Twitter Community detection

Unveiling the Twittersphere

IFQIR MOHAMED
SEDJARI YASSINE

Supervised by:
Pr. Mohamed LAZAAR

Outline

- Introduction
- Communities in social media
- Dataset description
- Feature extraction
- Clustering phase
- Results & Experimentation
- Conclusion

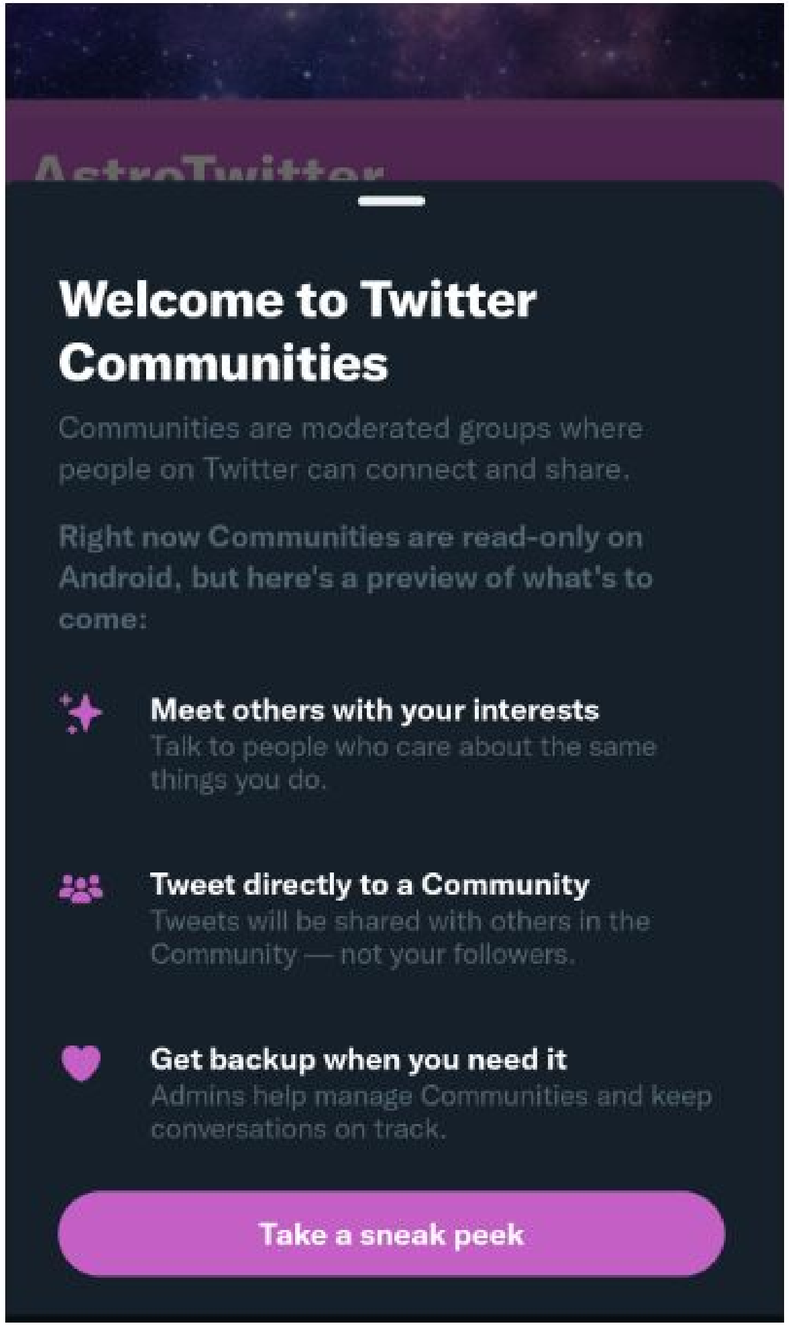
Introduction

Communities in social media

- The social hierarchy of social media users


2k
COMMUNITY


AstroTwitter
AstroTwitter
Nella ✨ @messylines1 · 25s
Taurus sun, Aquarius moon, and Scorpio rising!
ShawtyAstrology @sha... · 4h
I'm super curious.. what's everyone's big 3? (Sun, Moon, and Rising sign).
If you don't know yours, fill out your birth chart information on this website! ...
AstroTwitter
☆kat @thekxt · 1m
I'm so excited about this! This is the future
AstroTwitter
Aquarius ⚡ @aqua
thanks fo
ShawtyAstrology @sha
I'm super curious..
3? (Sun, Moon, and Rising sign).


