Business intelligence per big data

PROGETTO DI ANALISI DEI DATI

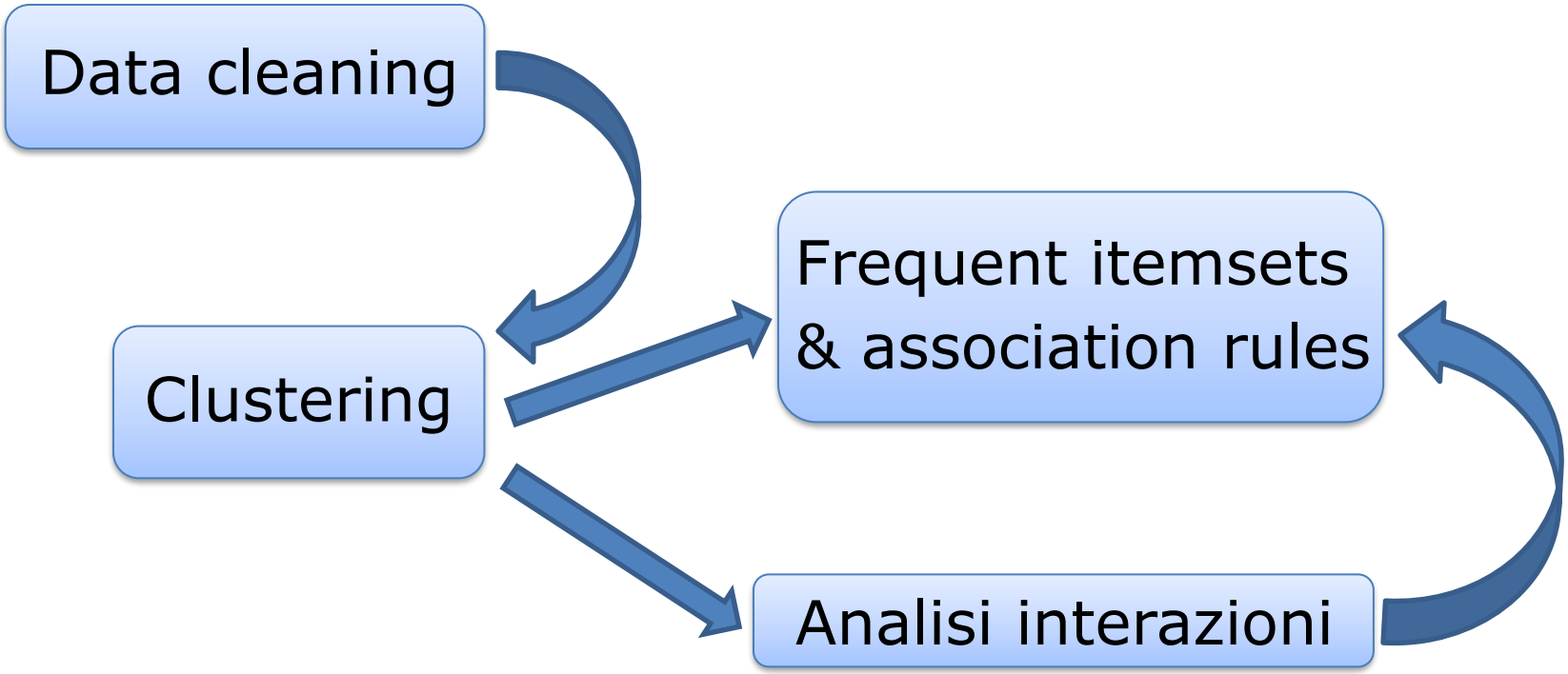
ANALISI DELLE INTERAZIONI SU DATABASE DI TWEET A TEMA COVID-19

