NYT Phrase Rank Lists

From left to right, below is the list of top 30, middle 30, and bottom 30 single word phrases for the NYT dataset with the suggested parameters of $HIGHLIGHT\_SINGLE = .9$ and $HIGHLIGHT\_MULTI = .5$.

wbc	0.8512157704
minneapolis	0.8472041311
dna	0.8470183261
cska	0.8468338448
cbs	0.8423605647
wellington	0.8421586470
iv	0.8421261359
fbi	0.8420124957
dc	0.8414233784
nba	0.8411841758
fx	0.8405590938
istanbul	0.8399355593
prague	0.8398781518
amc	0.8394182441
rbs	0.8389925043
clarence	0.8389280748
lima	0.8385657743
sacramento	0.8385282580
tripoli	0.8381023545
dublin	0.8379963028
lgbt	0.8379090817
bcs	0.8374214900
usa	0.8372757945
albany	0.8370727411
naples	0.8368999479
lexington	0.8368385291
cal	0.8363055371
milan	0.8362579035
amsterdam	0.8361695801

Table 1: Top 30

intentional	0.5870088101
prodded	0.5870083729
ordinarily	0.5870083729
confront	0.5870008670
soprano	0.5869853623
machinery	0.5869702128
riot	0.5869621683
waiting	0.5869557429
atrocities	0.5869556507
prevail	0.5869546957
here's	0.5869530878
interrupted	0.5869266819
smokers	0.5869212730
disruption	0.5868902329
originally	0.5868889273
slam	0.5868875053
poisoning	0.5868756404
desperately	0.5868640156
siege	0.5868636363
widow	0.5868611921
buy	0.5868535537
recreational	0.5868499478
hosting	0.5868189940
unusually	0.5867760126
wheelchairs	0.5867586481
torque	0.5867586481
strikes	0.5867349929
ribs	0.5867307076
endless	0.5867247648

Table 2: Mid 30

172	0.2744979322
06	0.2742260498
332	0.2739979322
299	0.2739979322
285	0.2739979322
139	0.2738766169
126	0.2734599502
214	0.2734599502
03	0.2728980737
320	0.2728980737
325	0.2726260498
157	0.2724599502
02	0.2723790519
375	0.2723790519
containing	0.2721296301
158	0.2719718549
380	0.2719718549
116	0.2710698596
specified	0.2696099502
124	0.2694850748
145	0.2691612652
280	0.2685555739
107	0.2683433052
111	0.2678667286
112	0.2673443406
114	0.2667911009
sincere	0.2666137439
ought	0.2653443406
102	0.2644190375

Table 3: Bottom 30

These lists above make sense because the top 30 are all common acronyms and proper nouns, the mid 30 are frequent and meaningful nouns and verbs, and the bottom 30 are mostly random numbers.

From left to right, below is the list of top 30, middle 30, and bottom 30 multi word phrases for the NYT dataset with the suggested parameters of $HIGHLIGHT\_SINGLE = .9$ and $HIGHLIGHT\_MULTI = .5$. These lists below for multi word phrases are intuitive because the top 30 are names of people and organizations, the middle 30 are common phrases in speech, and the bottom 30 are prepositions and transitional phrases.

kyrie irving	0.9859670300
adrian peterson	0.9858567268
justin bieber	0.9855607233
aston villa	0.9852497353
richie incognito	0.9851547916
marshawn lynch	0.9848564661
fidel castro	0.9845802860
napa valley	0.9845547987
graeme swann	0.9843076584
nick saban	0.9842913079
rafael nadal	0.9841839528
henrik lundqvist	0.9841120842
justin timberlake	0.9840859622
gareth bale	0.9840295465
tiger woods	0.9839738113
jamaal charles	0.9839603943
sears holdings	0.9839533040
dwight howard	0.9838893572
alan mulally	0.9838876767
derek stepan	0.9838231328
walt disney	0.9837392413
nick foles	0.9836625843
michel djotodia	0.9836394567
brook lopez	0.9835856279
hurricane katrina	0.9835716916
tim duncan	0.9835537588
ubs ag	0.9835314120
sebastian vettel	0.9834178661
janet yellen	0.9833784342
golan heights	0.9833098495