Welcome to Twitter Communities
Communities are moderated groups where people on Twitter can connect and share.
Right now Communities are read-only on Android, but here's a preview of what's to come:
Meet others with your interests
Talk to people who care about the same things you do.
Tweet directly to a Community
Tweets will be shared with others in the Community — not your followers.
Get backup when you need it
Admins help manage Communities and keep conversations on track.
Take a sneak peek


AstroTwitter
352 Members
Home
About
Nella ✨ @messylines1 · 51s
Taurus sun, Aquarius moon, and Scorpio rising!
ShawtyAstrology ✨ @sha... · 4h
I'm super curious.. what's everyone's big 3? (Sun, Moon, and Rising sign).
If you don't know yours, fill out your birth chart information on this website! ...
☆kat @thekxt · 1m
I'm so excited about this! This is the future
Aquarius ⚡ @aquababe_ · 2m
thanks for the invite!!
ShawtyAstrology ✨ @sha... · 4h

Dataset Description

SNAP

Dataset statistics	
Nodes	81306
Edges	1768149
Nodes in largest WCC	81306 (1.000)
Edges in largest WCC	1768149 (1.000)
Nodes in largest SCC	68413 (0.841)
Edges in largest SCC	1685163 (0.953)

Average clustering coefficient	0.5653
Number of triangles	13082506
Fraction of closed triangles	0.06415
Diameter (longest shortest path)	7
90-percentile effective diameter	4.5

