APOLLO : APOLitical Linguistic Overview
Week 07 meeting

정연환, 김구영, 권병훈

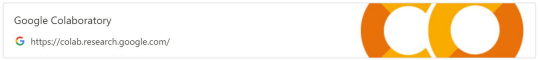
Progress report

week 06

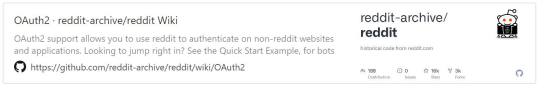
Todo

Setup

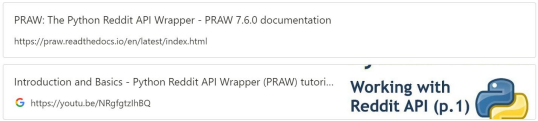
Jupyter notebook (or Google Colab)



Reddit API



PRAW



Playground

Subreddit

본인이 선정한 subreddit들의 hot, new, rising, top (month) 게시글 10개씩 title을 추출하여 list로 저장.

Submission

위 subreddit에서 저장한 게시글 40개의 attributes 저장.

저장할 attributes

- title
- author
- created_utc
- id
- score
- upvote_ratio

각 attributes들을 저장하여 dataframe으로 만드는 것을 추천.

User

위 submission에서 저장한 게시글 40개의 author들의 attributes 저장.

저장할 attributes

- created_utc
- name
- has_verified_email
- is_mod

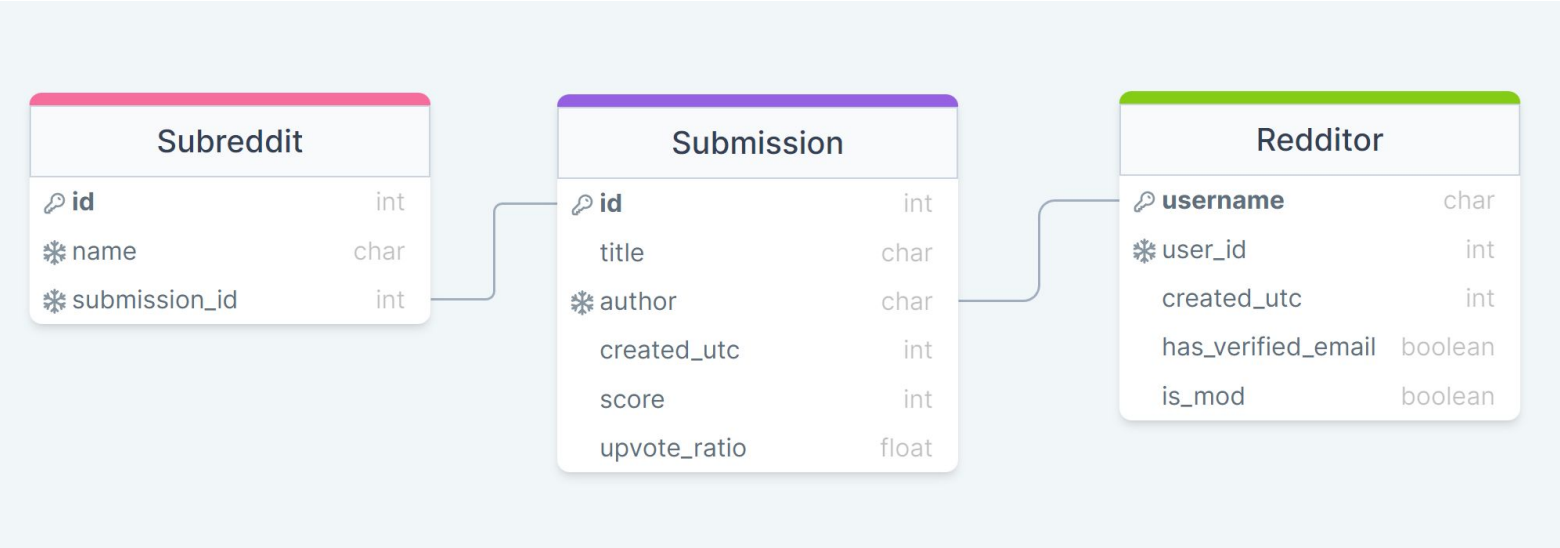
Output

Store data

위 playground에서 저장한 dataframe을 csv와 json으로 각각 drop.

각 subreddit 당 한 세트의 파일을 만들 것. (e.g., "wallstreetbets.csv", "wallstreetbets.json")

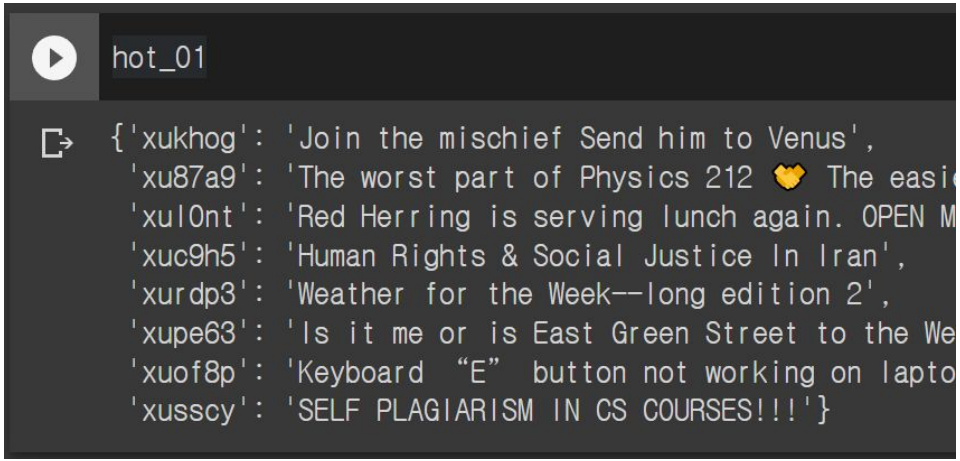
Database


Subreddit
id int
name char
submission_id int
Submission
id int
title char
author char
created_utc int
score int
upvote_ratio float
Redditor
username char
user_id int
created_utc int
has_verified_email boolean
is_mod boolean

Demo - Subreddit

- Extracted ids and titles from the r/UIUC - hot

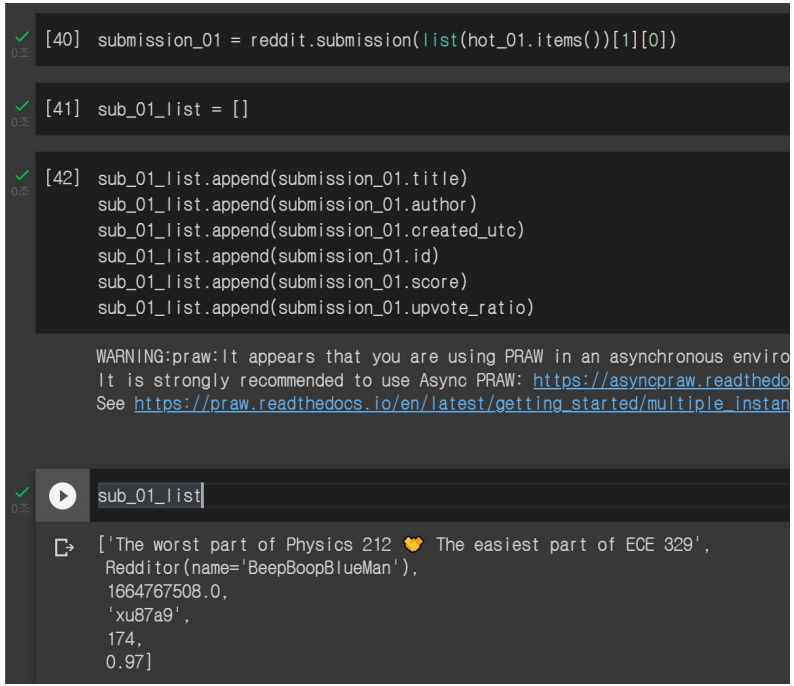
- Store their submission_id that could used as extracting from the submission



