Analyzing Reddit Comments With NLP

Nick Horton

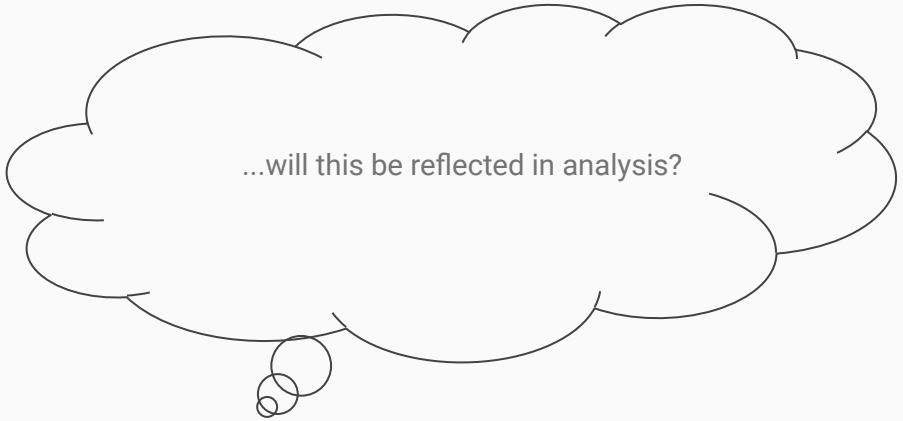
Project Focus

Lighthearted, non-serious:

- memes, funny, wallstreetbets

Serious, factual:

- science, psychology, worldnews



Data Acquisition

- Used PRAW (Python Reddit API Wrapper) to download comments from Reddit

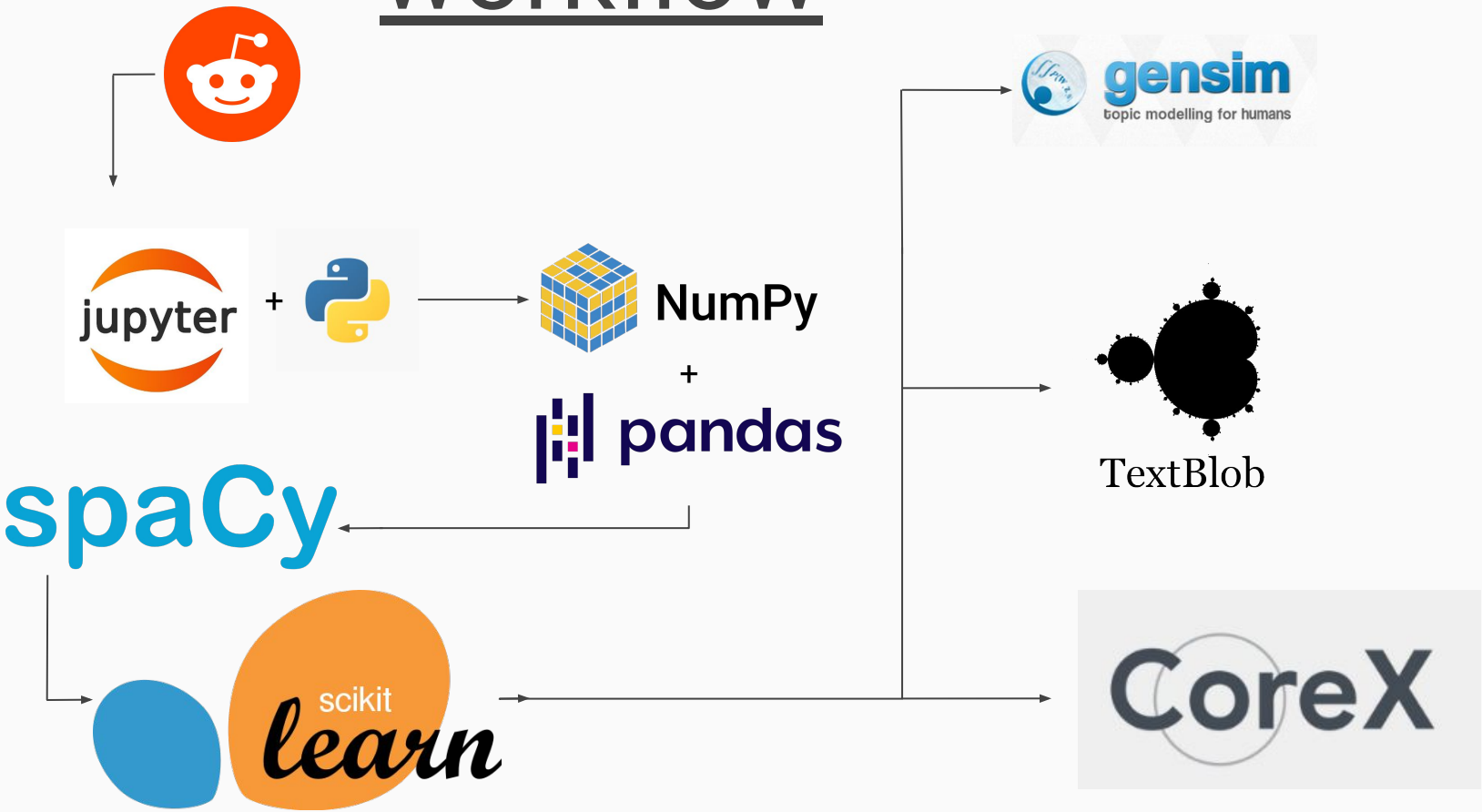
- High proportion of comments were from Askreddit

EDA

More serious subreddits have longer comments on average, while meme subreddits have the shortest.

	subreddit	total_words	total_comments	words_per_comment
12	writingprompts	268692	4026	66.74
11	cscareerquestions	326522	12703	25.70
9	psychology	32283	1482	21.78
1	science	979181	59966	16.33
4	worldnews	227072	15265	14.88
10	politics	534476	38821	13.77
3	askreddit	1160818	101141	11.48
8	nosleep	76217	6683	11.40
2	ama	18493	1885	9.81
5	funny	47841	6659	7.18
0	wsb	243668	36719	6.64
6	dankmemes	25725	4077	6.31
7	memes	37814	7332	5.16