Files hierarchy

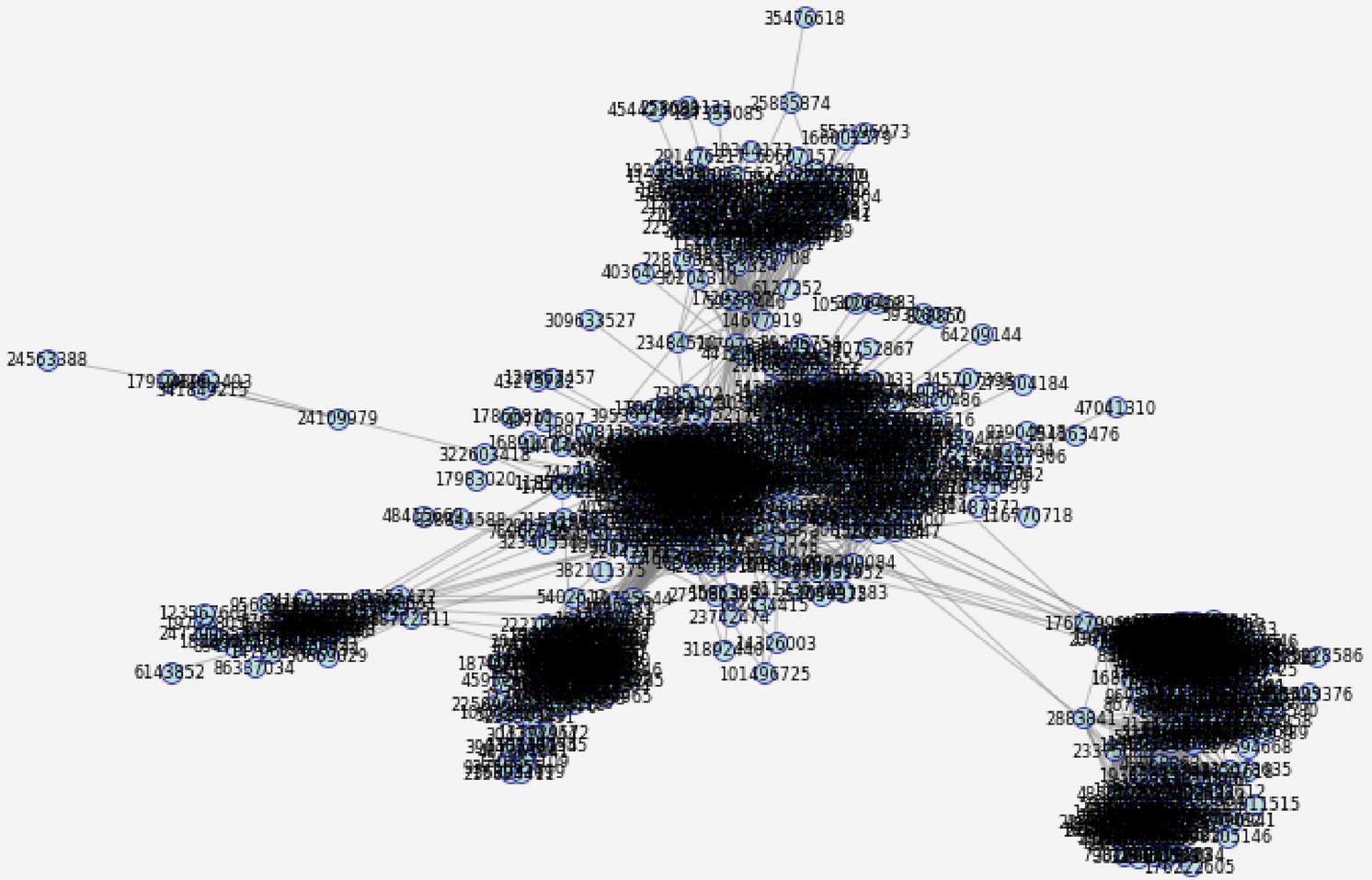
Project directory

- twitter_combined.txt
- Twitter directory

  NodeID.egofeat

  NodeID.featnames

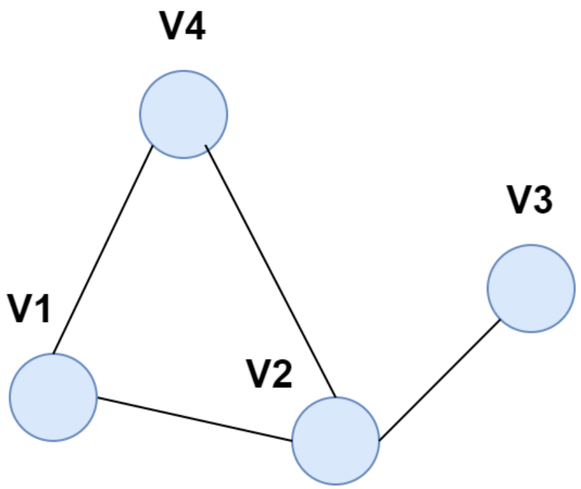
Graph network



Directed graph A--->B

Node A follows Node B

Feature Extraction



$$
\begin{array}{c|cccc}
 & V1 & V2 & V3 & V4 \\
\hline
V1 & 1 & 1 & 0 & 1 \\
V2 & 1 & 1 & 1 & 1 \\
V3 & 0 & 1 & 1 & 0 \\
V4 & 1 & 1 & 0 & 1
\end{array}
$$