Demo - Submission

- From the submission_id, we can extract useful information from the each submission

- We can store the author, which also can be used to extract the user information



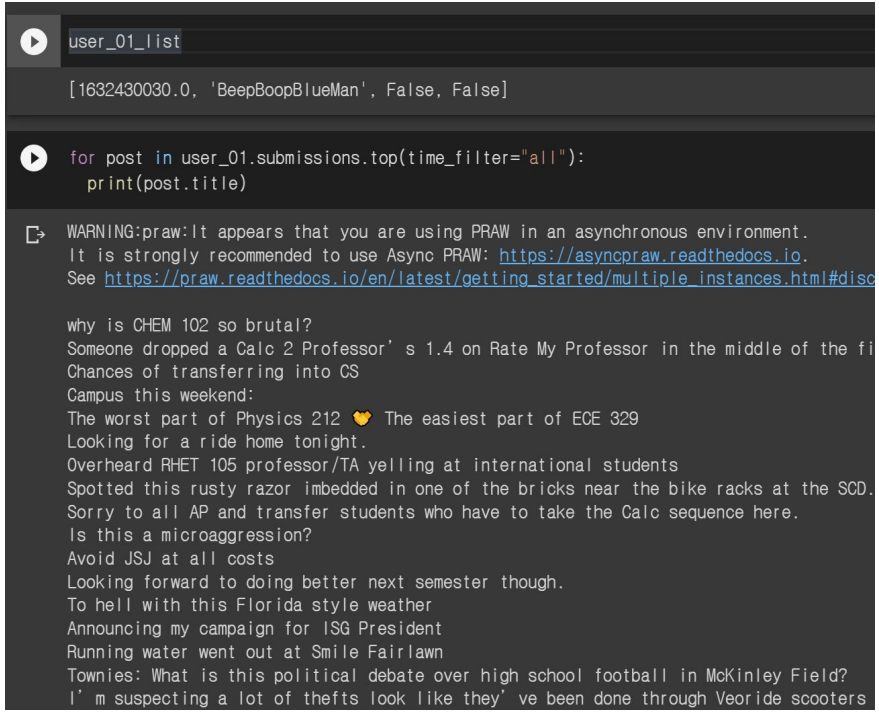
Demo - Submission (r/Forluma1)

	ID	Title	Author	Created (UTC)	Score	Upvote Ratio
0	xuc6c7	Ask /r/formula1 Anything - Daily Discussion - ...	F1-Bot	1.664780e+09	15	0.74
1	xuhhx7	2022 Singapore Grand Prix - Day after Debrief	F1-Bot	1.664798e+09	102	0.94
2	xuodvc	[@F1] At 11:11:11 the chequered flag was waved...	KaamDeveloper	1.664815e+09	2266	0.96
3	xugk3v	Russell asserts he left Schumacher space to av...	jovanmilic97	1.664796e+09	3241	0.93
4	xubpxo	I paid $84 dollars USD (was actually $130 AUD)...	HAMILT0N	1.664779e+09	8513	0.93

- We can store the information as a dataframe for the top submissions from r/Formula1

Demo - Redditor

- From the author, we can locate the user information
- We can track the posts that specific user posted, and extract useful information as well
- We can get a snapshot of the posts



Demo - Running Sentiment Analysis on post titles

Run sentiment analysis on a user's post titles

In [1]:
```
from nltk.sentiment import SentimentIntensityAnalyzer
sia = SentimentIntensityAnalyzer()
```

In [4]:
```
words = [w for w in nltk.corpus.state_union.words() if w.isalpha()]
stopwords = nltk.corpus.stopwords.words("english")
words = [w for w in words if w.lower() not in stopwords]
```

In [5]:
```
sia.polarity_scores("Wow, NLTK is really powerful!")
```

Out[5]:
```
{'neg': 0.0, 'neu': 0.295, 'pos': 0.705, 'compound': 0.8012}
```

In [8]:
```
user_03 = reddit.redditor('sssssstonkssss')
# print(user_03.created_utc)
# print(user_03.name)
# print(user_03.has_verified_email)
```

In [9]:
```
user_03_posts=[]
for post in user_03.submissions.top(time_filter="all"):
    user_03_posts.append(post.title)
    user_03_posts.append(sia.polarity_scores(post.title))
```

In [10]:
```
user_03_posts
```

```
 'Highly accurate... your purchase of this merchandise at concert pricing is in fact a donation to your billionaire grifter chairperson',
 {'neg': 0.0, 'neu': 1.0, 'pos': 0.0, 'compound': 0.0},
 'Just a reminder. Note that apes typically leave off that little paragraph at the bottom of the page.',
 {'neg': 0.07, 'neu': 0.93, 'pos': 0.0, 'compound': -0.0516},
 'Behold, our future overlords, the True And Noble Arbiters Of All That Is Good And Right™',
 {'neg': 0.0, 'neu': 0.599, 'pos': 0.401, 'compound': 0.8271},
 'Is their bot smart enough to recognize duplicate or add-on submissions? Or is it counting every single submission as adding that many shares to the galaxy tank?',
 {'neg': 0.0, 'neu': 0.826, 'pos': 0.174, 'compound': 0.644},
 'Why is GPU mining profitability crashing?',
 {'neg': 0.0, 'neu': 0.704, 'pos': 0.296, 'compound': 0.2732},
 'Is high debt and negative FCF typical for mortgage companies?',
 {'neg': 0.437, 'neu': 0.563, 'pos': 0.0, 'compound': -0.7351},
 'Is the finra margin debt retail or institutional?',
 {'neg': 0.263, 'neu': 0.737, 'pos': 0.0, 'compound': -0.3612},
 'is it just me, or is there a distinct correlation between TSLA dropping and Elon releasing the latest gimmick? 🤔',
 {'neg': 0.0, 'neu': 1.0, 'pos': 0.0, 'compound': 0.0},
 'rYAN cOhen poAcHINg tOP MANAgErS fROM OThEr cOMPaNiES',
 {'neg': 0.0, 'neu': 0.795, 'pos': 0.205, 'compound': 0.2023},
```

Next step

- Find ways to analyze deeper
  - We may need to find a way to sentiment-analyze a user's posts AND comments so we can better understand their political / social agenda

- Better improve dataframe structures
  - Storing information of subreddits, redditors, and posts into a easily manageable dataframe is essential for future uses

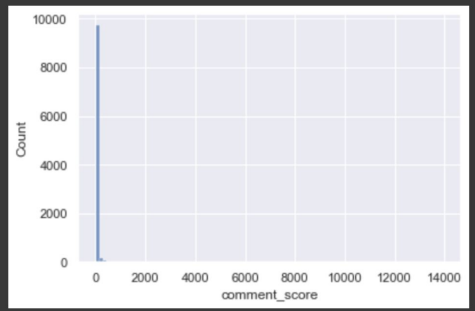
Progress report

Week 09

Research question thinking

Data cleaning

```
1 sns.set(style="darkgrid")
2 sns.histplot(data=pol_comment_df, x="comment_score",  bins=100)
3 plt.show()
```

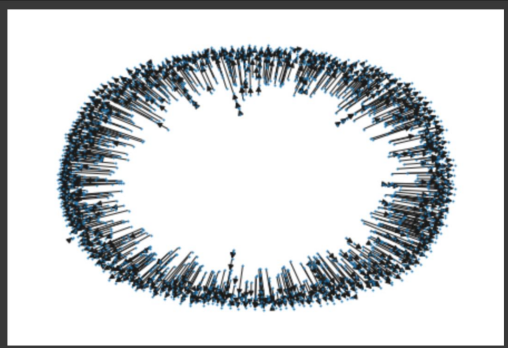


We need to clean above data, right?