DOCENTE: Tania Cerquitelli

GRUPPO 30: Enrico Agrippino, Sara Giovannini

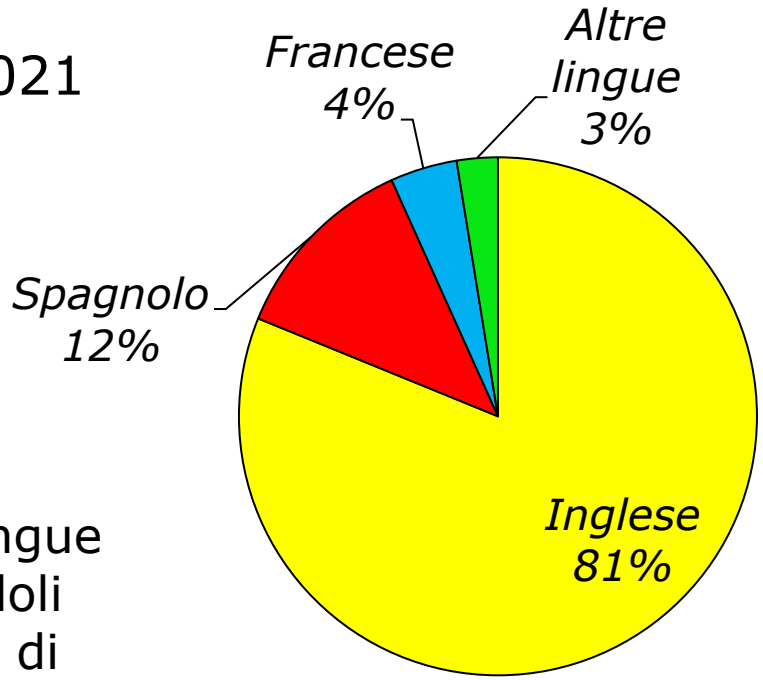
ANNO SCOLASTICO 2020-2021

PROCESSO DI ANALISI



DATA EXPLORATION

- 4000 tweet
- Settimana dal 1 al 8 Gennaio 2021
- Tweet in diverse lingue



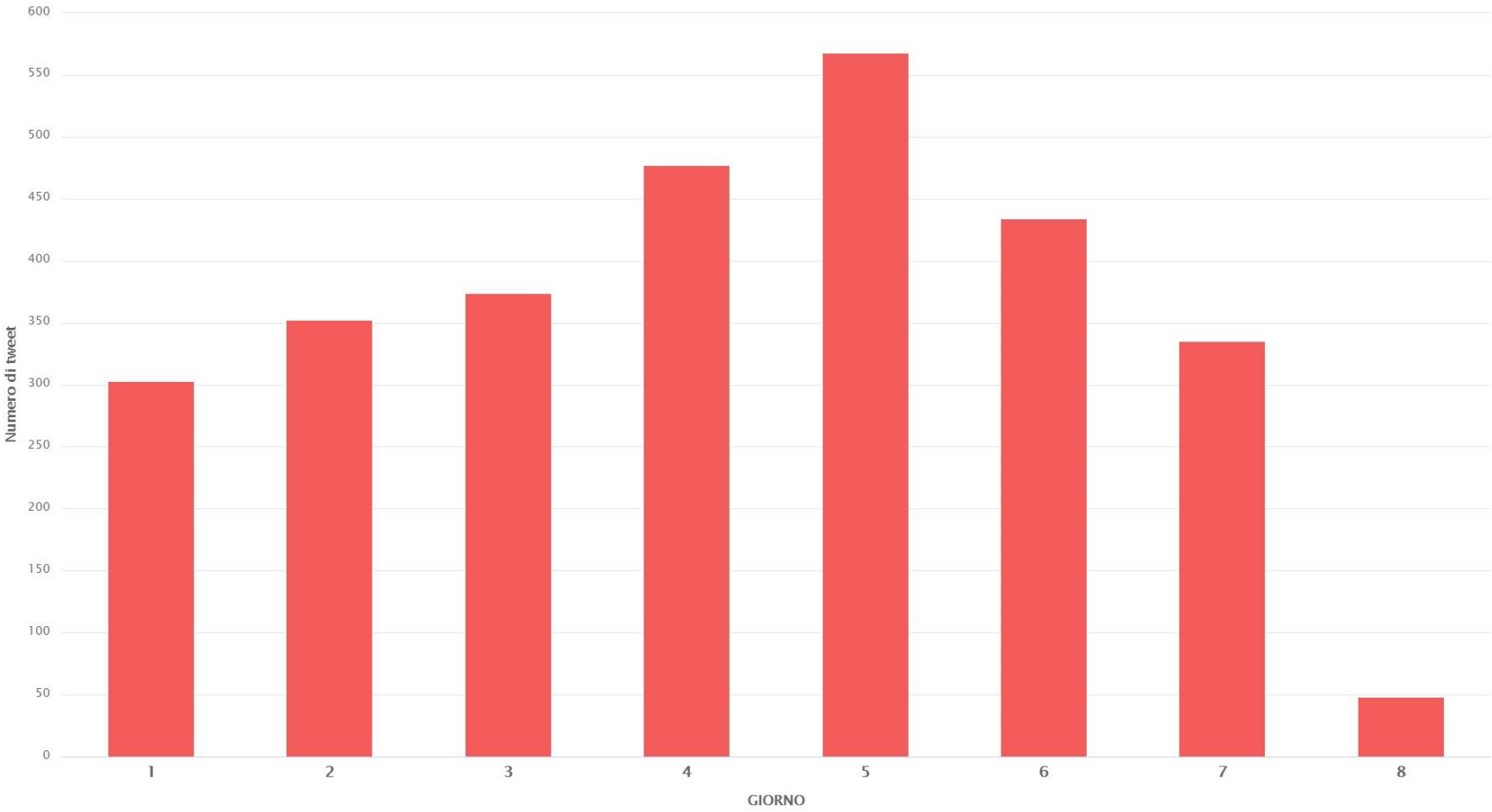
Language detection:

1. Si selezionano i tweet con 'Language=en'.
2. Si rimuovono i tweet di altre lingue etichettati come inglese, avendoli identificati grazie alla presenza di accenti o di espressioni emblematiche.