Edge-based Approach

Extract Adjacency Matrix

Load the dataset

- Extract the edges from the twitter_combined.txt

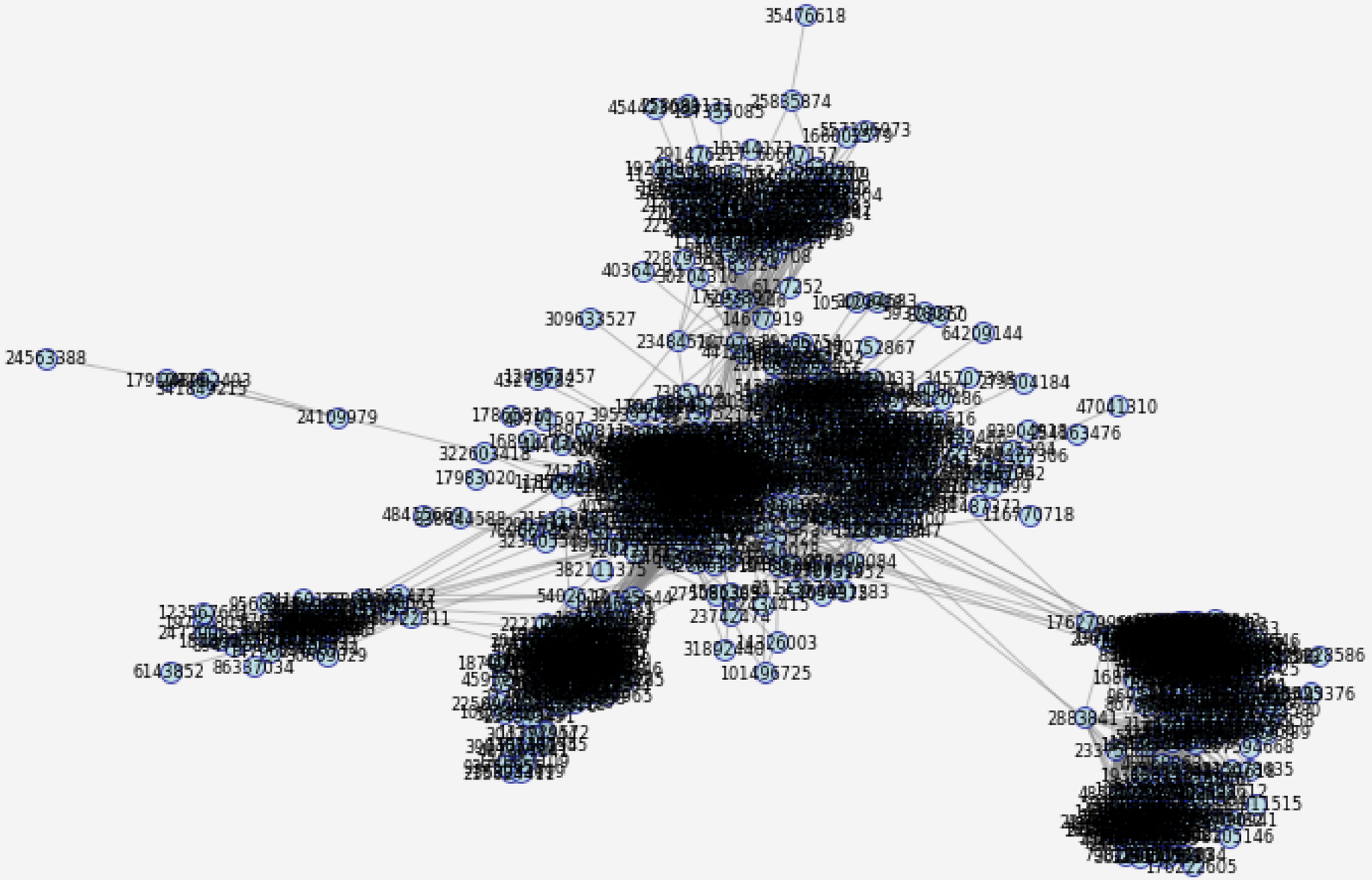
Form the list of edges (Node i ,Node j)

- These edges will be used in the process of creating the adjacency matrix

Construct the Adjacency

- Check the connection between nodes and fill in the adjacency matrix

Graph network



Feature-based approach

Extract Adjacency Matrix

Load the dataset

- Extract the edges from the twitter_combined.txt

Form the list of edges (Node i ,Node j)