Table 4: Top 30

handed down	0.1278334721
stand up	0.1278236174
ranges from	0.1278187086
from public life	0.1278089004
of comcast corp	0.1277916275
wars in iraq and	0.1277562480
doesn't include	0.1277504663
energy bill	0.1277337884
as south africa's	0.1277202886
well aware	0.1277180673
35 minutes	0.1277144797
in tax breaks	0.1277116959
israeli military said	0.1276519684
get hurt	0.1276420641
gave me	0.1276355675
of political prisoners	0.1276305578
announced tuesday that	0.1275690246
in june 2012	0.1275658157
dominican republic and	0.1275312384
slightly better	0.1275279922
political system	0.1275074171
to enlist	0.1275043352
left behind	0.1274784624
last four years	0.1274718913
still alive	0.1274561571
without power	0.1274171262
reunite with	0.1273897374
former employees	0.1273891631
to be	0.1273639799
60 years	0.1273436734

Table 5: Mid 30

one hand and	0.0075959662
of months of	0.0075959662
of pennsylvania and	0.0075959662
to syria to	0.0075959662
from home and	0.0075959662
three games with	0.0075959662
of meetings with	0.0075959662
for consumers to	0.0075959662
a lot to	0.0075959662
to happen in	0.0075959662
of murder in	0.0075959662
in court on	0.0075959662
for oil and	0.0075959662
nine years in	0.0075959662
to sign with	0.0075959662
to grant a	0.0075959662
to file for	0.0075959662
to die in	0.0075959662
in mind that	0.0075959662
of violence in the	0.0075959662
of millions of people	0.0075959662
for sale in	0.0075959662
to study in	0.0075959662
to turn this	0.0075959662
in court that	0.0075959662
of billions of	0.0075959662
to reporters at	0.0075959662
of money in	0.0075959662
of state for	0.0075959662
said she	0.0070537286

Table 6: Bottom 30

Yelp Phrase Rank Lists

From left to right, the lists on the following page are of the top 30, middle 30, and bottom 30 single word phrases for the NYT dataset with the suggested parameters of $HIGHLIGHT\_SINGLE = .9$ and $HIGHLIGHT\_MULTI = .5$.
These lists make sense because the top 30 are mostly all names of foods and drinks, the middle 30 are common nouns and verbs associated with restaurants and services, and the bottom 30 are mostly prepositions that don't have much meaning.

The lists on page 4 are the top, middle and bottom 30 multi word phrases in the Yelp dataset. Again these lists are intuitive because the top 30 are organization names or names of specific foods, the middle 30 are common foods and speech phrases, and the bottom 30 are more prepositions and transitions.

trail	0.9283774739
ale	0.9279517675
sangria	0.9251320512
yelper	0.9231119693
hangover	0.9215918356
eggplant	0.9210159397
gold	0.9208643534
library	0.9205310201
stew	0.9190822358
church	0.9189934626
latte	0.9187752054
cherry	0.9184212028
espresso	0.9181488972
yelpers	0.9176200685
vegas	0.9171503716
viet	0.9168684378
brie	0.9168375948
mole	0.9162541174
lifetime	0.9160440823
flatbread	0.9158635588
parmesan	0.9155271740
brewery	0.9152714963
ribeye	0.9152654524
zin	0.9152449625
meatball	0.9151991072
fitness	0.9151448717
farms	0.9150685900
coach	0.9150399127
ranch	0.9150369352

Table 7: Top 30

kinda	0.7716520296
basement	0.7716468228
fricken	0.7716241467
posts	0.7716208485
caliber	0.7716175362
diverse	0.7716133299
dimsum	0.7716003917
famed	0.7715747751
cheezy	0.7715329133
preschool	0.7715329133
isolated	0.7715109722
hope	0.7714825538
traditionally	0.7714700443
haircut	0.7714621814
lengthy	0.7714562774
attendance	0.7714516128
resteraunt	0.7714475621
pigeons	0.7714475621
app’s	0.7714475621
ashes	0.7714475621
leadership	0.7714475621
issue	0.7714269907
43rd	0.7714168844
ubiquitous	0.7713882299
softener	0.7713855170
chipper	0.7713338390
wage	0.7713290252
sites	0.7713123338
days	0.7713087739