→ 2891 tweet selezionati

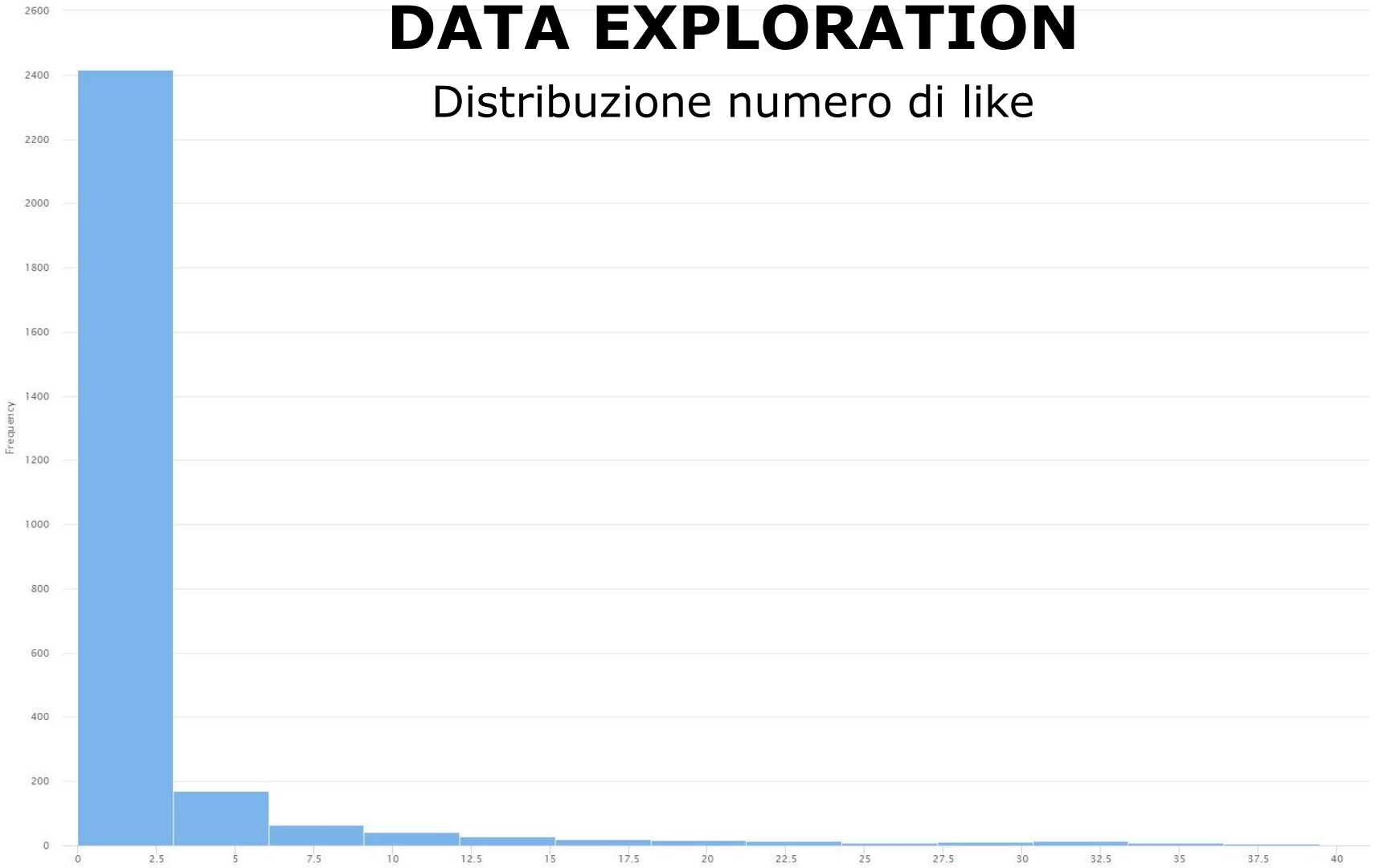
DATA EXPLORATION

Distribuzione tweet per giorno



DATA EXPLORATION

Distribuzione numero di like



PREPROCESSING

- Viene eliminato l'attributo IS_RETWEET: nel dataset assume solo il valore FALSE.