위에서 보이는 것처럼 우리의 데이터는 굉장히 skewed 되어 있는데, 이 데이터를 어떻게 정리할 것인가?

```
[ ]  1 G_01.remove_nodes_from(list(nx.isolates(G_01)))
```

```
1 nx.draw(G_01, node_size = 1)
```



Seems quite perplex. How can we extract meaningful information here?

위 그림은 추출한 댓글들의 네트워크를 추출한 것이다.

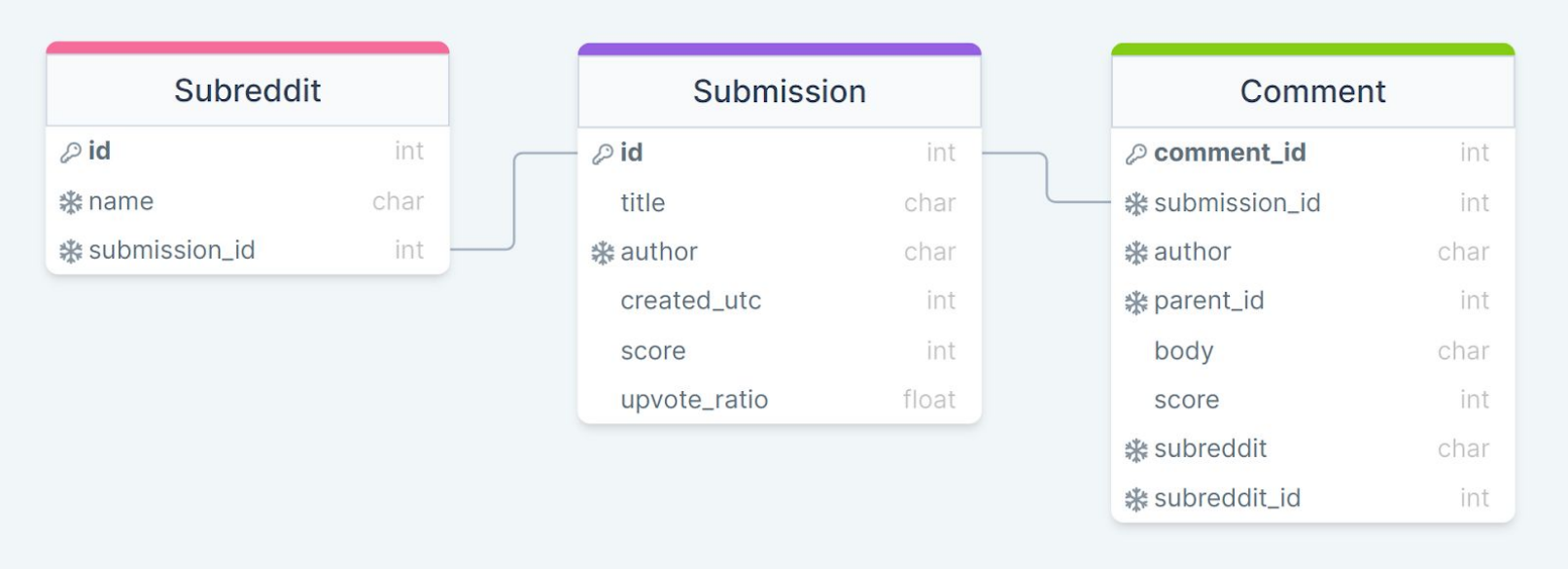
이 네트워크는 굉장히 무분별하게 분류되어져 있는데, 우리가 유의미한 데이터를 찾기 위해서는 어떻게 해야 할까?

Additional question

많은 추천을 받은 게시글들의 upvote_ratio 는 어떠한가? subreddit이 controversial 한가?

많은 추천을 받은 게시글들을 적은 user의 특성은? 게시글을 많이 쓴 사람인가? 적게 쓴 사람인가? 그 user의 평균 score는?

Database


Subreddit
id int
name char
submission_id int
Submission
id int
title char
author char
created_utc int
score int
upvote_ratio float
Comment
comment_id int
submission_id int
author char
parent_id int
body char
score int
subreddit char
subreddit_id int

Data collection

Collected subreddit

5

r/Politics 포함 5 서브레딧

Collected submission

4105

대부분 selftext 대신 링크

Collected comments

3 M

엄청 시간 오래걸림

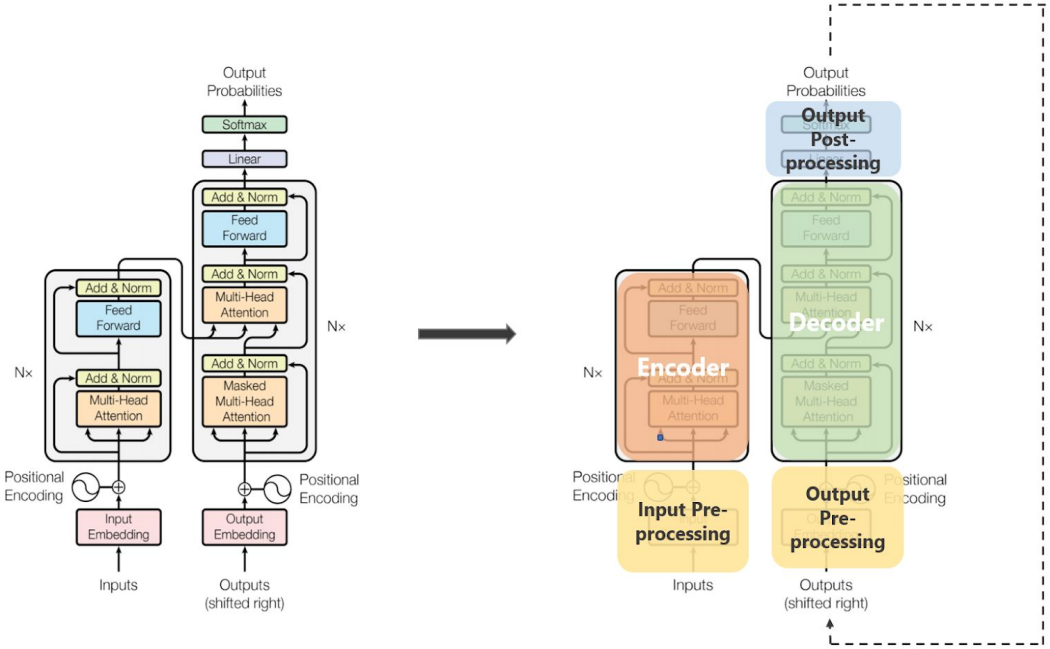
Politics vs Neutralpolitics

```
pol_df[:10]
```