Table 8: Mid 30

amongst	0.3745894890
06	0.3734362821
werent	0.3732478184
been	0.3732182178
which	0.3727403290
containing	0.3723831115
09	0.3723831115
yourself	0.3710446346
me	0.3708902248
awfully	0.3707345023
02	0.3707345023
becomes	0.3707275842
can	0.3704885490
merely	0.3704784044
could	0.3700245072
05	0.3674529488
be	0.3664271359
gives	0.3662908717
themselves	0.3659551358
him	0.3632958684
have	0.3631863165
ourselves	0.3618827637
is	0.3612317711
has	0.3610493581
them	0.3608067799
are	0.3587614793
were	0.3577612489
was	0.3570231537
myself	0.3561846821

Table 9: Bottom 30

jamba juice	0.9855316025
grand marnier	0.9837130519
hyatt regency	0.9834038475
botanical gardens	0.9819158655
jimmy johns	0.9815106738
hobby lobby	0.9798741782
palo verde	0.9795939574
del rey	0.9795771212
orange blossom	0.9792773294
peter piper	0.9791724839
taliesin west	0.9789654644
cheesecake factory	0.9788442154
cabernet sauvignon	0.9786717397
del mar	0.9785584259
english muffin	0.9782791894
chile relleno	0.9782021344
el paso	0.9781792604
pollo asado	0.9781085274
daily dose	0.9780325640
red robin	0.9780062009
squaw peak	0.9779615017
union hills	0.9779233649
upper crust	0.9777963502
delhi palace	0.9777607192
carl's jr	0.9777348202
johnny rockets	0.9776970828
tater tots	0.9776615624
michael mina	0.9776529286
ann taylor	0.9776391505
dos equis	0.9775953971

Table 10: Top 30

the wedge salad	0.0885694854
sitting on top of	0.0885663980
the old days	0.0885654616
a bad choice	0.0885610444
bent on	0.0885559050
especially since	0.0885509178
a shopping center	0.0885478251
the mediterranean	0.0885436086
with arugula	0.0885353958
flow of	0.0885310512
an incredible	0.0885294980
that point	0.0885292634
driving into	0.0885289634
the easiest	0.0885265352
starting to get	0.0885263643
particularly care for	0.0885217908
to order drinks	0.0885201542
to dance	0.0885161429
by weight	0.0885105557
a medium	0.0885099993
diagnosed with	0.0885098387
away from	0.0885085989
at ikea	0.0885037011
a nice outdoor patio	0.0884998262
does not matter	0.0884994810
all night	0.0884984343
for sunday brunch	0.0884845028
a yummy	0.0884817247
every other	0.0884773489
an enthusiastic	0.0884728286

Table 11: Mid 30

to venture to	0.0067325702
a search for	0.0067175661
to go again and	0.0066756286
some research on	0.0065839591
got seated at	0.0065818086
no issues with	0.0065001504
and it's fun to	0.0064707003
1 star because	0.0064616889
and just wanted to	0.0064463948
the nail on	0.0064420109
and can't wait to go	0.0064151448
and very easy to	0.0063960832
people complain about	0.0063864553
i'm working on	0.0061836335
usually sit at	0.0061836335
got tired of	0.0061514703
actually care about	0.0061280780
very reminiscent of	0.0061191174
so i'll keep	0.0061012530
the peak of	0.0060635619
i've worked in	0.0059302285
a touch more	0.0059302285
to relax with	0.0059064550
a block of	0.0058506258
to call ahead and	0.0058439550
to kick back	0.0057881258
the only drawback to	0.0057675661
a cheeseburger with	0.0057606217
a scale of	0.0057047924
any combination of	0.0049302285

Table 12: Bottom 30

Phrasal Segmentation Metrics

The below table shows the number of unique qualified phrases and average number of phrases per sentence for each dataset for $HIGHLIGHT\_SINGLE = .9$ and $HIGHLIGHT\_MULTI = .5$. These metrics include both single and multi word phrases.