- Vengono prima creati gli attributi binari **HAS_LINK**, **HAS_TAG**, poi eventuali link e tag vengono rimossi dai tweet.

- **Process Document**: viene creata la matrice TF-IDF con pruning percentuale >0.35%.

  1. Rimozione *stopwords.*
  2. Rimozione dei token COVID, CORONAVIRUS, COVID19, CORONA.
  3. Stem (Snowball): i token vengono ricondotti alla radice.

CLUSTERING: DBSCAN

Problemi:

1) Alta cardinalità del dataset → Filtro su giorno

2) Elevato numero di attributi → Aumento soglia TF-IDF

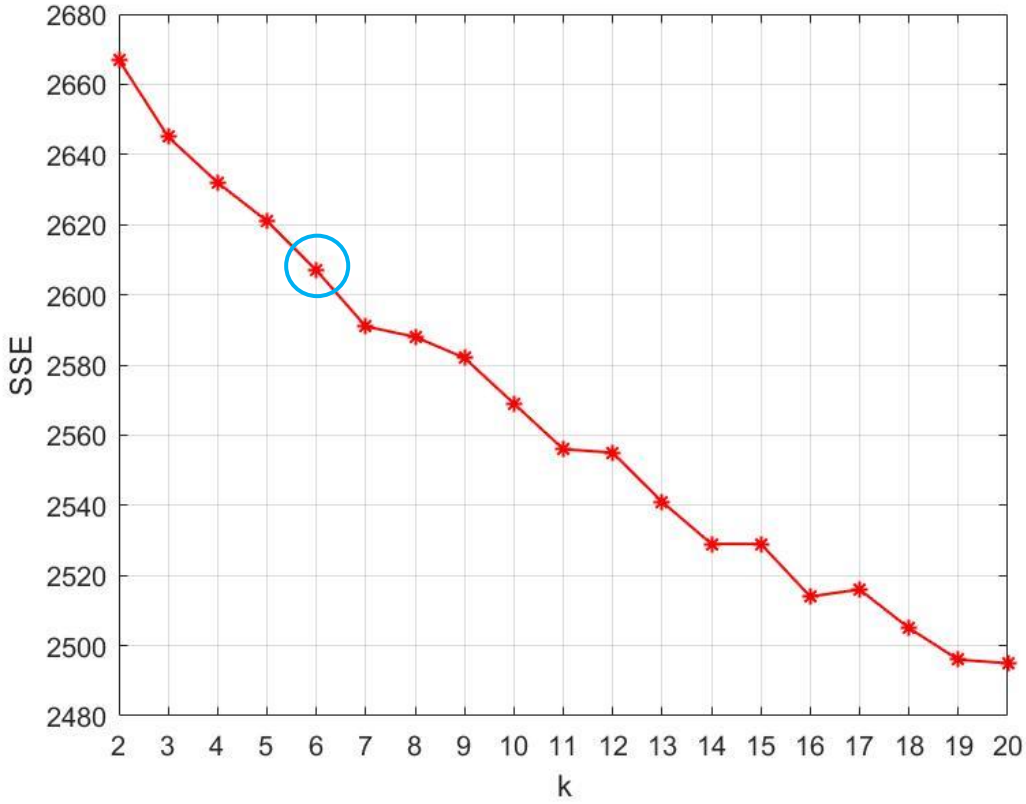
Risultati deludenti:

1) L'algoritmo riconosce solo il noise.

2) L'algoritmo riesce nella creazione di clusters ma questi sono di scarsa qualità: argomenti e dimensioni disomogenee.