- These edges will be used in the process of creating the adjacency matrix

Extract account features from nodeId.feature names

- Check the connection between nodes and fill in the adjacency matrix

Compute degree, closeness and betweenness centrality

Store extracted data in a Dataframe (972,24250)

Final Dataset

	nodeId	#OCTAVIA	#THEHELP	#ff	@BAFTA	@FuckYesEmma	@JUDAOcombr	@astowellcom	@emmastonebr	@helpmovie	...	@chococat
968	14528221	0	0	1	0	0	0	0	0	0	...	0
969	14840869	0	0	0	0	0	0	0	0	0	...	0
970	82726142	0	0	0	0	0	0	0	0	0	...	0
971	255790981	0	0	0	0	0	0	0	0	0	...	0
972	36618690	0	0	0	0	0	0	0	0	0	...	1

@helpmovie	...	@chococat	@dannyBstyle,	@jon_blaze55	@kylepulver:	@terrycavanagh	@twitchtv.	Degree Centrality	Closeness Centrality	Betweenness Centrality
0	...	0	0	0	0	0	0	0.001030	0.238165	0.000000
0	...	0	0	0	0	0	0	0.003090	0.239990	0.001701
0	...	0	0	0	0	0	0	0.002060	0.206245	0.000958
0	...	0	0	0	0	0	0	0.009269	0.273984	0.002988
0	...	1	1	1	1	1	1	0.010299	0.294064	0.010228

Clustering Phase

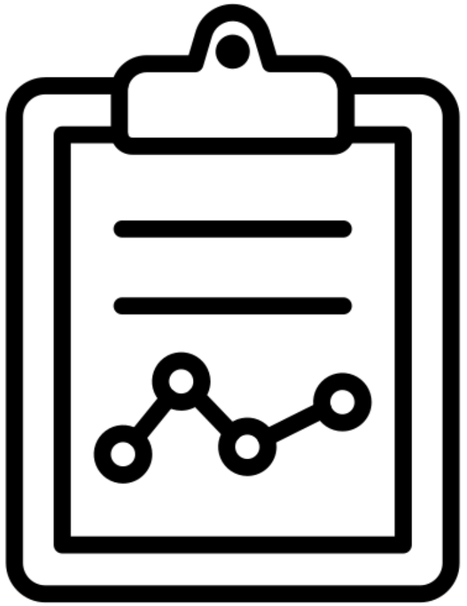
Edge-based Approach

Parameter	Value
Number of clusters	3
Affinity	Precomputed
Algorithm	Spectral Clustering

