hard
bloodborne
done
sorry
much
isnt
ones
birthday
quick
wireless
update
hardware
appreciate
million
experience
improvements
japanese
thats
grand
great
early
legend
today
talk
hoops
went
crazy
epic
many
awesome
find
pvp
mobile
codenamed
mans
thank
lenses

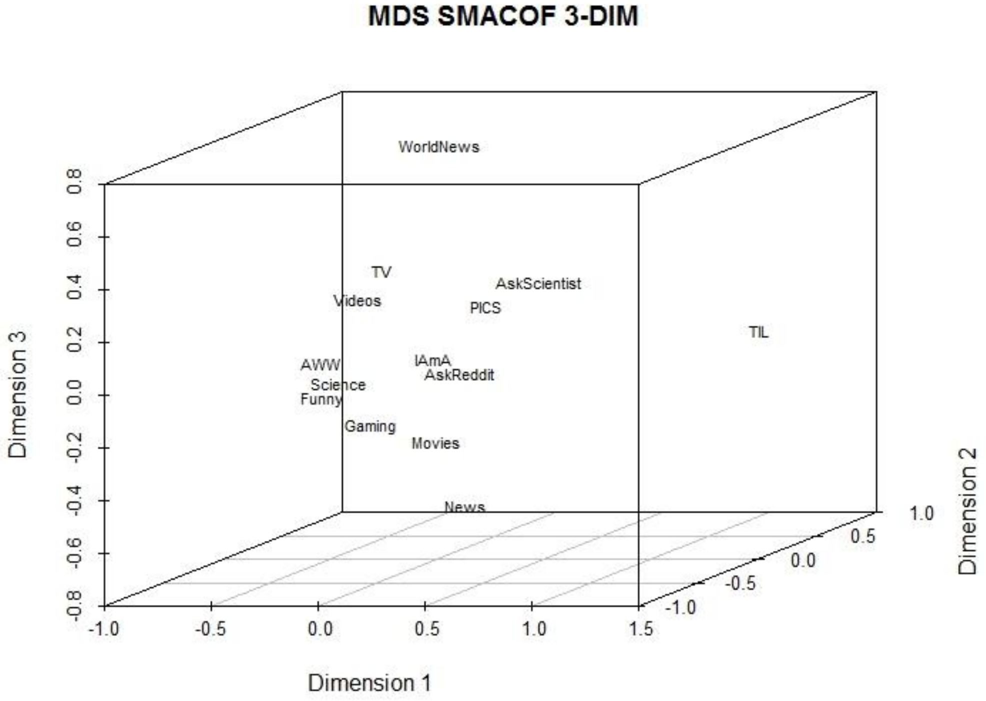

show
season
new
series
game
shows
discussion
netflix
thrones
thread
anyone
spoilersv2blast
star
prince
television
watching
trailer
will
dead
one
best
time
night
episode
episodes
original
favorite
comedy
walking
hbo
good
like
live
kimmy
premiere
tonight
ever
top
â€˜
can
real
bad
david
gets
2016
made
week
whats
find
paul
april
julia
ripa
even
john
greatest
cable
dont
wire
idea
cant
list
teen
else
end
bill
bee
still
air
men
die
set
million
cbs
itâ€™s
due
art
world
know
bbc
trek
saul
post
gear
man
red
today
guy
tribute
actually
thinkgolden
old
â€
katic
return
call
stars
scenes
based
get
talk
havent
women
fans
fillion
film
american
crime
mtv
every
airing
viral
opening
rock
finale
hes
watch
kevin
amc
last
full
park
finds
girl
got
may
join
date
fear
work
life
fan
now
just
host
love
usa
dan
party
pete
back
first
must
fun
silicon
wife
kelly
right
path
schmidt
scene
amazon
better
snl
oliver
became
cast
ive
neil


little
dog
cat
baby
just
meet
new
puppy
like
happy
old
reddit
shes
cute
time
hes
kitten
adopted
get
first
found
can
got
best
loves
home
one
today
name
think
good
dont
guy
pup
friend
looks
cant
pretty
kitty
love
know
two
big
day
say
eyes
play
ever
look
sleeping
likes
boy
face
still
sweet
great
help
belly
girl
work
ready
leave
rescue
sit
hot
guys
turned
will
ago
human
find
favorite
dogs
fox
makes
tiny
cake
now
saw
much
fell
lion
bite
lay
hello
mother
tired
hard
corgi
cats
photo
man
sure
shelter
told
ears
took
next
local
stop
friends
made
car
high
life
keep
loki
last
year
almost
ball
missed
gone
takes
banjo
going
cutest
let
well
jax
want
bed
top
fur
lil
red
knew
sometimes
sits
see
sir