	NYT	Yelp
Avg per sentence	.94149	.780041
Total unique phrases	65552	41988

Table 13: Phrase metrics

Since NYT is professionally written and edited, I would expect a natural language processor to more easily identify phrases, which is reflected in the data. Yelp reviews contain a lot of slang and bad grammar, which might make the task of identifying phrases more difficult, yielding fewer phrases identified overall.

NYT Phrase Clusters

Below are samples of six clusters obtained from using the k means algorithm with 75 centers on the NYT dataset. I only clustered phrases returned by AutoPhrase and used each phrase's vector from word2vec to do the clustering. I've labeled the concept each cluster is describing in the table captions.

government
state
country
public
city
capital
Syria
corruption
politics
army
Iraq
Muslim
Islamic
Islamist
wave
Assad
ethnic
suicide
Libya
toll

Table 14: Middle East Politics

revenue
range
auction
_billion_euros_
_million_euros_
_million_pounds_
euros
acres
_billion_pounds_
lease
temperature
_square_feet_
_ticket_sales_
_trillion_yen_
_miles_per_hour_
_million_shares_
_contemporary_art_
_square_foot_
_million_Americans_
_million_viewers_

Table 15: Quantities

federal
health
Obama
tax
website
Americans
insurance
cover
bills
_federal_government_
_health_insurance_
insurers
Medicaid
Medicare
_Affordable_Care_Act_
_minimum_wage_
HealthCare.gov
Obamacare
premiums
_Social_Security_

Table 16: Obamacare

_U.S._
_United_States_
American
China
Chinese
region
Israel
Japan
Western
Japanese
Beijing
_North_Korea_
Saudi
Iranian
_South_Korea_
Pakistan
Turkey
_Middle_East_
Egypt
Arab

Table 17: State Entities

Republican
Senate
House
Republicans
_White_House_
Democrats
Democratic
Democrat
Senator
_President_Barack_Obama_
_President_Obama_
congressional
legislative
Capitol
conservatives
liberal
senator
Christie
_Tea_Party_
Legislature

Table 18: American Politics

history
football
manager
star
college
sports
English
soccer
sport
basketball
professional
baseball
golf
hockey
tennis
_national_team_
cricket
Argentine
athlete
_college_football_

Table 19: Sports

Yelp Phrase Clusters

On this and the following page I've used the same process as described for NYT on the Yelp dataset.

food
Food
_pleasantly_surprised_
_dining_experience_
_blown_away_
Atmosphere
_poor_service_
_slow_service_
Taste
_Customer_service_
Average
_mediocre_food_
_bottom_line_
_entire_experience_
_noise_level_
non-existent
_pretty_slow_
_wait_times_
_terrible_service_
_Wait_staff_

Table 20: Food Service/Experience

pizza
salad
sandwich
sandwiches
salads
pasta
hummus
Pizza
bruschetta
pita
wrap
mushroom
soups
pepperoni
_thin_crust_
gyro
_grilled_cheese_
italian
falafel
panini

Table 21: Italian/Greek Food

_ok_
average
_OK_
_Ok_
basic
_pretty_decent_
stellar
_bar_food_
_pretty_tasty_
_pretty_darn_
_pretty_damn_
sub-par
_pretty_awesome_
_pretty_standard_
_pretty_bad_
_looked_pretty_
downhill
_sub_par_
_below_average_
_pretty_solid_

Table 22: Descriptions

_Orange_Table_
_La_Grande_Orange_
_Two_Hippies_
_Z_Pizza_
_5th_and_Wine_
_Cheba_Hut_
_Roaring_Fork_
_Local_Breeze_
_Mellow_Mushroom_
Bagels
Chino
_Buffalo_Wild_Wings_
_Tilted_Kilt_
_Melting_Pot_
_PF_Chang's_
_Al's_
_BBQ_place_
_original_location_
_Jimmy_Johns_
Ray's