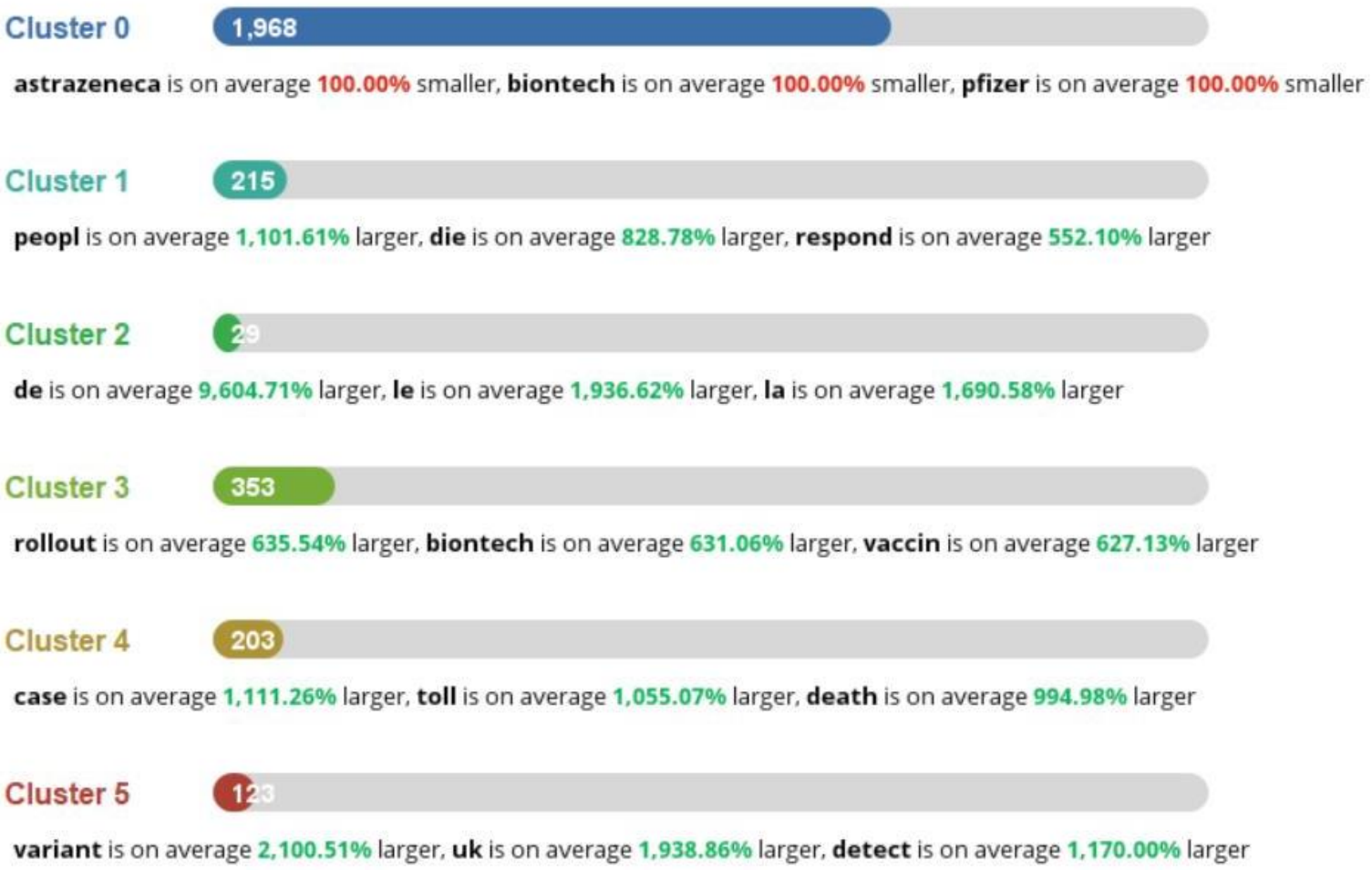
CLUSTERING: k-means

Vengono calcolati gli SSE al variare di k utilizzando la distanza euclidea.