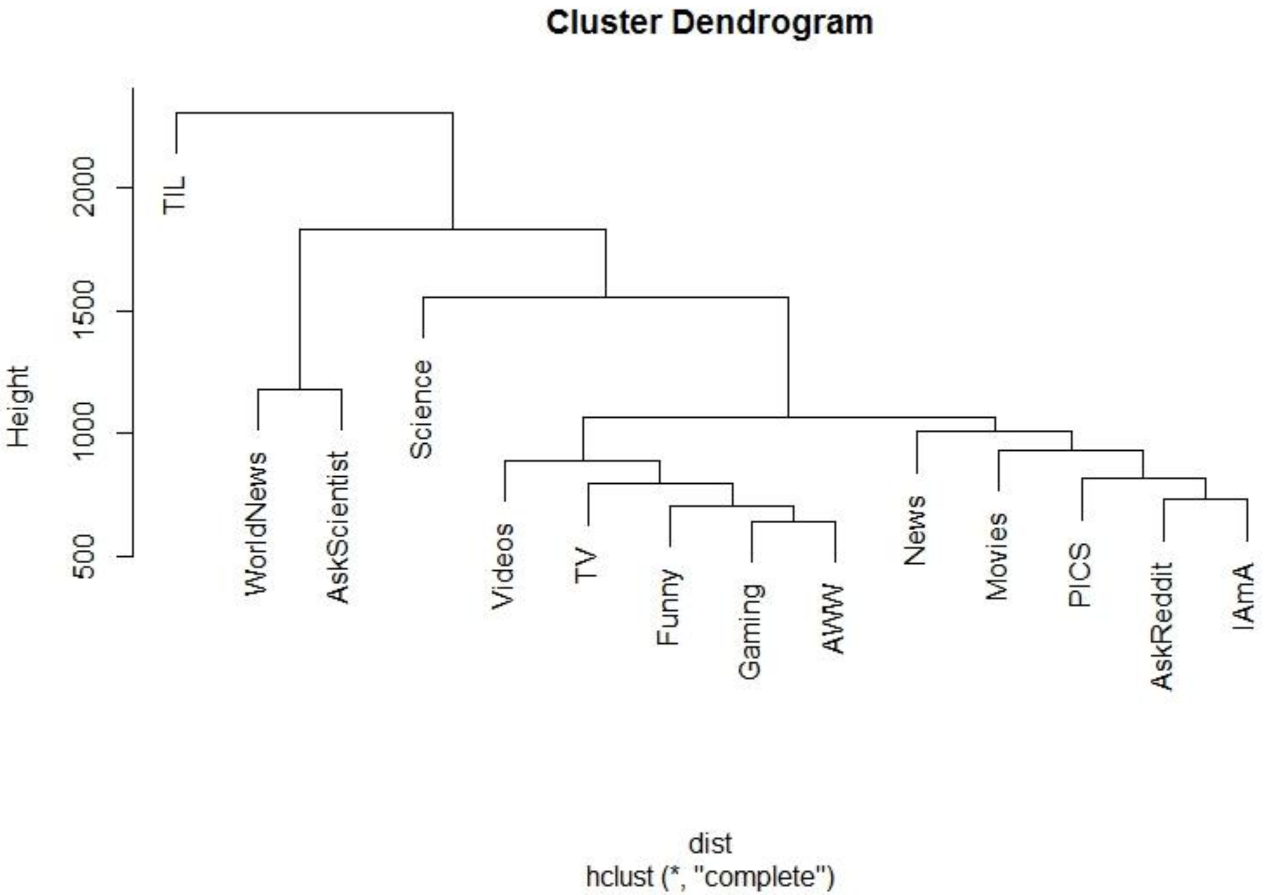
MDS


MDS SMACOF 2-DIM
AskScientist
WorldNews
Gaming
IAmA
AWW
AskReddit
PICS
TV
Funny
TIL
Videos
Movies
News
Science
Dimension 2
0.5
0.0
-0.5
-1.0
-0.5
0.0
0.5
1.0
1.5
Dimension 1