	id	author	body	parent_id	score	subreddit	subreddit_id	submission_id
0	hynym3j	AutoModerator	₩nAs a reminder, this subreddit [is for civil ...	t3_t2suj2	1	politics	t5_2cneq	t2suj2
1	hyotmts	Lunar_Deer	Americans who support Putin can get fucked.	t3_t2suj2	7055	politics	t5_2cneq	t2suj2
2	hyo3bc6	workerbee77	Bob Mueller laid out the charges that Donald T...	t3_t2suj2	3709	politics	t5_2cneq	t2suj2
3	hyo265c	bobface222	He is almost correct	t3_t2suj2	15571	politics	t5_2cneq	t2suj2
4	hyoiog1	Jacob_C	I don't think people here understand the word ...	t3_t2suj2	1582	politics	t5_2cneq	t2suj2
5	hyo0agn	Bess_Marvin_Curls	Take out "almost". And trump is #1 in treason.	t3_t2suj2	3454	politics	t5_2cneq	t2suj2
6	hyo0m2p	notyomamasusername	I hope this is the crucible that finally clean...	t3_t2suj2	696	politics	t5_2cneq	t2suj2
7	hyoofxx	-Economist-	The good news, my dad is switching to voting b...	t3_t2suj2	220	politics	t5_2cneq	t2suj2
8	hyo1m3j	nodustspeck	Your votes need to reflect what you say, Romne...	t3_t2suj2	770	politics	t5_2cneq	t2suj2
9	hyp8991	Zezin96	I'm beginning to regret dragging this guy thro...	t3_t2suj2	44	politics	t5_2cneq	t2suj2

11	idpu0vb	Urgullibl	I think it's	t3_vjxj2b	5	NeutralPoli	t5_2tk0i	vjxj2b
12	idlxrao	canekicker	One set of	t3_vjxj2b	32	NeutralPoli	t5_2tk0i	vjxj2b
13	idng7v9	Illustrious-f	What	t3_vjxj2b	7	NeutralPoli	t5_2tk0i	vjxj2b
14	idq4xpi	bucky001	This	t3_vjxj2b	2	NeutralPoli	t5_2tk0i	vjxj2b
15	idoiphw	Jackpot777	I guess	t3_vjxj2b	5	NeutralPoli	t5_2tk0i	vjxj2b
16	idpotig		[deleted]	t3_vjxj2b	2	NeutralPoli	t5_2tk0i	vjxj2b
17	idm65db		[removed]	t3_vjxj2b	0	NeutralPoli	t5_2tk0i	vjxj2b
18	idnpesx		[removed]	t3_vjxj2b	-5	NeutralPoli	t5_2tk0i	vjxj2b
19	idnn9bm		[removed]	t3_vjxj2b	-5	NeutralPoli	t5_2tk0i	vjxj2b
20	idlv2nd		[removed]	t3_vjxj2b	0	NeutralPoli	t5_2tk0i	vjxj2b
21	idmbm1p		[removed]	t3_vjxj2b	1	NeutralPoli	t5_2tk0i	vjxj2b
22	idmdkzs		[removed]	t3_vjxj2b	0	NeutralPoli	t5_2tk0i	vjxj2b
23	idms5ug		[removed]	t3_vjxj2b	1	NeutralPoli	t5_2tk0i	vjxj2b
24	idmt2fc		[removed]	t3_vjxj2b	1	NeutralPoli	t5_2tk0i	vjxj2b
25	idn1j28		[removed]	t3_vjxj2b	1	NeutralPoli	t5_2tk0i	vjxj2b
26	idn7jyf		[removed]	t3_vjxj2b	1	NeutralPoli	t5_2tk0i	vjxj2b
27	idncg5w		[removed]	t3_vjxj2b	1	NeutralPoli	t5_2tk0i	vjxj2b
28	idos8dj		[removed]	t3_vjxj2b	1	NeutralPoli	t5_2tk0i	vjxj2b
29	idp2a1o		[removed]	t3_vjxj2b	1	NeutralPoli	t5_2tk0i	vjxj2b
30	idp3iem		[removed]	t3_vjxj2b	0	NeutralPoli	t5_2tk0i	vjxj2b
31	idpgb06		[removed]	t3_vjxj2b	0	NeutralPoli	t5_2tk0i	vjxj2b
32	idpjmb8		[removed]	t3_vjxj2b	1	NeutralPoli	t5_2tk0i	vjxj2b
33	idpprsr		[removed]	t3_vjxj2b	1	NeutralPoli	t5_2tk0i	vjxj2b
34	idq42ab		[removed]	t3_vjxj2b	1	NeutralPoli	t5_2tk0i	vjxj2b
35	idqavqd		[removed]	t3_vjxj2b	1	NeutralPoli	t5_2tk0i	vjxj2b
36	idx7as9		[removed]	t3_vjxj2b	1	NeutralPoli	t5_2tk0i	vjxj2b
37	iez2pib		[removed]	t3_vjxj2b	1	NeutralPoli	t5_2tk0i	vjxj2b
38	igl6zq1	vmerc	The link in	t3_vjxj2b	1	NeutralPoli	t5_2tk0i	vjxj2b
39	idmjauc		[removed]	t3_vjxj2b	-2	NeutralPoli	t5_2tk0i	vjxj2b
40	idm7qjc		[removed]	t3_vjxj2b	-1	NeutralPoli	t5_2tk0i	vjxj2b
41	idly8dc		[removed]	t3_vjxj2b	3	NeutralPoli	t5_2tk0i	vjxj2b

Transformers



Toxicity analysis

body	parent_id	score	subreddit	subreddit_id	submission_id	label	toxicity_sco
MAGA republicans: "Fuck Joe Biden!" ₩n₩nJoe B...	t3_x59p0m	11011	politics	t5_2cneq	x59p0m	toxic	0.998468
T1 diabetic here. Fuck you GOP. Fuck you and...	t3_wihq4u	9258	politics	t5_2cneq	wihq4u	toxic	0.997967
Get fucked Cannon.	t3_xwgido	12843	politics	t5_2cneq	xwgido	toxic	0.997483
Everything is fun and games until the whole wo...	t3_t4q4nz	7294	politics	t5_2cneq	t4q4nz	toxic	0.997282
Get this fucker off the bench. This is ridiculous	t3_vodnck	12812	politics	t5_2cneq	vodnck	toxic	0.997271

body	parent_id	score	subreddit	subreddit_id	submission_id	label	toxicity_score
Why the fuck is socialism wrong?!₩n₩nDonthey e...	t3_xxx3r7	-72	politics	t5_2cneq	xxx3r7	toxic	0.998709
Young people are fucking dumb lol	t3_xdicpa	-55	politics	t5_2cneq	xdicpa	toxic	0.998462
Imagine being scared of a dude in a fucking bu...	t3_sntmw2	-74	politics	t5_2cneq	sntmw2	toxic	0.998447
Lmfao anyone listening to Noam fucking Chomsky...	t3_sk8ffr	-71	politics	t5_2cneq	sk8ffr	toxic	0.997456
Thats still the stupidest fucking nickname ever	t3_x77612	-111	politics	t5_2cneq	x77612	toxic	0.996905

Network analysis

Next step

- Investigate network more
  - Maybe we can find some interesting nodes
  - “Clustering” with other networks

- Investigate text data
  - NLP stuffs, maybe we can combine comments as tree
    - Like this