Table 23: Popular Restaurants

tex-mex
Cambodian
Taiwanese
_real_Mexican_food_
family-owned
family-run
_plate_lunch_
barrio
_highly_rated_
_lunch_option_
locally-owned
mom-and-pop
_family_owned_
_fast_casual_
mom _pop_
_authentic_Chinese_
_authentic_Japanese_
Viet
_Asian_cuisine_
_authentic_Chinese_food_

Table 24: Authentic Food

Hallmark
_Harkins_theater_
_Bath_and_Body_Works_
big-box
_online_shopping_
_Williams_Sonoma_
_Half_Price_
_Stein_Mart_
Kombucha
Brookstone
_Ted_Baker_
kayaking
randomness
_Grocery_store_
In-N-Out's
Axis/_Radius_
_Foot_Locker_
_Disney_Store_
non-fiction
_Weight_Watchers_

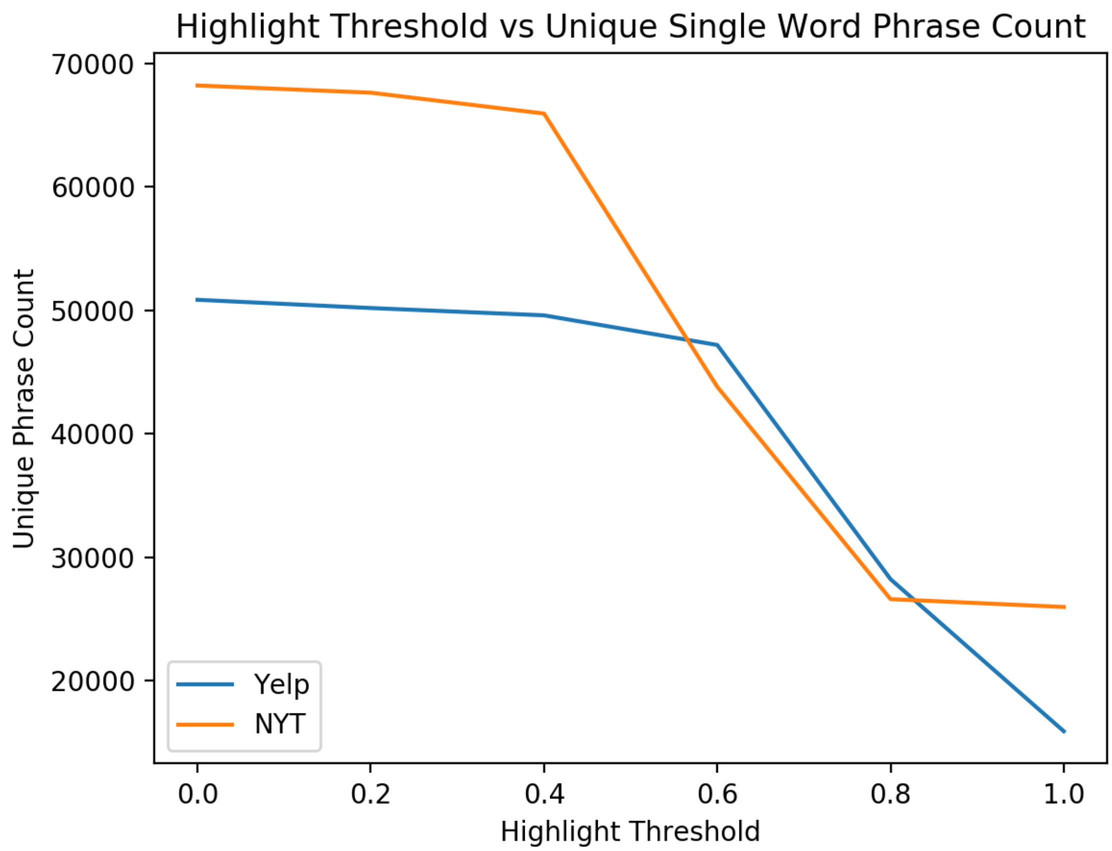
Table 25: Grocery/Department Stores

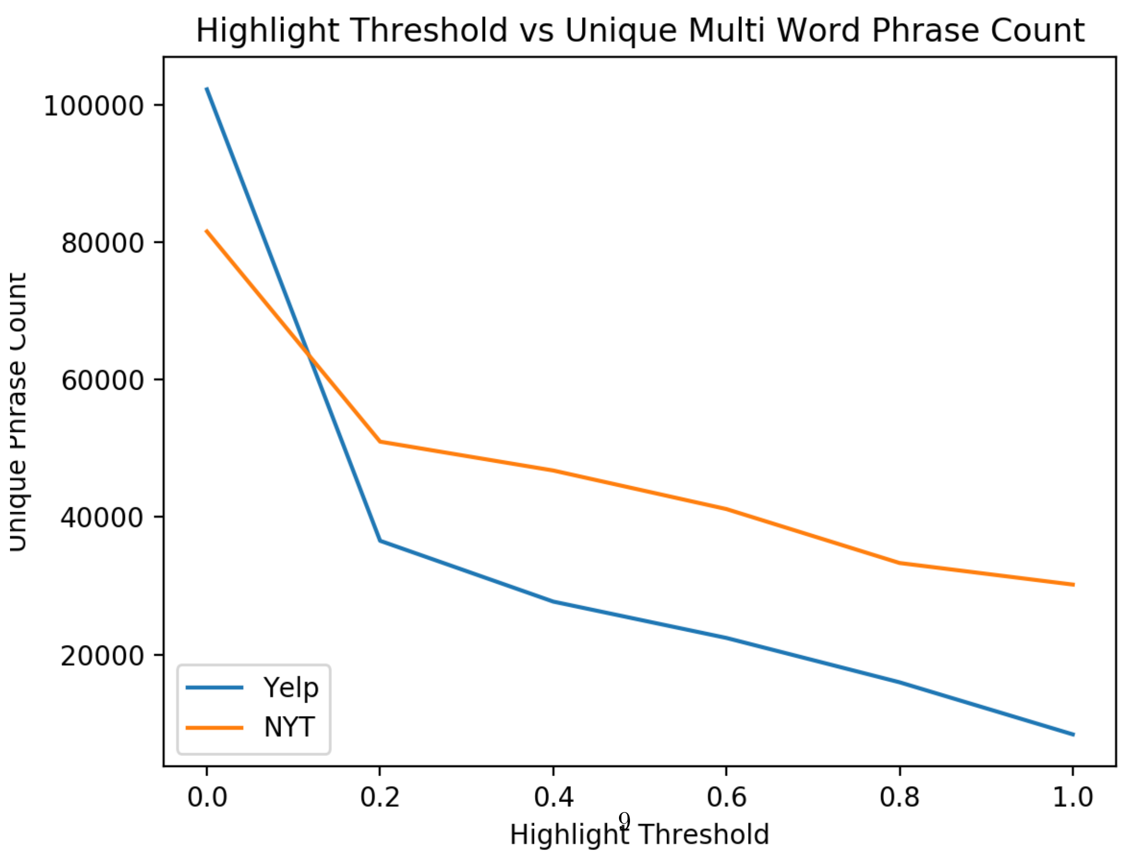
AutoPhrase Parameters

The images below show plots of the number of unique phrases returned by AutoPhrase with different highlight thresholds. The top figure shows the number of unique single word phrases and the bottom figure shows the number of unique multi word phrases.

There is a very clear inverse relationship between highlight threshold and number of unique phrases. The shapes of the graphs are very similar for both datasets with the NYT dataset yielding more phrases for both single and multi word, except with a very low threshold in the multi word curve. In the single word curve they both plateau for a while before dropping for a threshold in the .5-.6 range. In the multi word curve they both drop very steeply as the threshold rises from 0 to .2 before leveling off a bit.