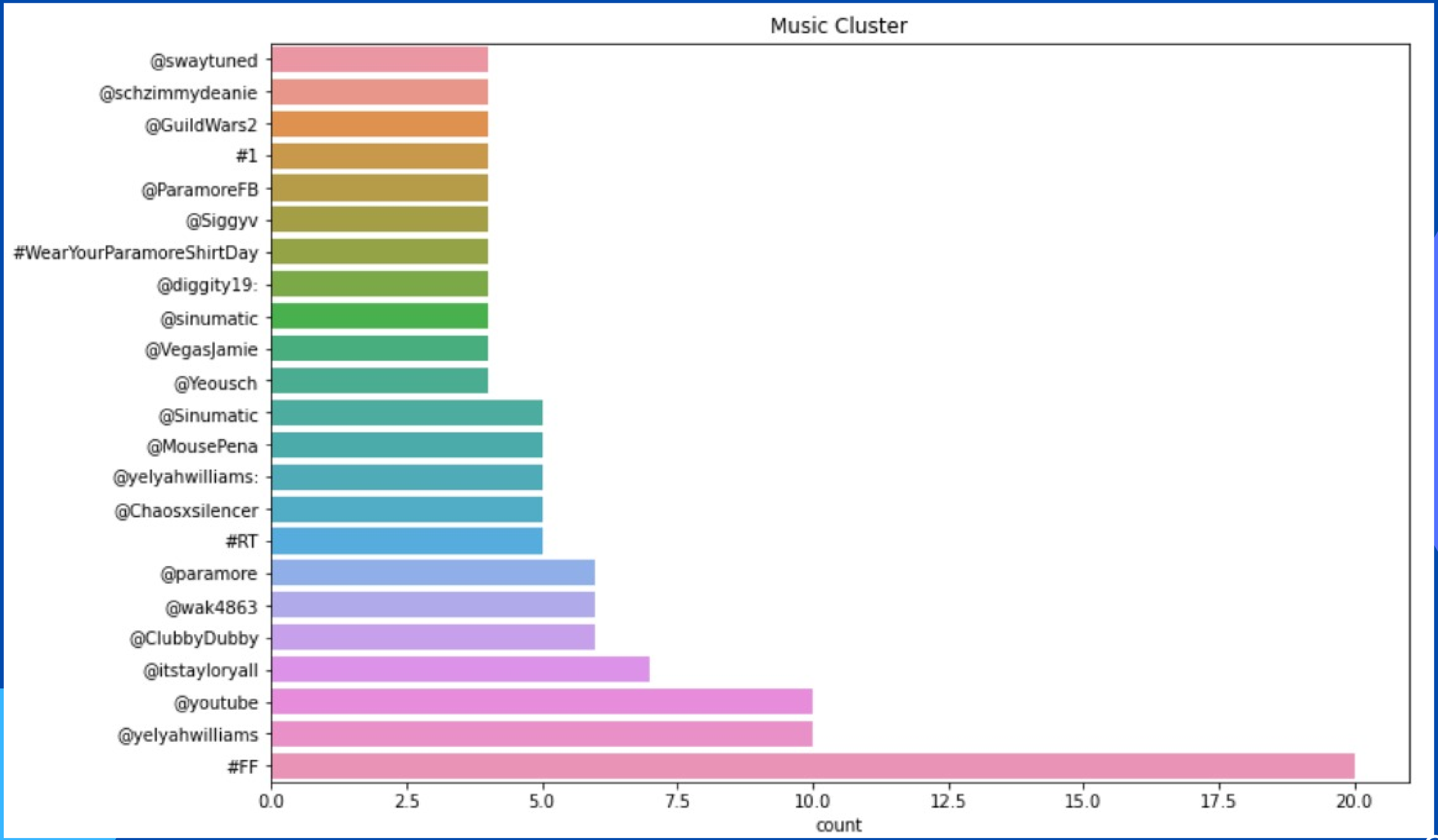
Feature-based approach

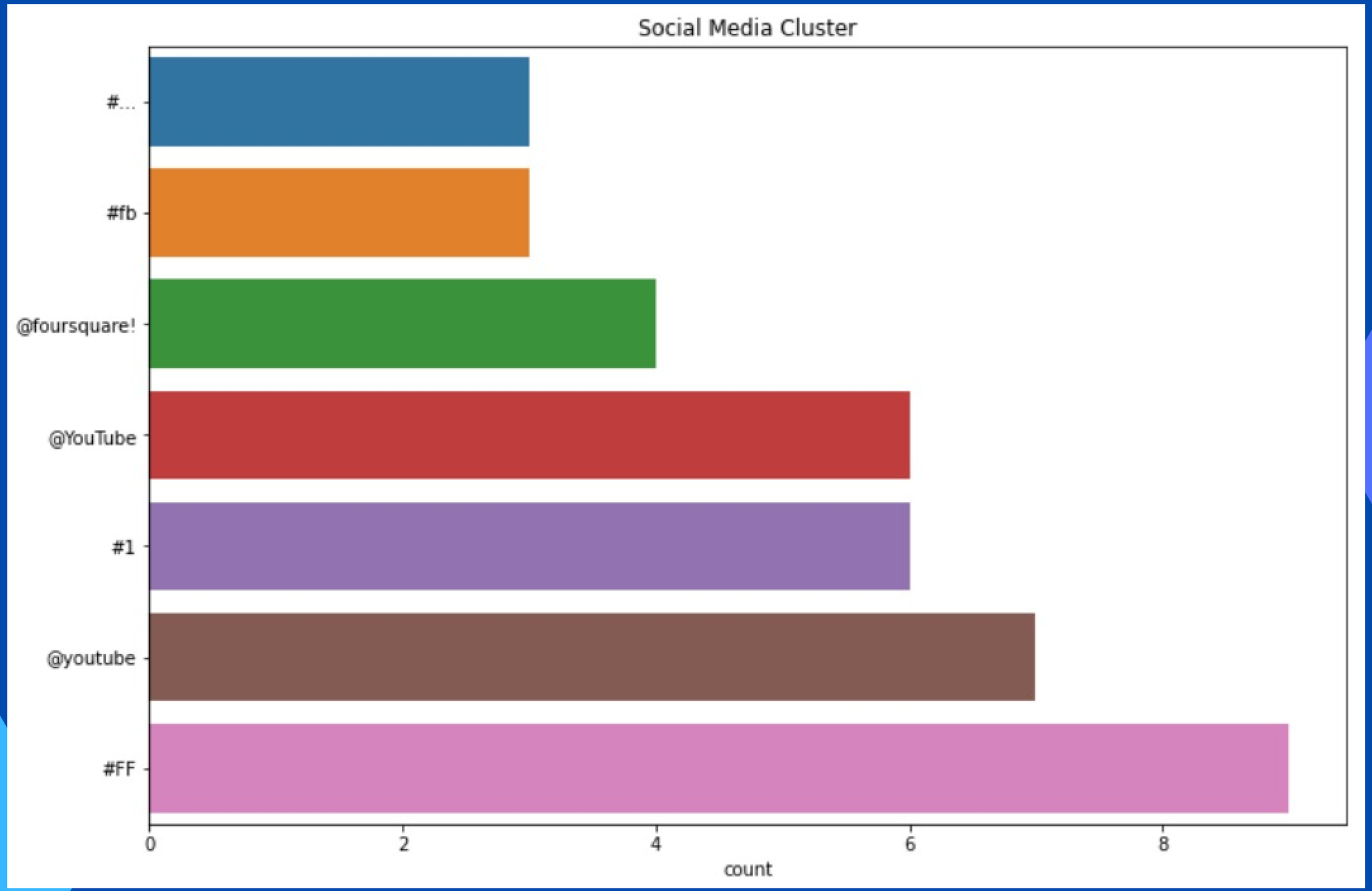
Algorithm	Number of Clusters	Data	Random Seed
k-means	3	df	42
hierarchical	3	df	-
spectral	3	df	42