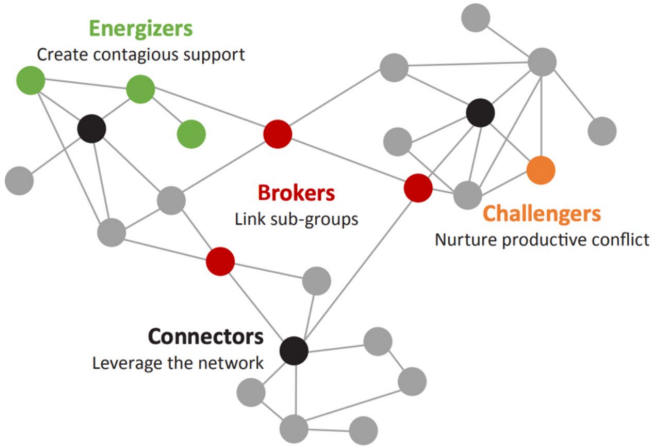
Network analysis

- So what
- Difficult to know node attribute
- Hard to identify the cluster

Characteristics of the network

week 10

Network Analysis



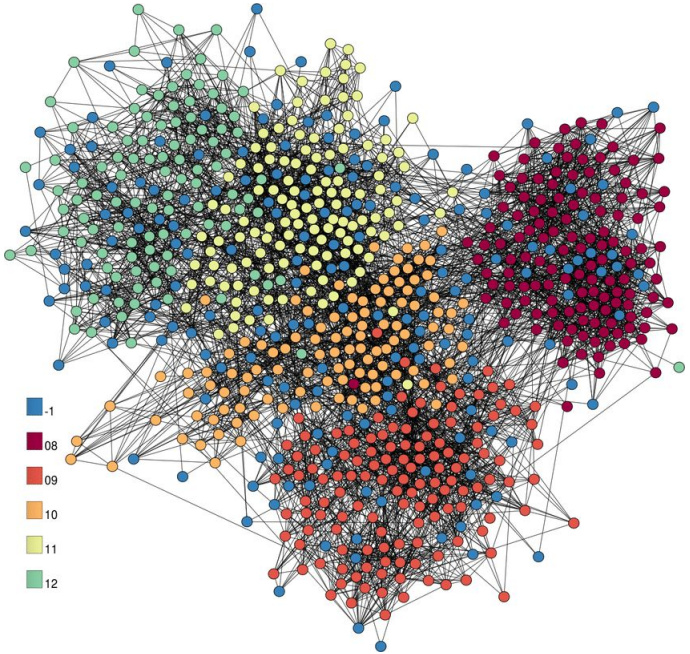
What is the network type?

Methodology

https://towardsdatascience.com/pyvis-visualize-interactive-network-graphs-in-python-77e059791f01

Methodology

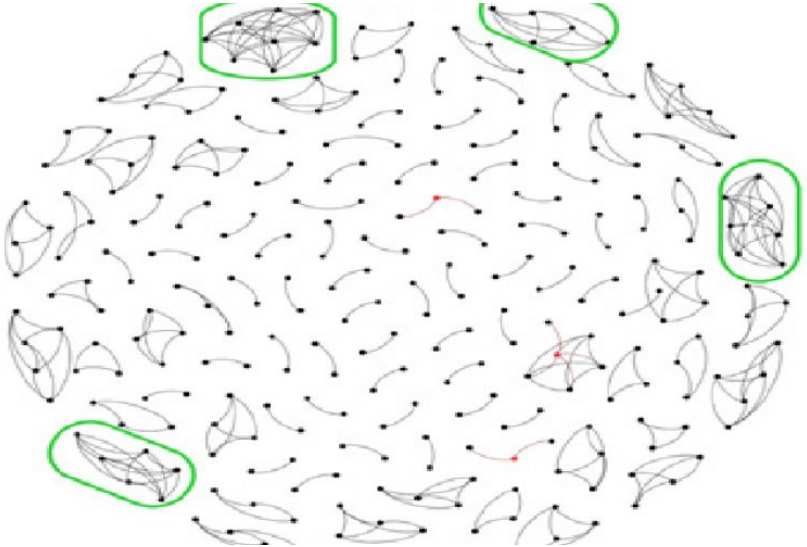
In [7]:
```
def network_to_df(net, threshold = 6):
    counter = Counter(net)
    counter = Counter({k: c for k, c in counter.items() if c >= threshold})
    common_relation = sorted(counter, key=counter.get, reverse=True)

    relation_df = pd.DataFrame.from_dict(counter, orient='index').reset_index()
    relation_df[['target','source']] = pd.DataFrame(relation_df['index'].tolist(
    relation_df = relation_df[relation_df['source'] != relation_df['target']]
    relation_df.drop(['index'], axis = 1, inplace = True)
    relation_df.rename(columns = {0:'weight'}, inplace = True)
    relation_df = relation_df[['source', 'target', 'weight']]

    return relation_df
```



Muelder, Chris et al. (2014). Multivariate Social Network Visual Analytics. 10.1007/978-3-319-06793-3_3.

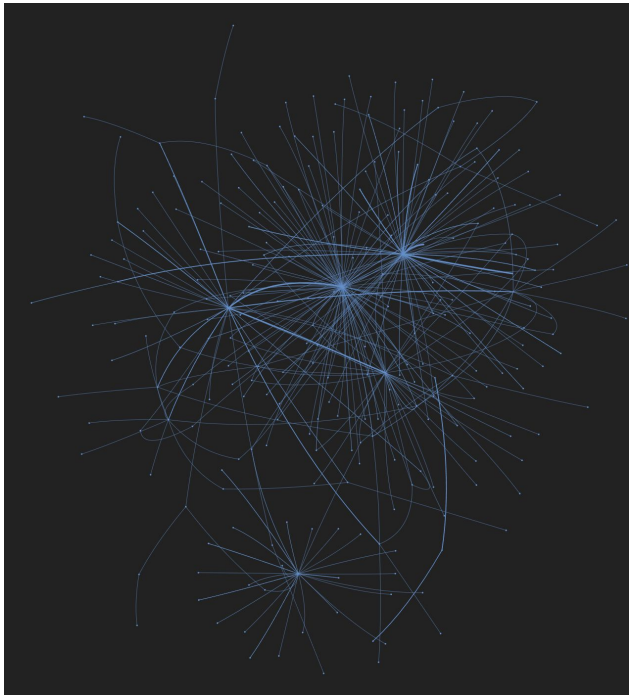
Methodology

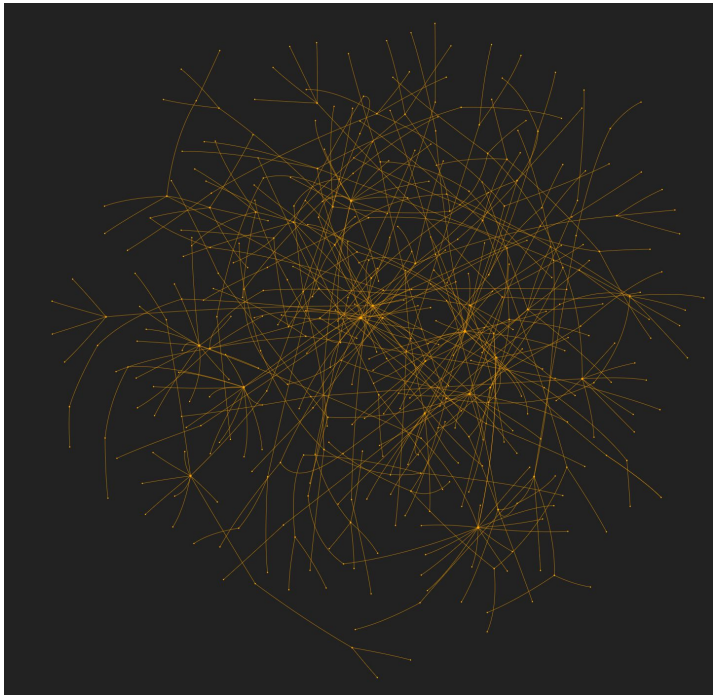
```
def network_builder(relation_df, threshold = 5, net_type = "simple", sub = None):
    filename = 'output/r_' + sub + '_' + net_type + '.html'

    G = nx.from_pandas_edgelist(relation_df, 'source', 'target', 'weight')

    for component in list(nx.connected_components(G)):
        if len(component) < threshold:
            for node in component:
                G.remove_node(node)
```