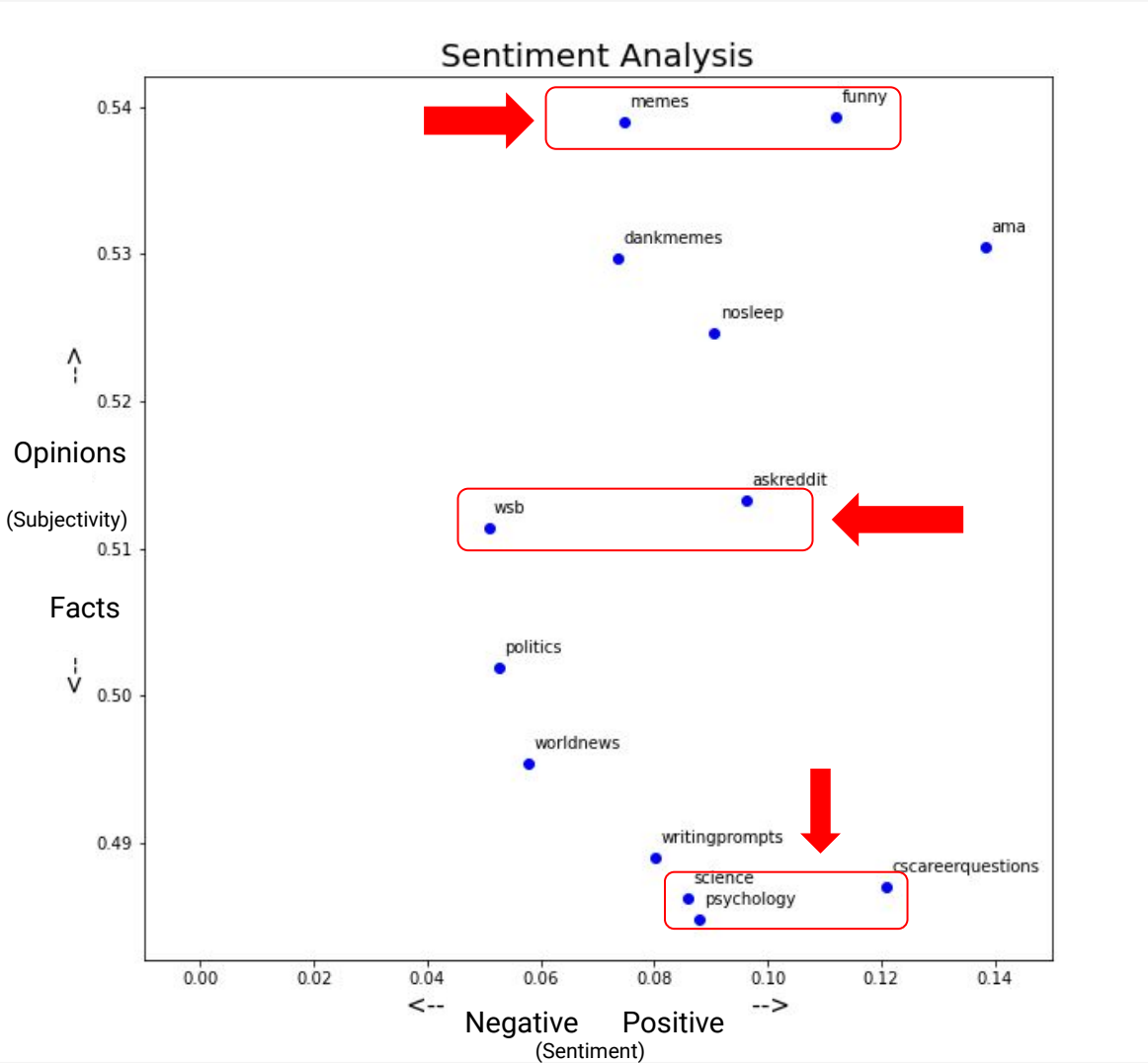
workflow


jupyter
+
NumPy
+
pandas
spaCy
scikit
learn
gensim
topic modelling for humans
TextBlob
CoreX

Word2Vec Sentiment Analysis

Sentiment analysis worked mostly as expected:

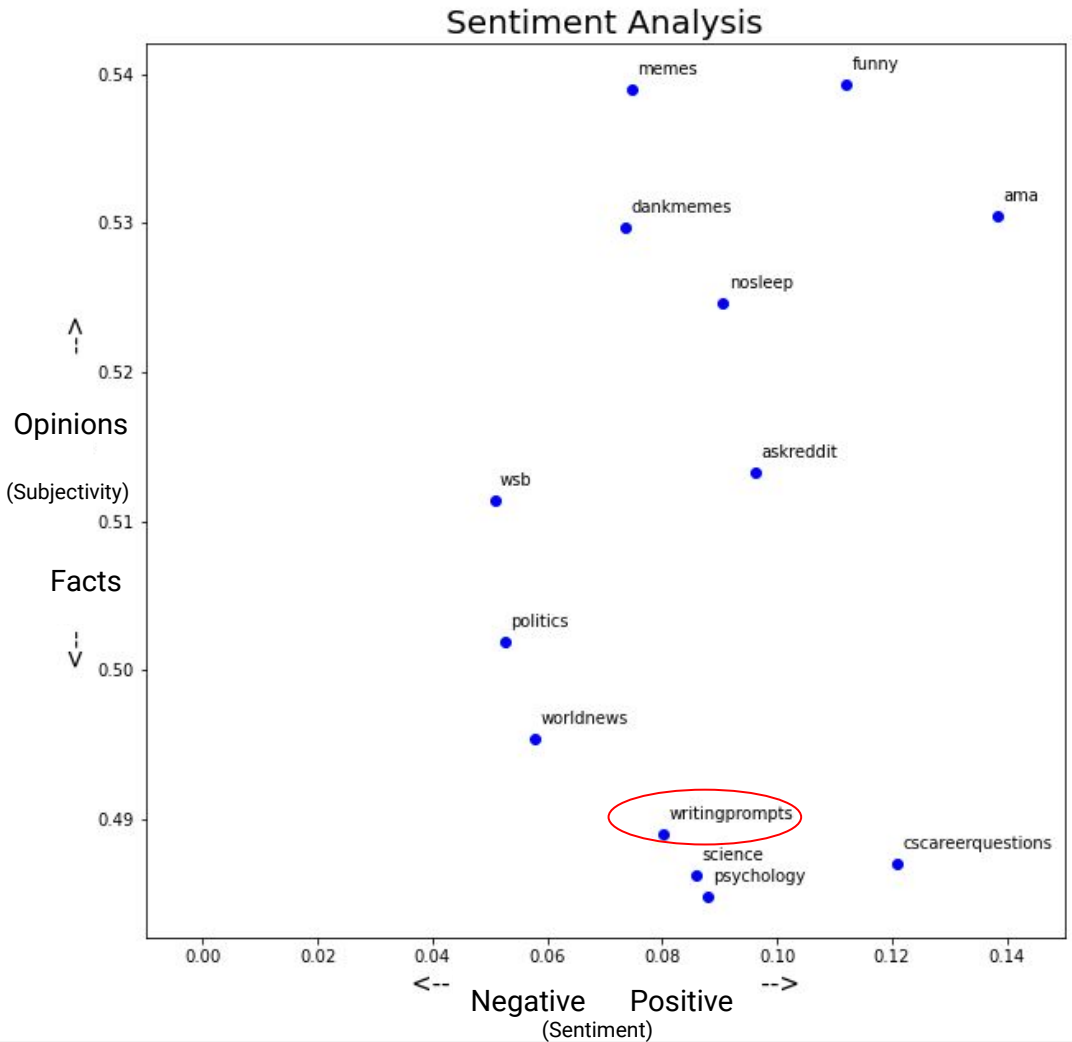
- funny, memes → subjective
- science, cscareerquestions → objective
- askreddit, wallstreetbets → neutral



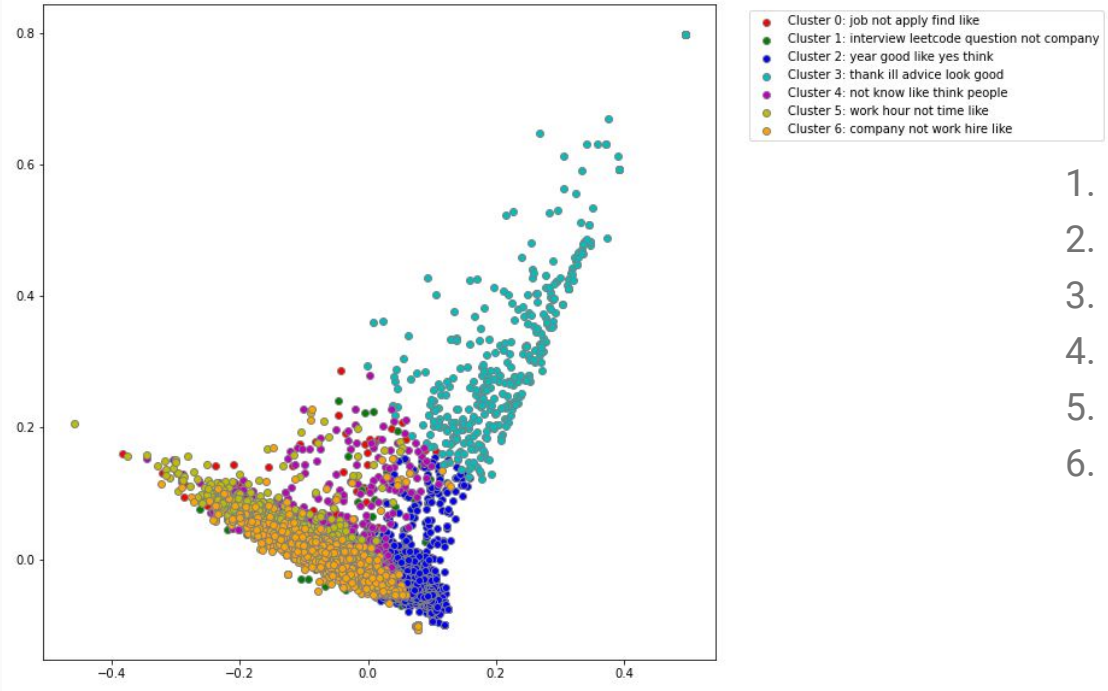
Word2Vec Sentiment Analysis

Unexpected:

- Writingprompts → factual?
  actually, is completely fictional
- TextBlob can't differentiate fictional stories from factual retellings of events



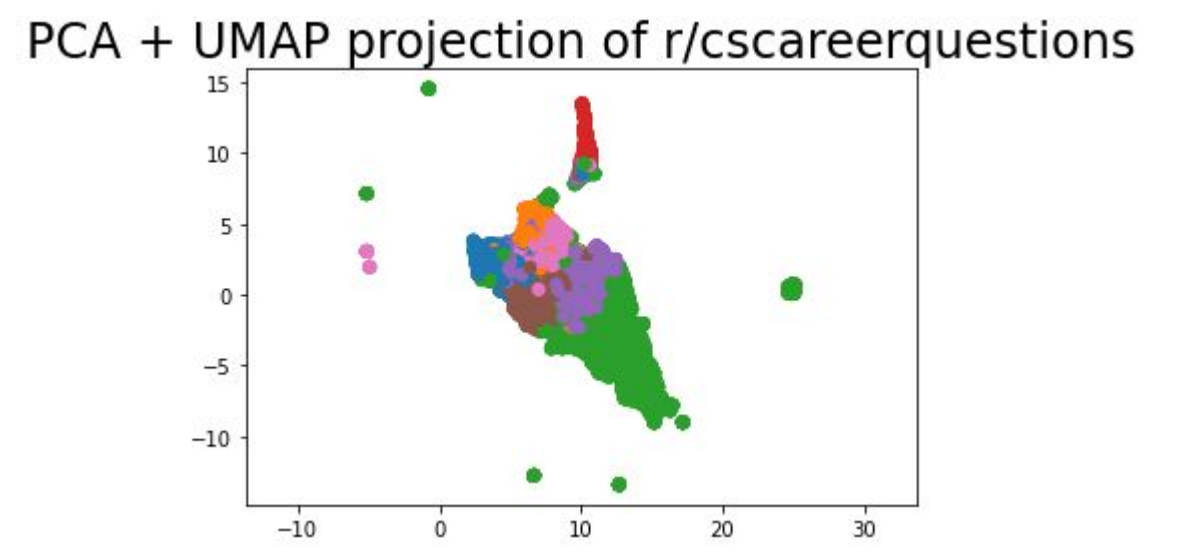
K-Means: r/cscareerquestions



Topics n=7

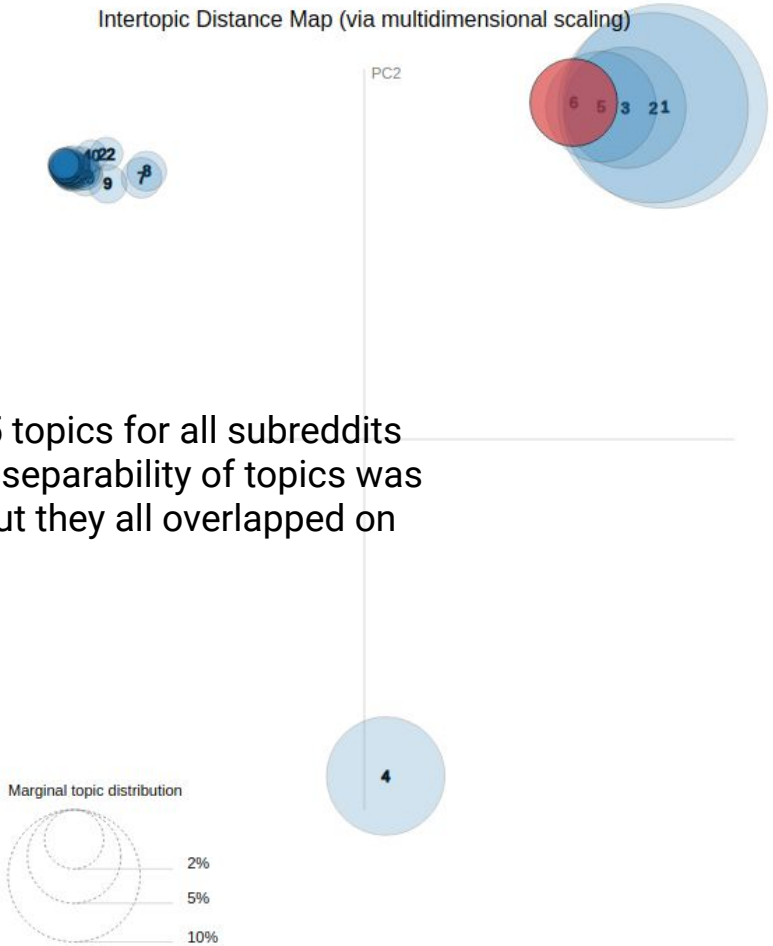
1. job not apply find like
2. interview leetcode question not company
3. thank ill advice look good
4. not know like think people
5. work hour not time like
6. company not work hire like

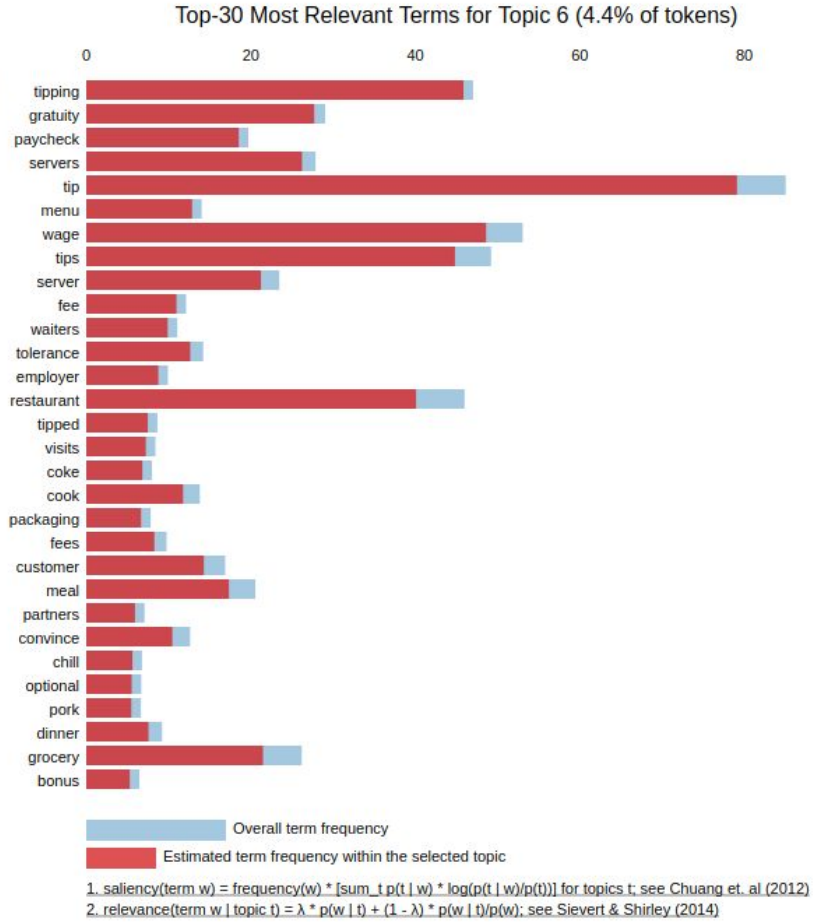
Dimensionality Reduction!