This shows that AutoPhrase detects more phrases in the NYT dataset, presumably because it is written with correct grammar and less slang than Yelp.


Highlight Threshold vs Unique Single Word Phrase Count
Unique Phrase Count
70000
60000
50000
40000
30000
20000
0.0
0.2
0.4
0.6
0.8
1.0
Highlight Threshold
Yelp
NYT


Highlight Threshold vs Unique Multi Word Phrase Count
Unique Phrase Count
100000
80000
60000
40000
20000
0.0
0.2
0.4
0.6
0.8
1.0
Highlight Threshold
Yelp
NYT

NYT Clustering Parameters

Below I will show samples from 3 different clustering results on the NYT dataset. Each row is from the same clustering iteration and has the same number of centers, the centers are stated in the captions.
The first row of tables shows k=5, second row shows k=20, and the third row shows k=40. You can see the clusters become more coherent as you read through each level. For example, at k=5 the Justice/Politics cluster mixes police, family and politics while a similar cluster at k=20 (U.S Government/Foreign Entities) provides more closely related phrases related to the U.S and foreign policy. Furthermore, there is a k=40 cluster (U.S Government) that narrows this down to only the U.S.

government
_U.S_
state
_United_States_
country
long
public
city
American
political

Table 26: k=5: Government/Politics

game
season
play
games
lead
led
lost
shot
free
history

Table 27: k=5: Sports/Games

police
family
law
court
federal
death
party
Republican
South
President

Table 28: k=5: Justice/Politics

_United_States_
president
China
European
President
Iran
Europe
Russia
trade
France

Table 29: k=20: U.S Government/Foreign Entities

city
military
security
French
British
violence
war
corruption
_United_Nations_
Afghanistan

Table 30: k=20: Military/War

company
business
director
research
firm
_chief_executive_
general
organization
Department
Center

Table 31: k=20: Business

long
high
fell
average
rose
short
stock
lower
unemployment
reading

Table 32: k=40: Economy/Stocks

Republican
Senate
House
Republicans
Congress
_White_House_
Texas
Democrats
Democratic
Democrat

Table 33: k=40: U.S Government

England
captain
Chelsea
_New_Zealand_
_Premier_League_
_Champions_League_
Arsenal
Liverpool
Barcelona
squad

Table 34: k=40: Soccer

Yelp Clustering Parameters

Below I will show samples from 3 different clustering results on the Yelp dataset. Each row is from the same clustering iteration and has the same number of centers, the centers are stated in the captions.
The first row of tables shows k=5, second row shows k=20, and the third row shows k=40. You can see the clusters become more coherent as you read through each level. For example, at k=5 there is a Miscellaneous cluster in which the phrases don't seem to be related, while the clusters at k=20 are more conceptually similar. Furthermore, there is a k=40 cluster (Decorations) that is much more specific than any cluster in k=20 or k=5.

food
restaurant
stars
sushi
_happy_hour_
star
Thai
average
Food
Chicken

Table 35: k=5: Restaurants/Food

order
long
free
water
music
car
money
glass
manager
game

Table 36: k=5: Miscellaneous

nice
friendly
bit
atmosphere
high
sat
Love
variety
park
Nice

Table 37: k=5: Descriptions

food
restaurant
stars
sushi
star
Thai
average
Food
Mexican
Italian

Table 38: k=20: Food Types

sandwich
fries
shrimp
steak
sandwiches
appetizer
salads
pasta
egg
eggs

Table 39: k=20: Food Items

order
friendly
friends
family
sat
manager
party
bartender
talk
chef

Table 40: k=20: Food Service

friendly
manager
_customer_service_
bartender
talk
company
chef
professional
management
_super_friendly_

Table 41: k=40: Customer Service

sandwich
fries
bread
steak
sandwiches
bacon
salads
eggs
onion
sausage

Table 42: k=40: Food Items

art
center
brand
_dining_room_
furniture
_bar_area_
private
stock
bakery
design

Table 43: k=40: Decorations