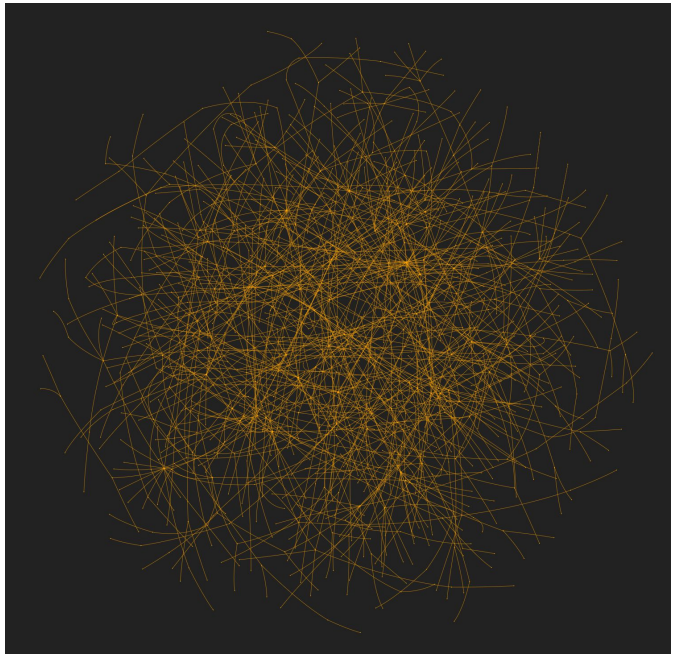
Gaskó, N., Bota, F., Suciu, M. A., & Lung, R. I. (2020). A Game Theoretical Analysis of Academic Writing Co-authorship Networks. J. Sci. Res., 9(3), 319-325.

Result



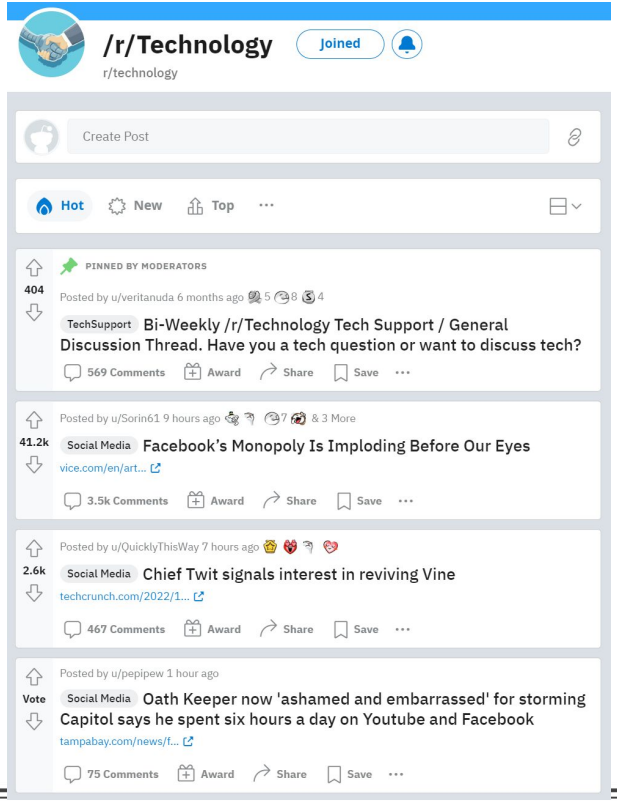


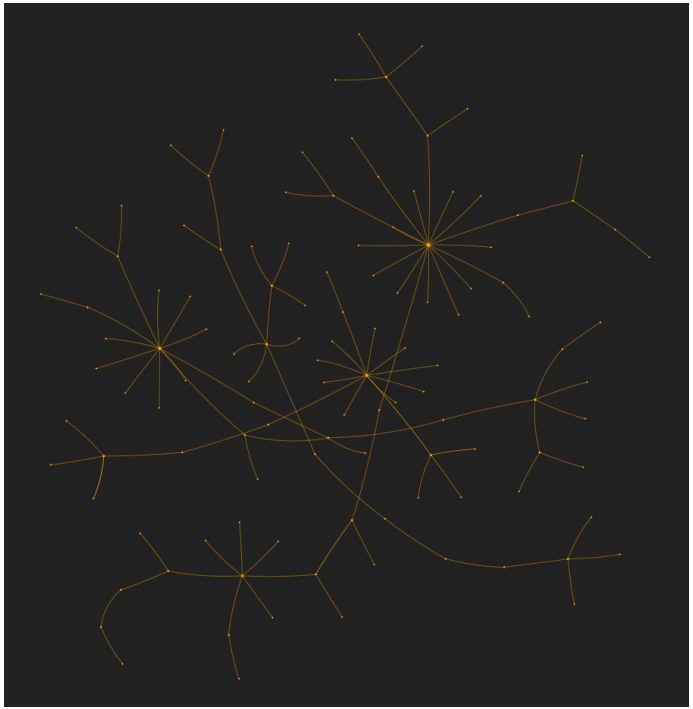
Result



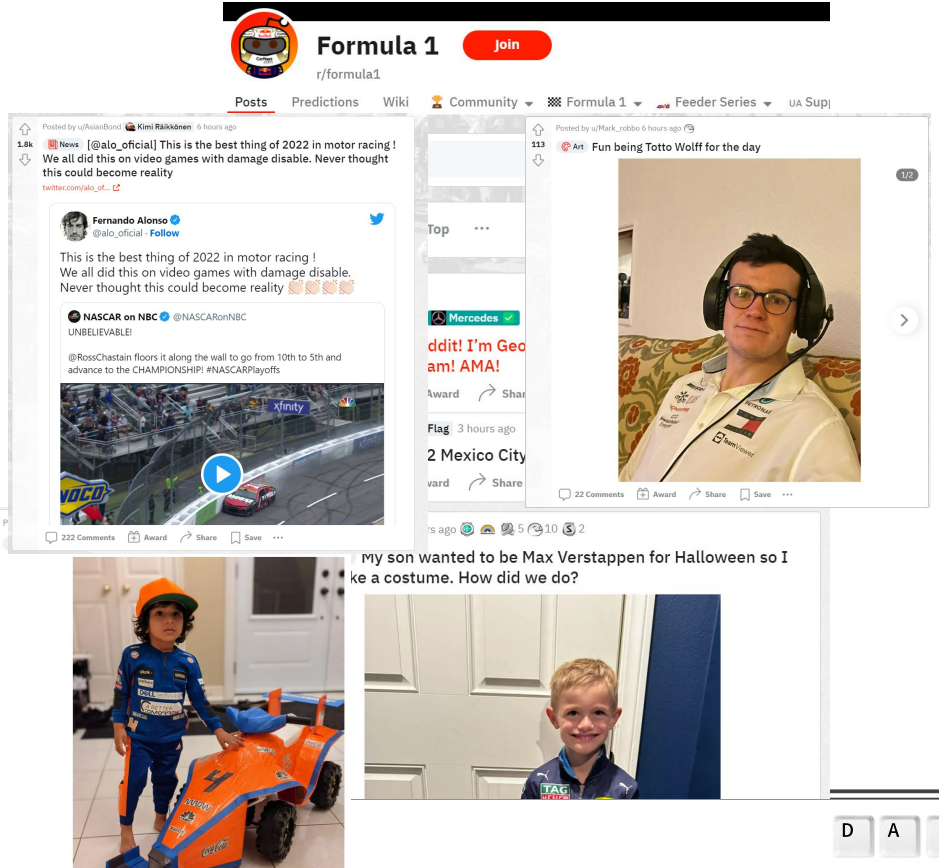
흡사...