The topics look much less arbitrary after PCA + UMAP - clusters are clearly defined

When using 75 topics for all subreddits combined, the separability of topics was much better, but they all overlapped on the plot

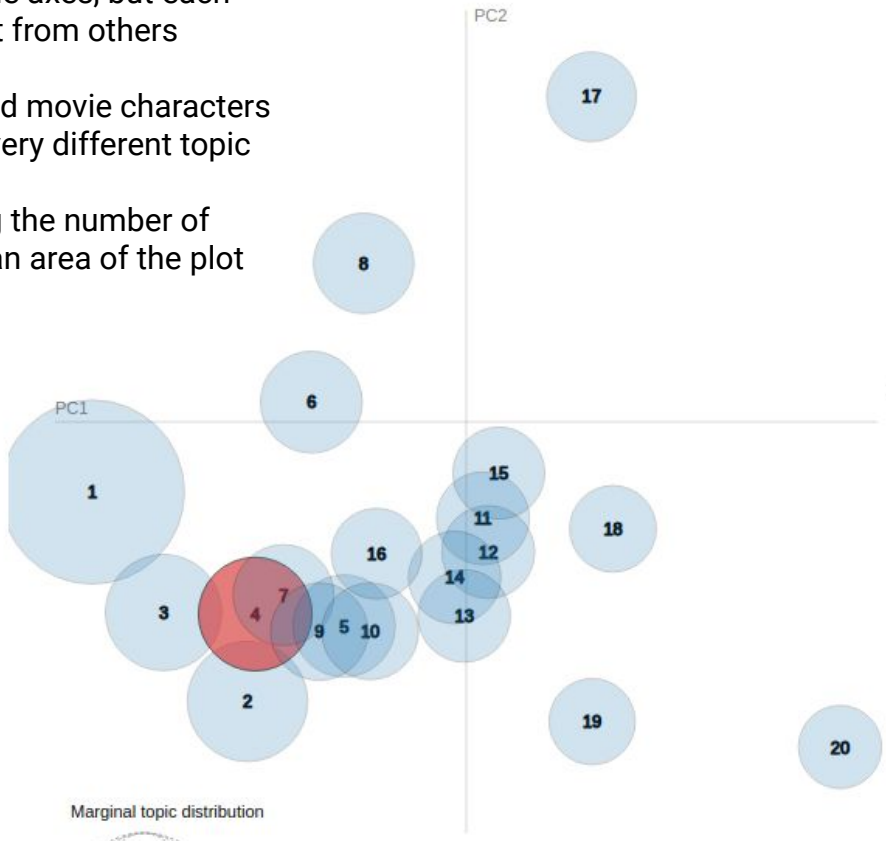


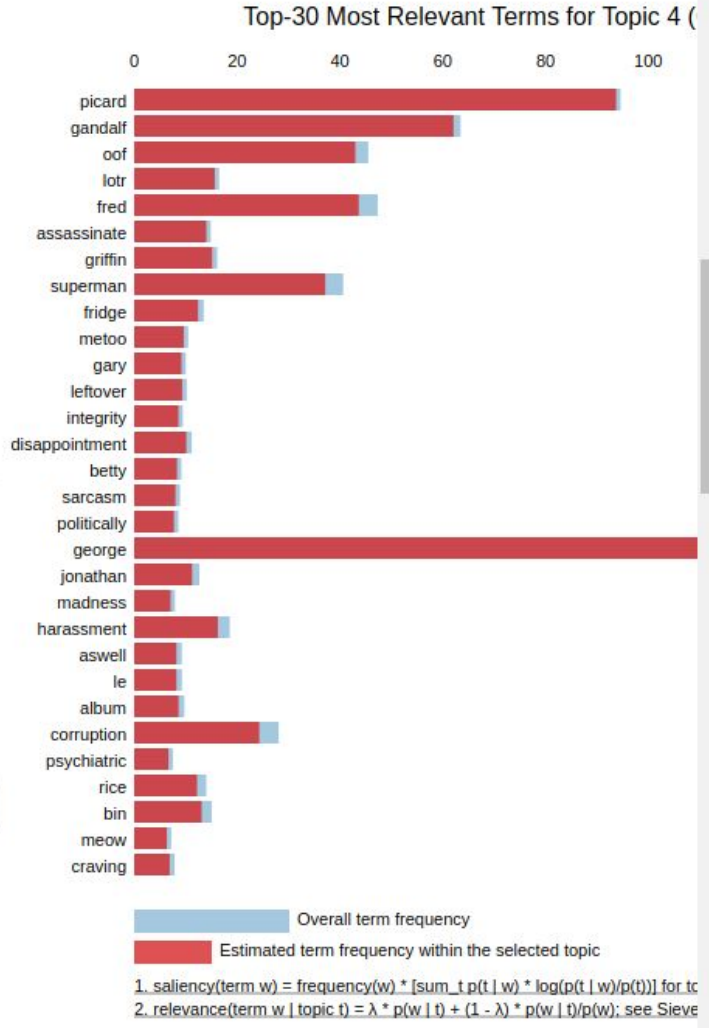


Reducing the number of topics spread them out more over the axes, but each topic was less distinct from others

In this instance, TV and movie characters appeared in almost every different topic

It's likely that reducing the number of topics "zooms in" on an area of the plot





Take Away

Dimensionality Reduction is a useful tool for topic modeling.

Comments across different subreddits have significant variation in distribution of topics, number of words per comment, and user activity.

Future Work

- Increase scope - find other highly trafficked subreddits with widely varying content

- Implement a faster, homemade K-Means algorithm to speed up workflow, allowing for more experimentation

- Make Flask app for playing around with pyLDAvis

Thanks!

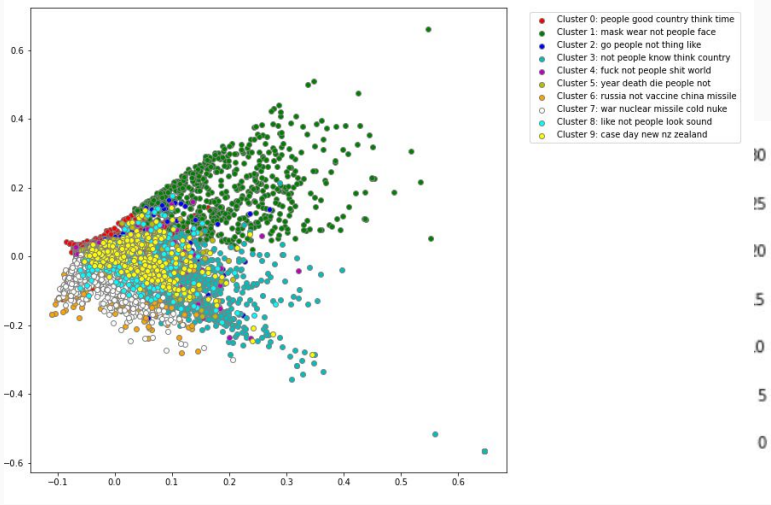
Contact me:

nhorton04@gmail.com
www.github.com/nhorton04

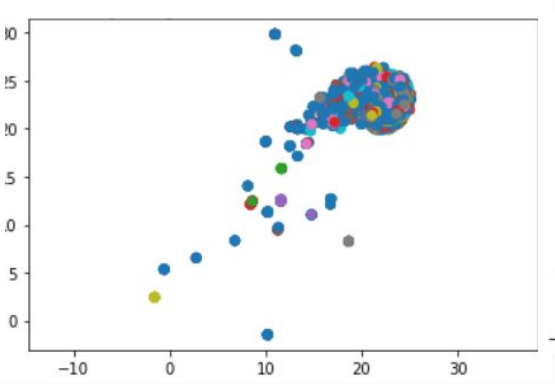
Appendix

r/worldnews n_clusters = 10

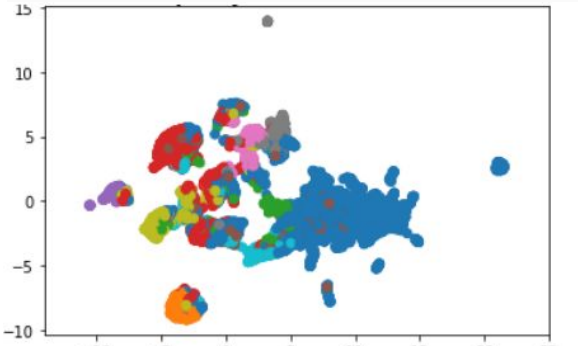
K-Means:



UMAP:

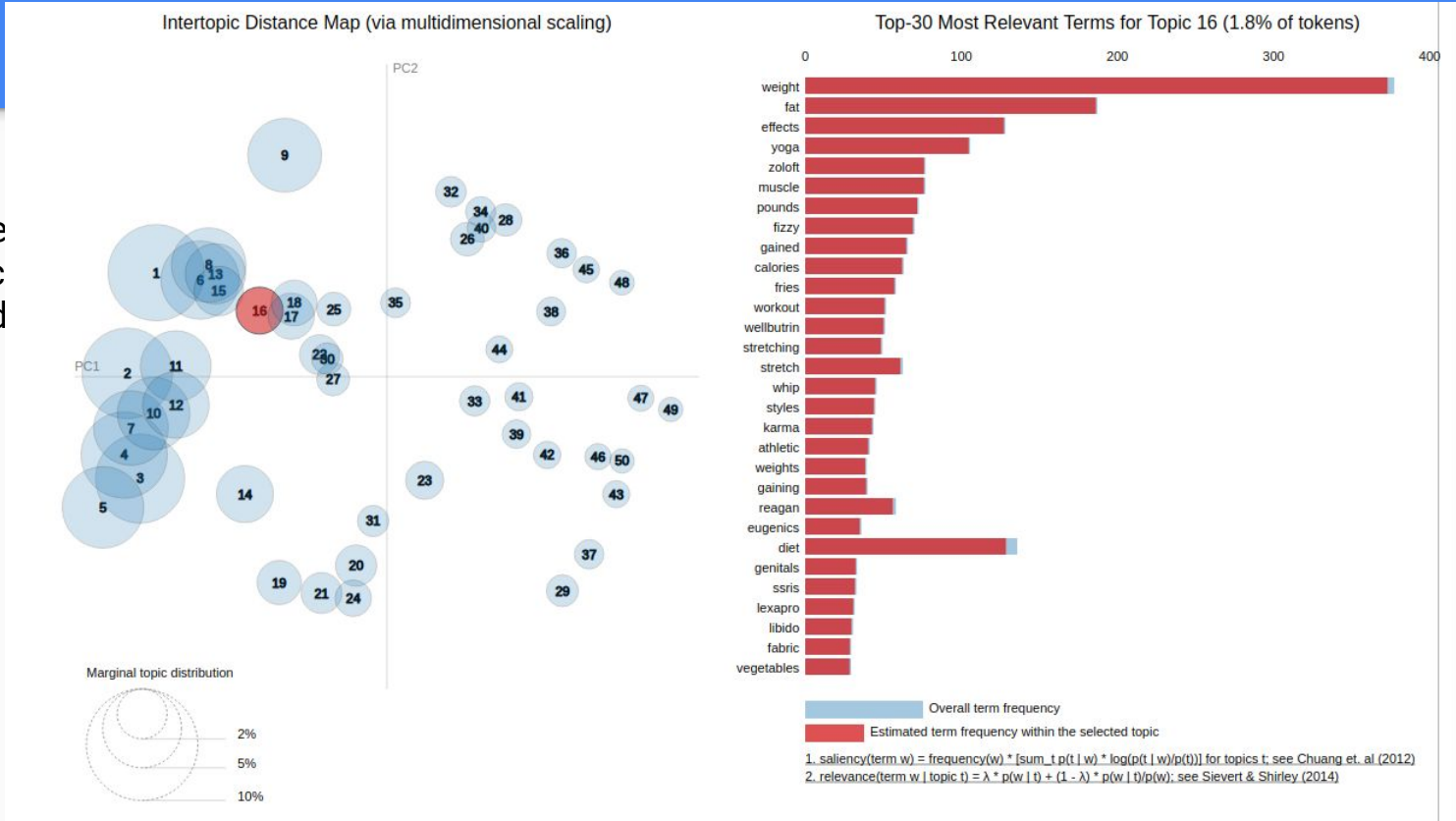


PCA + UMAP:



Visualizing Topic Modeling with pyLDAvis

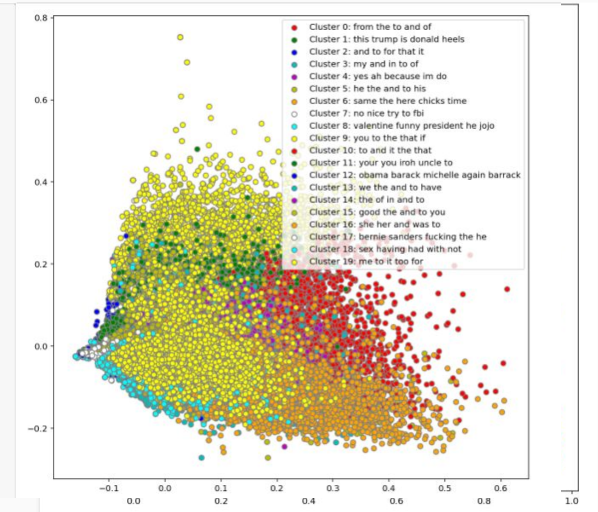
There is a wide varie
askreddit. This topic
antidepressants and
(weight gain, libido)



K Means Clusters

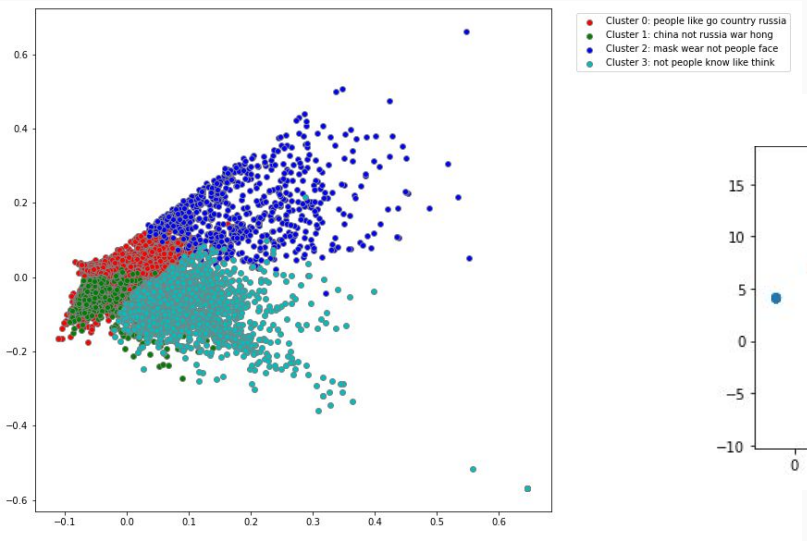
Askreddit had the widest distribution, likely due to the higher volume of text data from them. Topics varied widely, but didn't fit into many clusters.