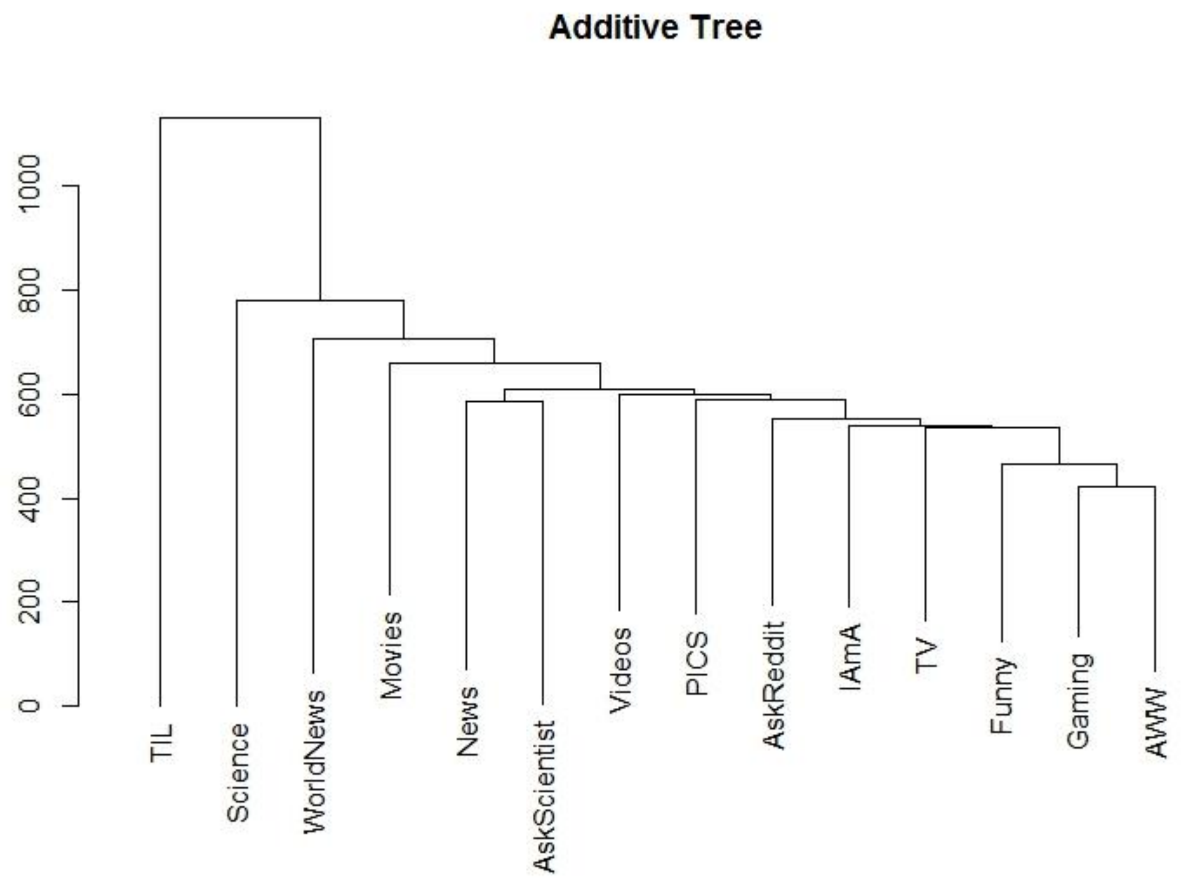

Stress Plot
Stress
0.40
0.35
0.30
0.25
0.20
0.15
0.10
1
2
3
4
5
Dimension


MDS SMACOF 3-DIM
WorldNews
TV
Videos
AskScientist
PICS
TIL
AWW
IAmA
AskReddit
Science
Funny
Gaming
Movies
News
Dimension 3
0.8
0.6
0.4
0.2
0.0
-0.2
-0.4
-0.6
-0.8
-1.0
-0.5
0.0
0.5
1.0
1.5
Dimension 1
1.0
0.5
0.0
-0.5
-1.0
Dimension 2

Hierarchical Clustering



Additive Tree



Conclusions

- We can see that the Additive Tree and Hierarchical Clustering solutions produced sensible clusters
  - Videos/TV
  - Movies/Pics
  - AskReddit/IAmA

- The 2-Dimension MDS results showed:
  - <u>Dimension 1:</u> possibly a measure of number of unique words in subreddit
  - <u>Dimension 2</u>: possibly a measure of overlap between subreddits

Conclusions

- Limitations:
  - No control on number of pages and posts captured
  - Do not have the universe of posts on Reddit
  - Temporal effect
    - Gaming for instance mentions “Dark Souls” a game recently released
    - TV mentions “Game” and “Thrones” a show recently released
  - Don’t understand python & scrapy fluently
- Directions for Future Research
  - Include comments from subreddit posts
  - Consider other measures of similarity between subreddits
  - Study username patterns (SNA)
  - Examine change in word use patterns

Questions?