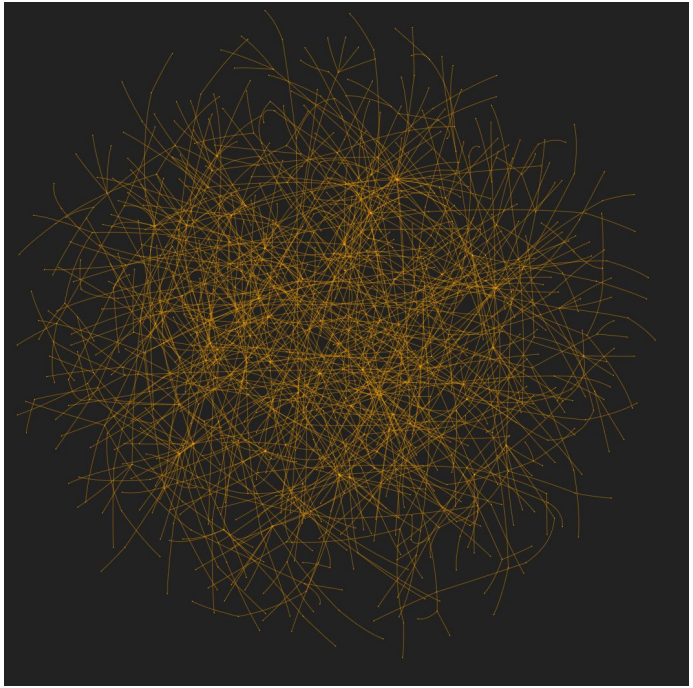
Shape differs by subreddit





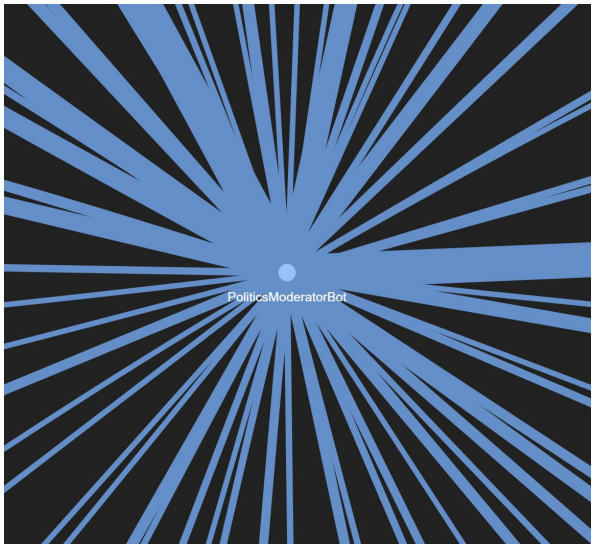
Shape differs by subreddit

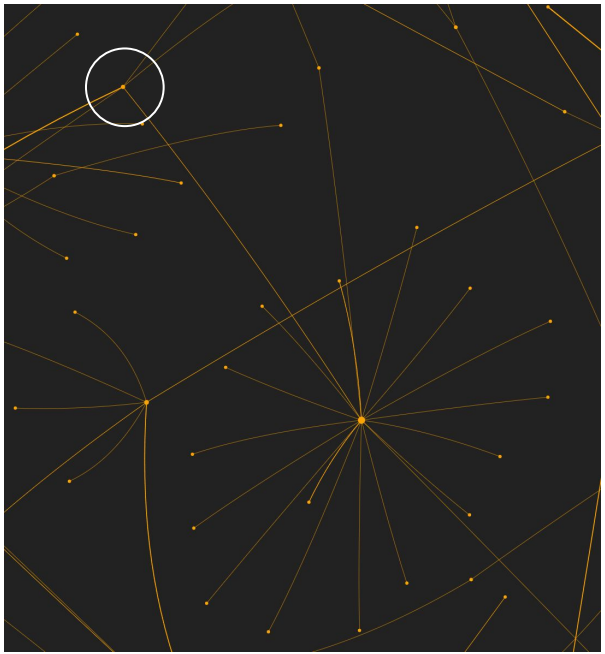




D A I S

Demo

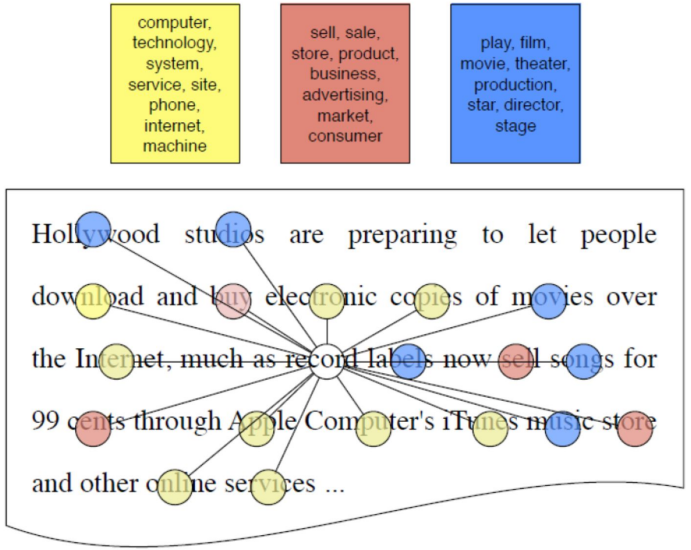




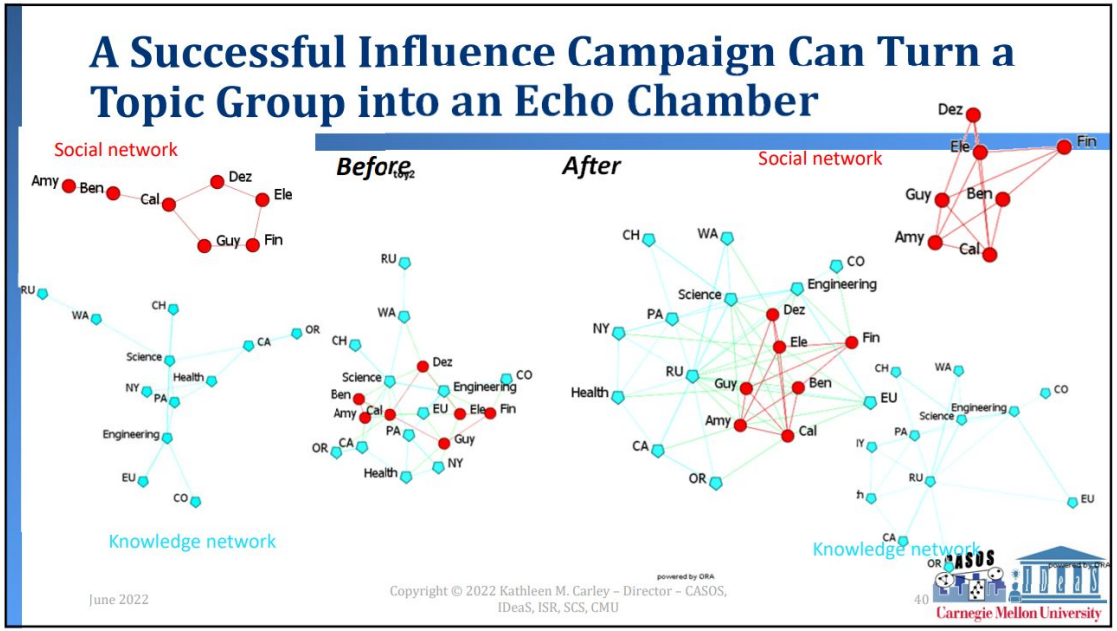
Next step

TF-IDF

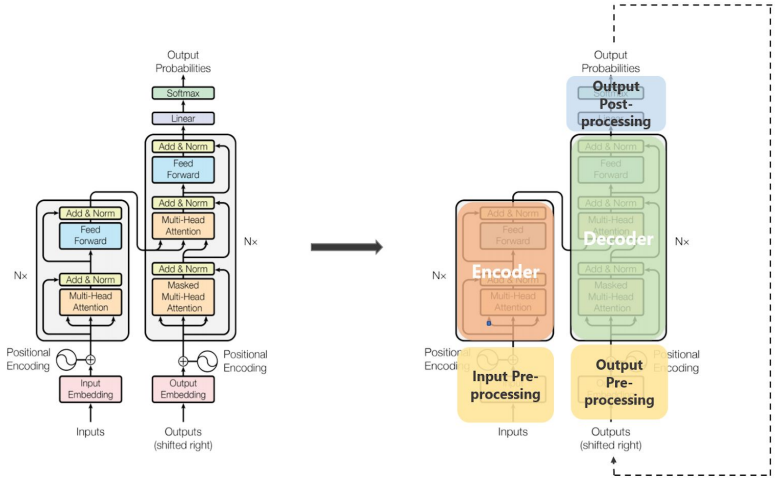
LDA



Introduction

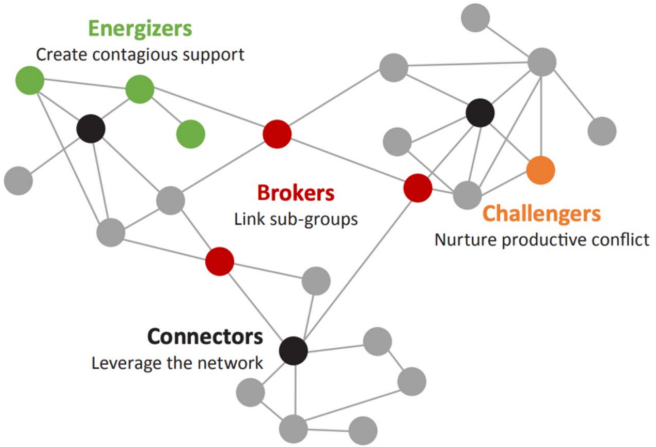


Method



week 10

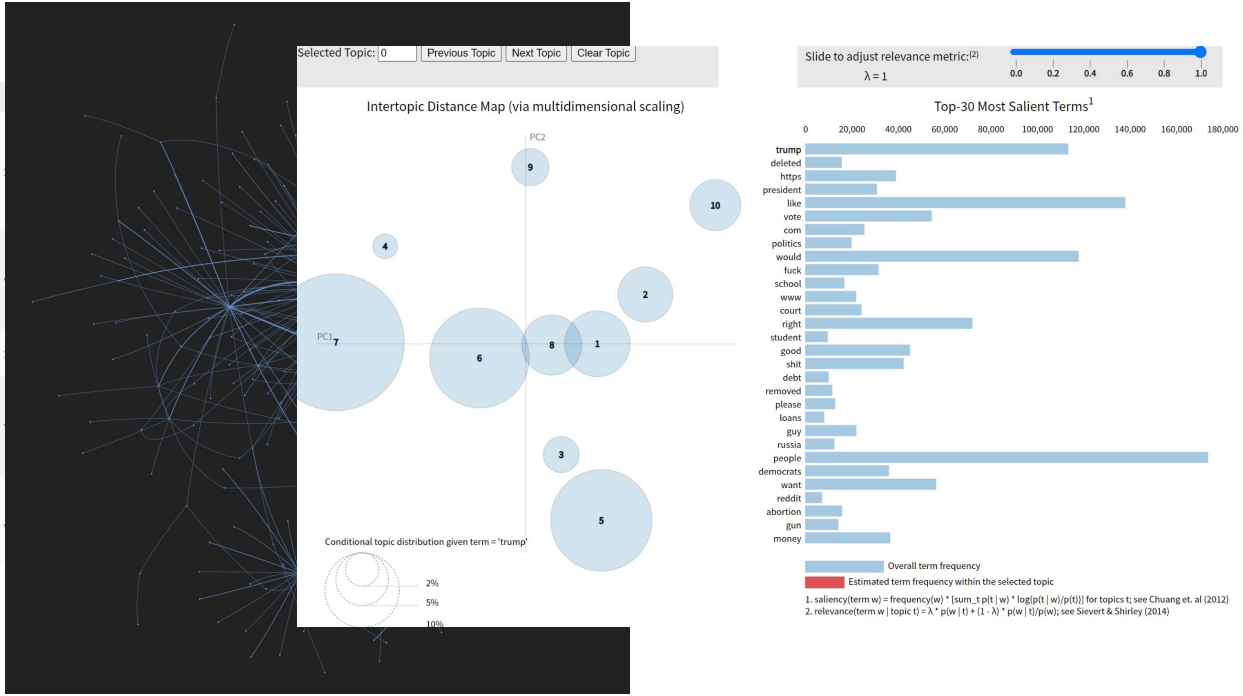
Network Analysis



What is the network type?

Result

body	parent_id	score	subreddit	subreddit_id