Other subreddits like r/atheism had a narrower distribution, but still had variation in direction (represented here in the y axis).

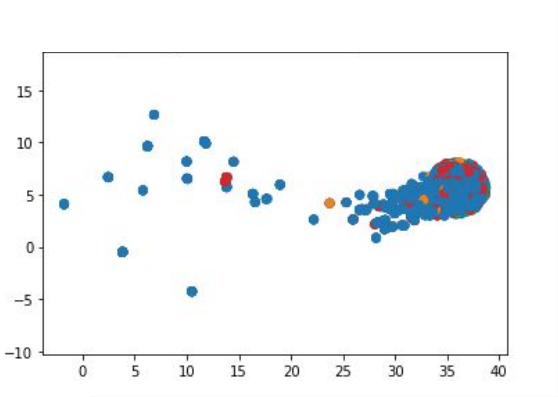


r/worldnews n_clusters = 4

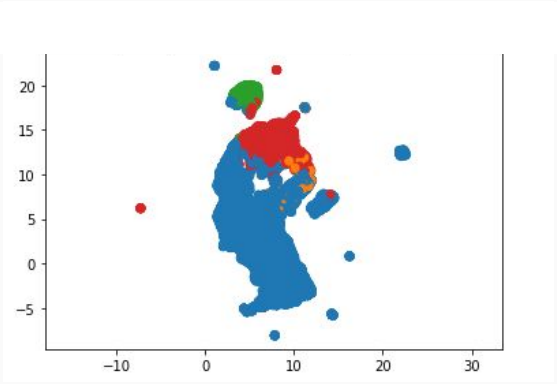
K-Means:



UMAP:



PCA + UMAP:



Topic Modeling with Corex

Corex seemed to give me more clearly differentiated topics than the other topic modeling methods. However, I had trouble getting the corex visualization package to work

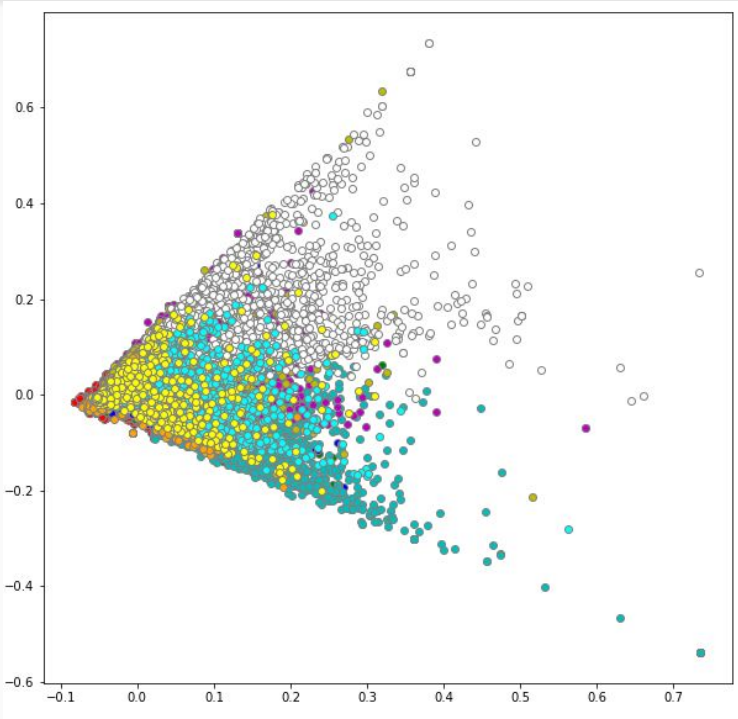
```
0: not,year,like,know,thing,start,want,work,life,way
1: people,need,person,bad,care,problem,actually,job,health,issue
2: time,tell,friend,girl,day,sex,say,night,later,ask
3: pay,money,buy,credit,bank,save,card,company,account,loan
4: president,world,leader,country,war,george,roosevelt,washington,power,political
5: have,feel,get,try,good,come,help,talk,make,week
6: harry,magic,potter,wizard,muggle,speak,english,office,hogwart,teacher
7: home,old,kid,car,parent,family,house,away,drive,young
8: be,go,think,sure,to,pretty,sorry,shit,remember,read
9: jesus,united,states,west,jed,bartlet,wing,vote,christ,joe
10: wear,heel,game,high,play,man,woman,pink,video,shoe
11: eat,use,water,clean,body,originally,weight,ride,blood,healthy
12: bed,break,tooth,drink,brush,morning,smoke,floor,fuck,pull
13: place,big,see,social,deal,kind,call,cost,especially,drug
14: new,run,set,completely,plan,bring,order,course,public,education
15: darth,park,terry,crew,vader,van,mark,der,trailer,assimilate
16: hard,relationship,give,love,let,sexual,partner,hit,feeling,hair
17: valentine,funny,bernie,mr,rogers,iroh,sander,uncle,bob,leslie
18: orgasm,minute,head,face,second,hand,finger,star,climax,blow
19: look,little,bit,child,reason,cause,control,abuse,hold,sort
```

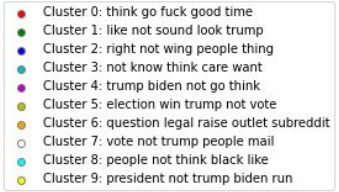
NMF Topic Modeling

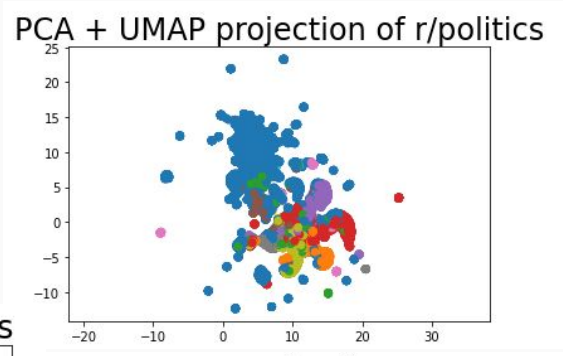
Debatably better separation of topics than LDA, but way too slow

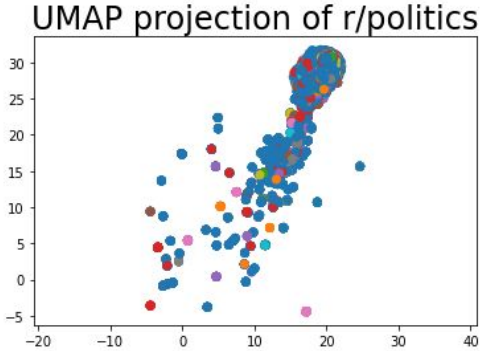
```
Topics in NMF model (generalized Kullback-Leibler divergence):
Topic #0: birthday happy cheers patient celebrate day fun hope today hello cake favourite hey home friend stay bda
y coming fellow im
Topic #1: like just kids quit come ve long people doesn smoke things trying help way told really family smoking ge
tting room
Topic #2: yes actually believe times room didn sure point live haunted sadly mosquito generally lie scared absolut
ely expensive friends asia thier
Topic #3: did stop taxes exactly cat win invest sign sex addicted differently whim np www lottery 2013 24th dog co
ntrol appear
Topic #4: good mate okay guess idea luck pretty power bad honest thanks potatoes deserve jerked feeling friends so
unds causes sense enjoy
Topic #5: don want karma know honestly born world just coffee job look proof guilty won working school college ric
h ghost girlfriend
Topic #6: thanks lot ll stuff doing house great friend questions kind man response comment definitely interesting
appreciated answering idk hope situation
Topic #7: deleted finances finally films film figured figure fights fighting zoompleaños field fiction fiance ferr
ari female felt fellow fight financial feels
Topic #8: think ama ask answer makes mean alters comments numbers wtf qualified quick make stand going adolf https
hitler like special
Topic #9: ve got right experience shit hard type seen different close left working state bring caught little conne
ction grateful difficult explore
Topic #10: nice person really cool lying called subreddit interested tho used liked car ferrari social diagnosed c
oke guy nazi media bugatti
Topic #11: life job want human like bit sucks fuck hell color favorite start friends entire enjoy buy resume livin
g man professional
Topic #12: lol love yeah pizza info yep thats identify dumb small cool welcome aren thoughts kinda crime need game
meme town
Topic #13: years 10 reason pay sister ago going abroad land transcripts condo 11 tea 20s club far pm 25 need posts
Topic #14: like people know post question tell dont sir water boredom bored curiosity favorite die answer normal d
isorder overcome personality checks
Topic #15: say sorry just think oh said question thumbs people ok bro probably wanted business hate op congratulat
ions username wasn funny
Topic #16: time thank day tell happened hired great god friend hope story learn try movie removed positive grand s
aid happen learning
Topic #17: money thing best family wish bought hope awesome millions buy worst happiness better lmao like thought
big spent asked caused
Topic #18: true feel make reddit work half congrats money does biggest new current tax enjoy use change gold says
gay expect
Topic #19: 13 old girl year allegations false marvellady24 claims respond marvel 24 number true ms like alright dr
ugs sleep mr autistic
```