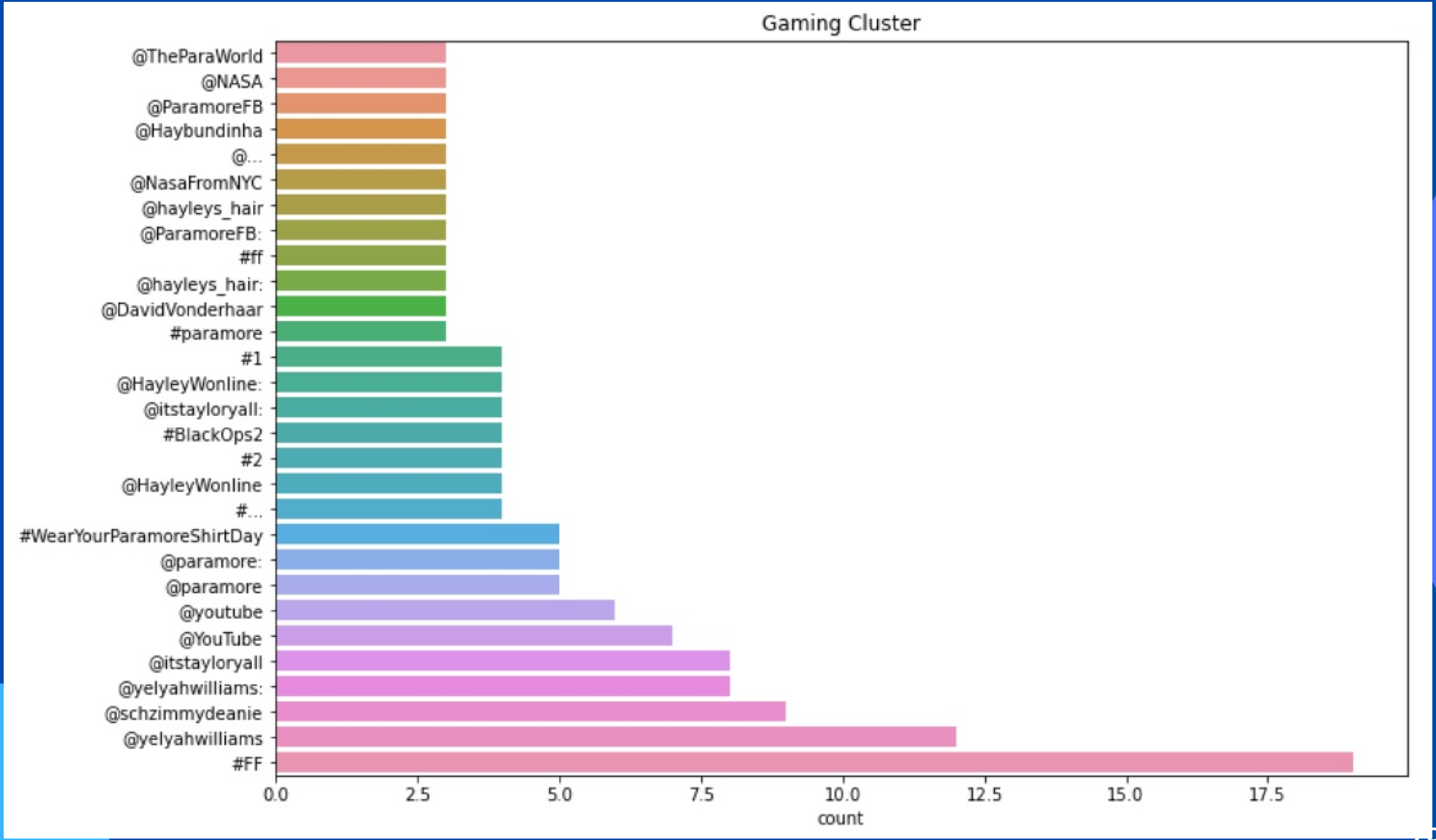
Results & Experimentation



Method	Clustering Algorithm	Silhouette Score
Method 1: Edge-based Approach	Spectral Clustering	-0.1887
Method 2: Feature-based Approach	Spectral Clustering	-0.0343
	Hierarchical Clustering	0.6756
	K-MEANS	0.6917


Music Cluster
@swaytuned
@schzimmydeanie
@GuildWars2
#1
@ParamoreFB
@Siggyv
#WearYourParamoreShirtDay
@diggity19:
@sinumatic
@VegasJamie
@Yeousch
@Sinumatic
@MousePena
@yelyahwilliams:
@Chaosxsilencer
#RT
@paramore
@wak4863
@ClubbyDubby
@itstayloryall
@youtube
@yelyahwilliams
#FF
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
count


Social Media Cluster
#...
#fb
@foursquare!
@YouTube
#1
@youtube
#FF
0
2
4
6
8
count


Gaming Cluster
@TheParaWorld
@NASA
@ParamoreFB
@Haybundinha
@...
@NasaFromNYC
@hayleys_hair
@ParamoreFB:
#ff
@hayleys_hair:
@DavidVonderhaar
#paramore
#1
@HayleyWonline:
@itstayloryall:
#BlackOps2
#2
@HayleyWonline
#...
#WearYourParamoreShirtDay
@paramore:
@paramore
@youtube
@YouTube
@itstayloryall
@yelyahwilliams:
@schzimmydeanie
@yelyahwilliams
#FF
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
count

Conclusion

Thank you for your attention

Feel free to ask questions