MAGA republicans: "Fuck Joe Biden!" ₩n₩nJoe B...	t3_x59p0m	11011	politics	t5_2cneq
T1 diabetic here. Fuck you GOP. Fuck you and...	t3_wihq4u	9258	politics	t5_2cneq
Get fucked Cannon.	t3_xwgido	12843	politics	t5_2cneq
Everything is fun and games until the whole wo...	t3_t4q4nz	7294	politics	t5_2cneq
Get this fucker off the bench. This is ridiculous	t3_vodnck	12812	politics	t5_2cneq

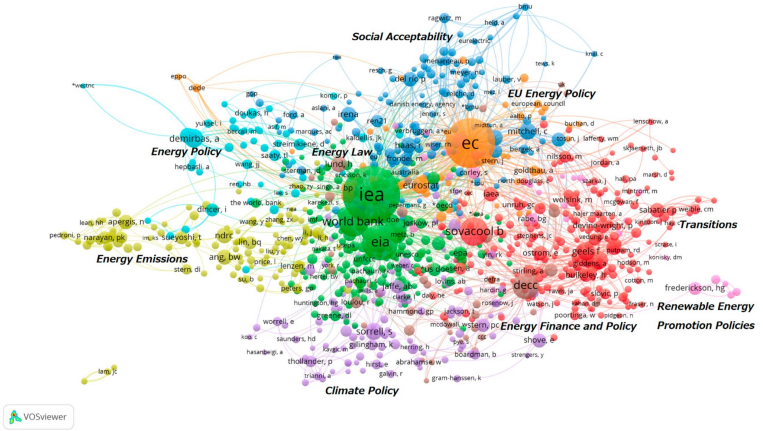


What we learned from our project

- Data cleansing is hard
- Data cleansing is annoying
- But data cleansing is essential

Further study

TF-IDF



Thank you