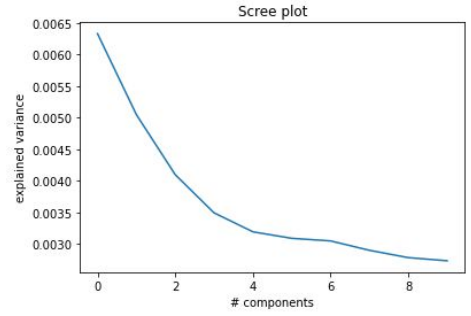
KMeans + PCA + UMAP r/politics

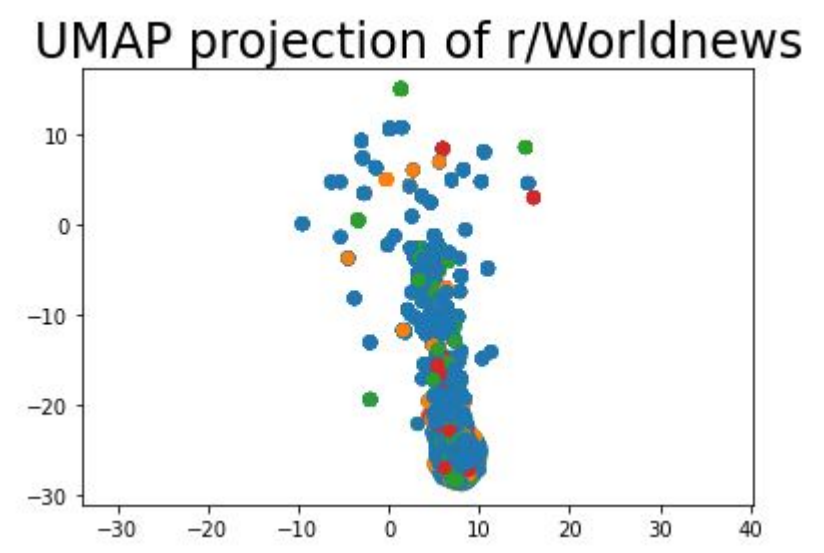


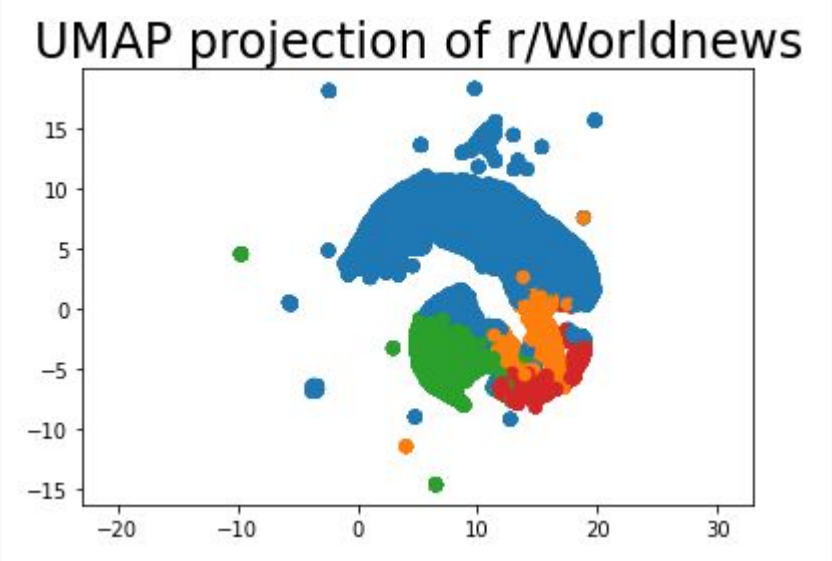






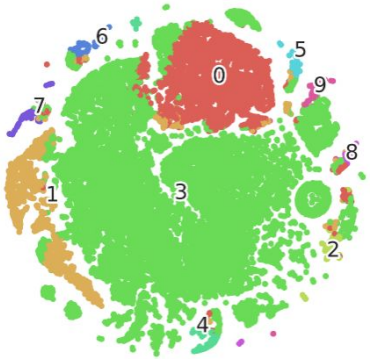


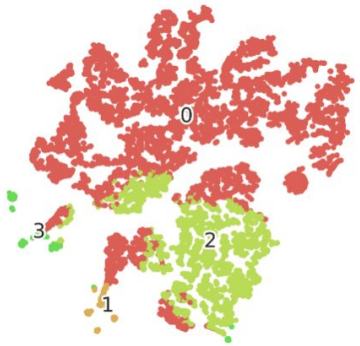

UMAP projection of r/Worldnews
10
0
−10
−20
−30
−30
−20
−10
0
10
20
30
40

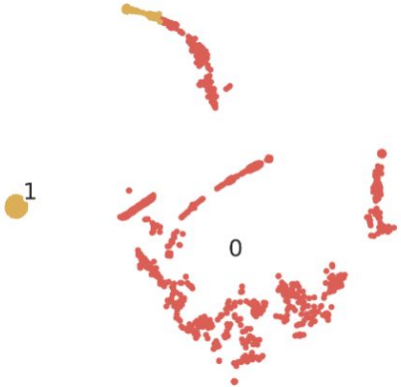

UMAP projection of r/Worldnews
15
10
5
0
−5
−10
−15
−20
−10
0
10
20
30

tSNE

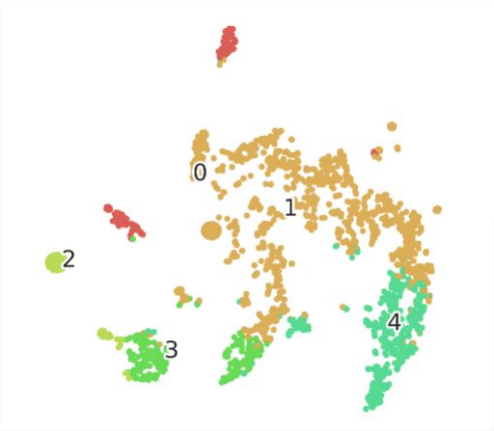
Would've been great if it was faster! Extremely slow with anything bigger than a small sample of my data

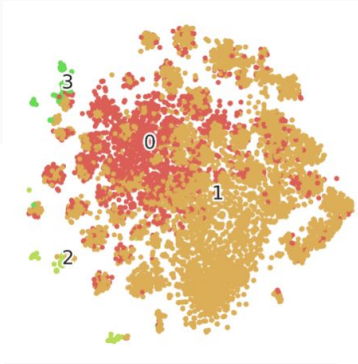


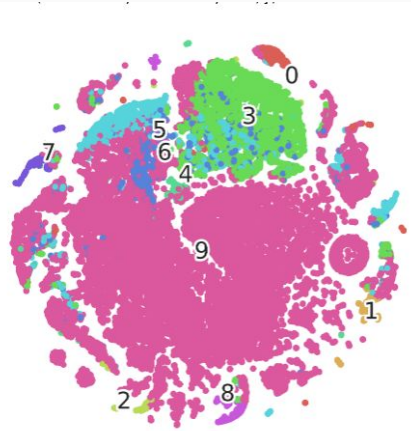


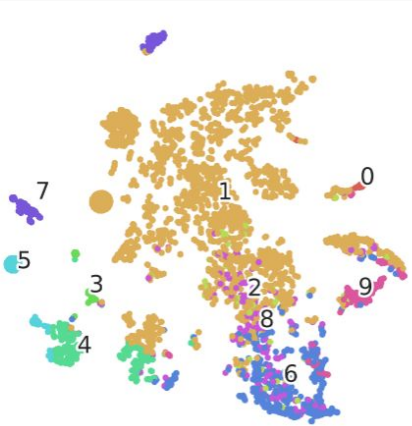


tSNE

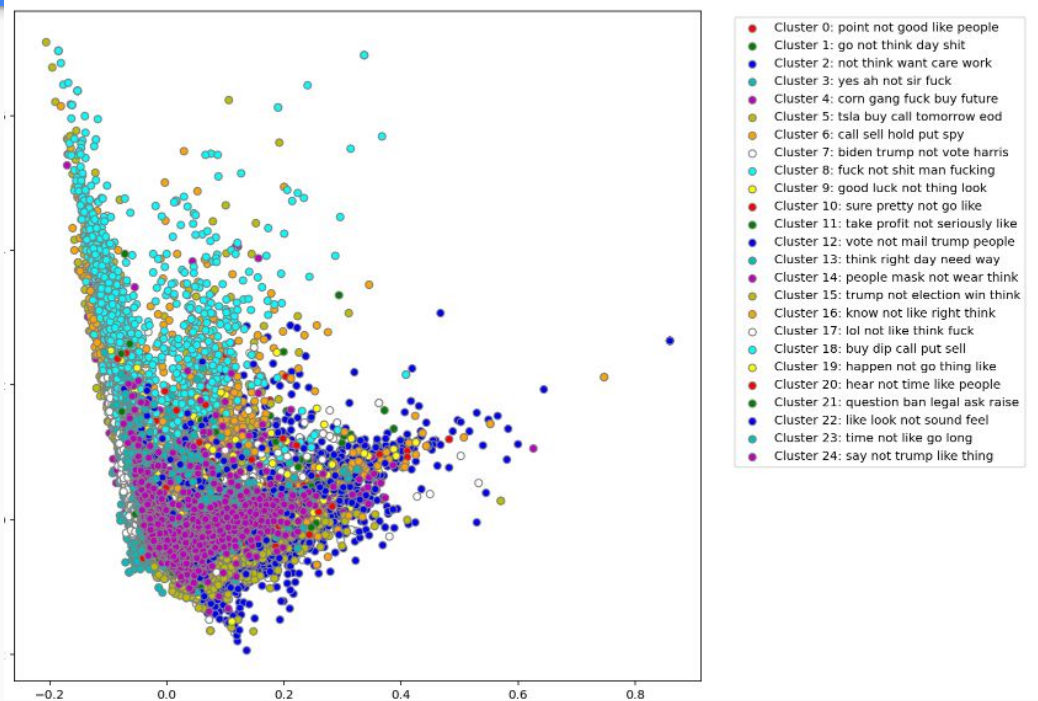






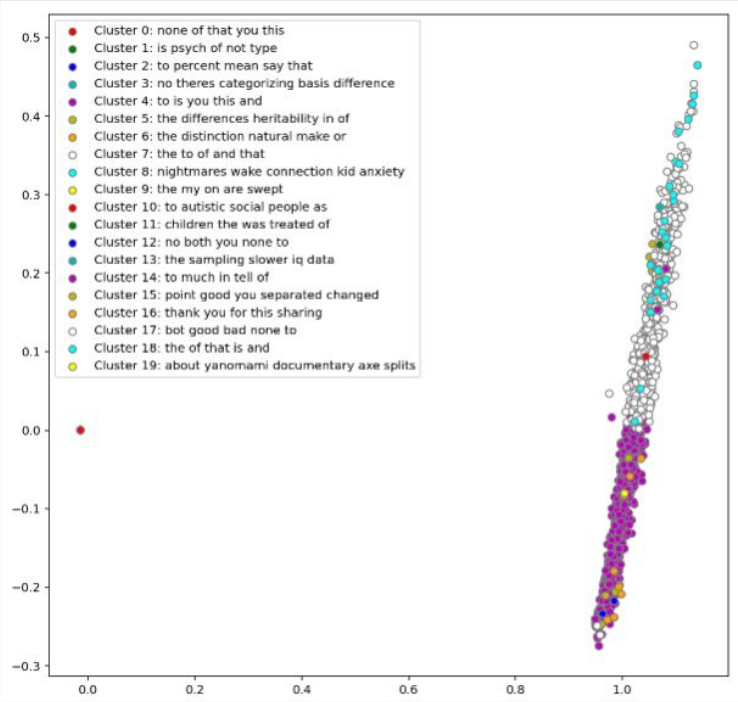


r/wallstreetbets KMeans



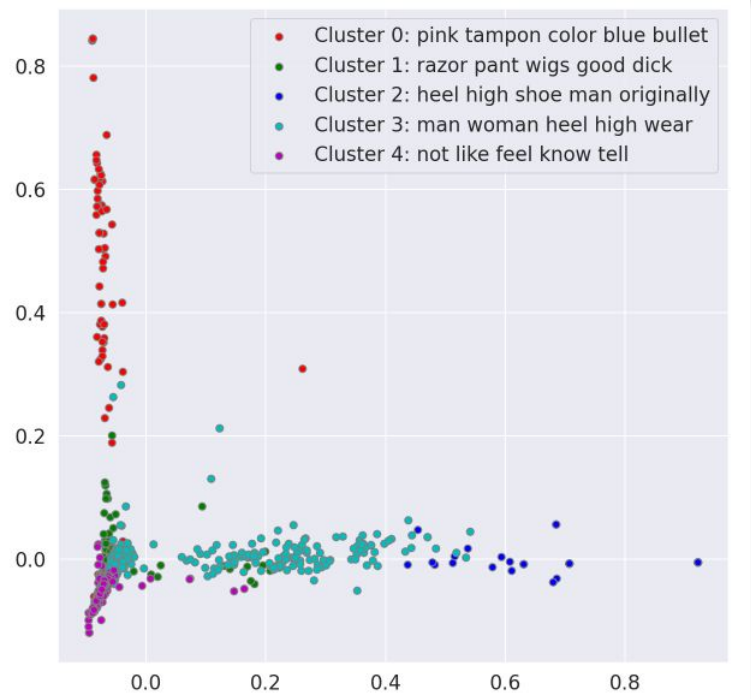
Wide variance, similar to r/askreddit. Different shape though, indicating slightly different topic distribution

r/psychology KMeans



r/psychology's distribution is almost as thin as r/memes, interestingly… but there is still a lot of variance in the vertical direction

KMeans from a tiny sample of r/askreddit for running some tests more quickly



The topics are kind of funny

Hypothesis

H_0:Comments from one subreddit will not be different than comments from another subreddit.

H_A:Comments from one subreddit will have different scores than comments from another subreddit.