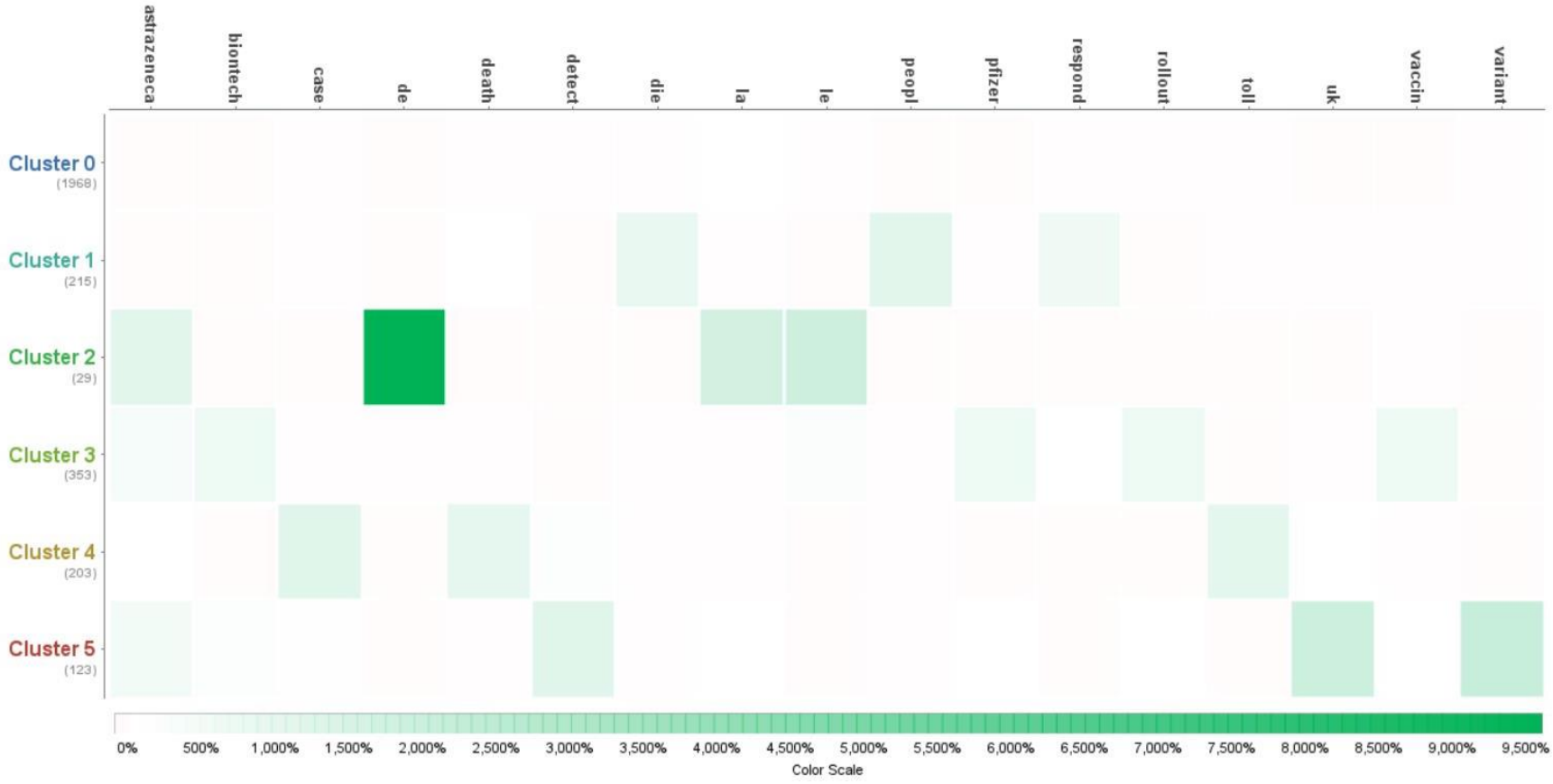


Si sceglie k=6.

OVERVIEW DEI CLUSTER OTTENUTI /1



OVERVIEW DEI CLUSTER OTTENUTI /2



ANALISI MACRO DIFFERENZE FRA CLUSTERS

Cluster	Like	Retweet	Has link	Has tag	Lunghezza media	Numero di tweet
0	7.67	1.5	82%	31%	131.9	1968
1	14.8	4.76	74%	35%	173.6	215
3	4.3	1.37	83%	29%	147.7	353
4	4.6	1.58	79%	19%	144.8	203
5	6.1	2.1	85%	37%	157.8	123
Dataset	7.45	1.78	82%	30%	138.5	2891

CLUSTER 1:

COVID & POLITICA

CLUSTER 1: FREQUENT ITEMSET

ITEMSET	Supporto
people	0.860
die	0.209
people, get	0.126
people, die	0.098
vaccin	0.093
govern	0.056
death	0.051
infect	0.051
help	0.047
spread	0.042

CLUSTER 1: ASSOCIATION RULES

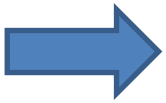
Premessa → Conclusione	Supporto	Confidenza	Lift
people, get → vaccin	0.033	0.259	2.787
work → die	0.023	0.833	3.981
get→ infect	0.023	0.147	2.874
investig → die	0.019	0.800	3.820
wear → mask	0.019	0.667	20.476
people, elect → why	0.014	0.600	11.72
house → white	0.014	0.500	17.91
health → people, risk	0.014	0.375	11.518

Filtro per numero di like e retweet

Alcuni degli itemset più significativi per i tweet che hanno un numero di **like** e **retweet** maggiori della **media** del cluster:

FAVOURITE_COUNT>=14

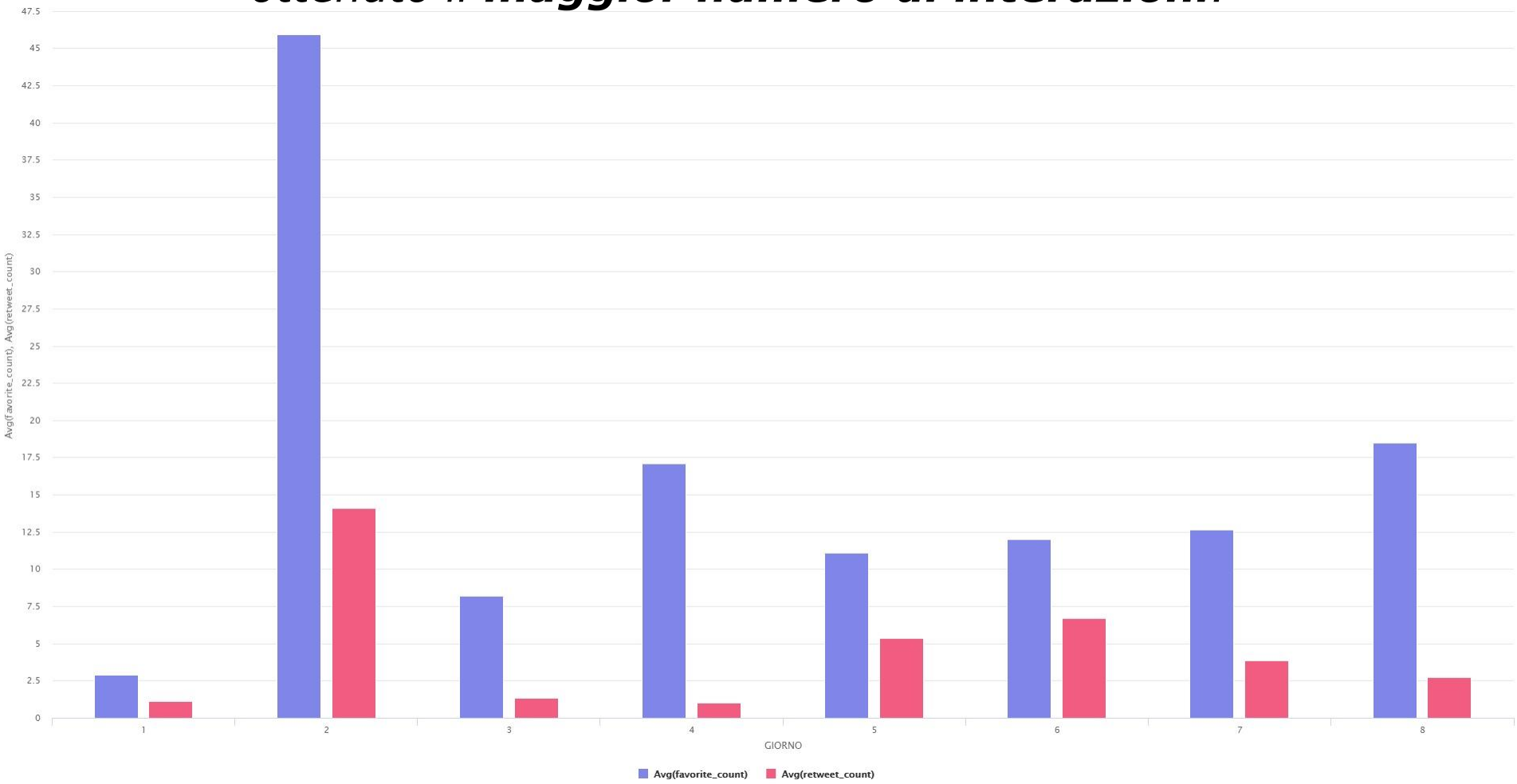
RETWEET_COUNT>=4

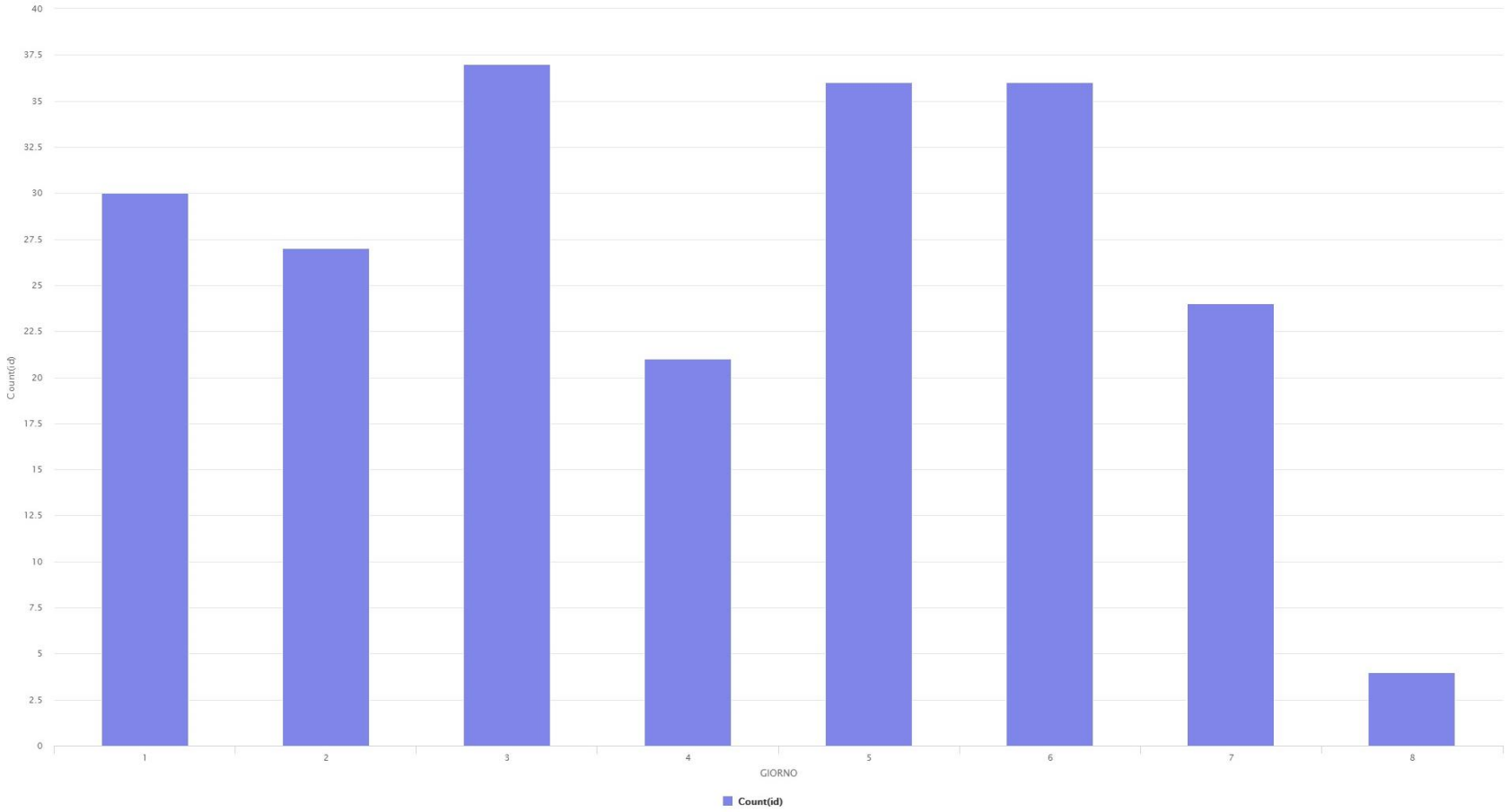


16 tweet

- Holiday (supporto=0.125)
- Leave (0.125)
- Read (0.125)

*Si considerano ora i tweet della **giornata** che ha ottenuto il **maggior numero di interazioni**.*





*I tweet selezionati (**giorno 2**) sono **27**.*

Cluster 1, 2 gennaio 2021: FREQUENT ITEMSET

ITEMSET	Supporto	Differenza	Differenza %
people	0.815	-0.045	-5.52%
die	0.222	0.013	5.86%
people, get	0.185	0.059	31,89%
govern	0.148	0.092	62.16
republican	0.111	0.088	79,3%
ask	0.074	0.060	81.08%
fact	0.074	0.055	74,32%
leader	0.074	0.065	87.84%

CLUSTER 3:

VACCINI

CLUSTER 3: FREQUENT ITEMSET

ITEMSET	Supporto
vaccin	1
get	0.159
health	0.096
dose	0.091
receive	0.068
world	0.062
india	0.057

ITEMSET	Supporto
know	0.051
country	0.045
rollout	0.045
pfizer	0.042
approv	0.040
health, worker	0.025
astrazeneca	0.023

CLUSTER 3: ASSOCIATION RULES

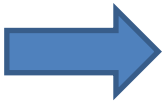
Premessa → Conclusione	Supporto	Confidenza	Lift
worker→ health, vaccin	0.025	0.750	7.787
approve → india, vaccin	0.023	0.571	10.086
biontech → pfizer, vaccin	0.023	0.889	20.92
astrazeneca → rd, vaccin	0.020	0.875	30.89
biotech → bharat, india, vaccin	0.017	0.750	44.125
israel → vaccin, word	0.014	0.455	7.29
please → get, vaccin	0.011	0.571	3.60
refuse, vaccin → health, worker	0.011	0.800	31.38
health, care → vaccin, worker, refuse	0.011	0.571	50.43
leader → vaccin, world	0.011	0.571	9.17
slow → vaccin, rollout	0.011	0.800	17.65

Filtro per numero di like e retweet

Alcuni degli itemset più significativi per i tweet che hanno un numero di **like** e **retweet** maggiori della **media** del cluster:

FAVOURITE_COUNT>=4

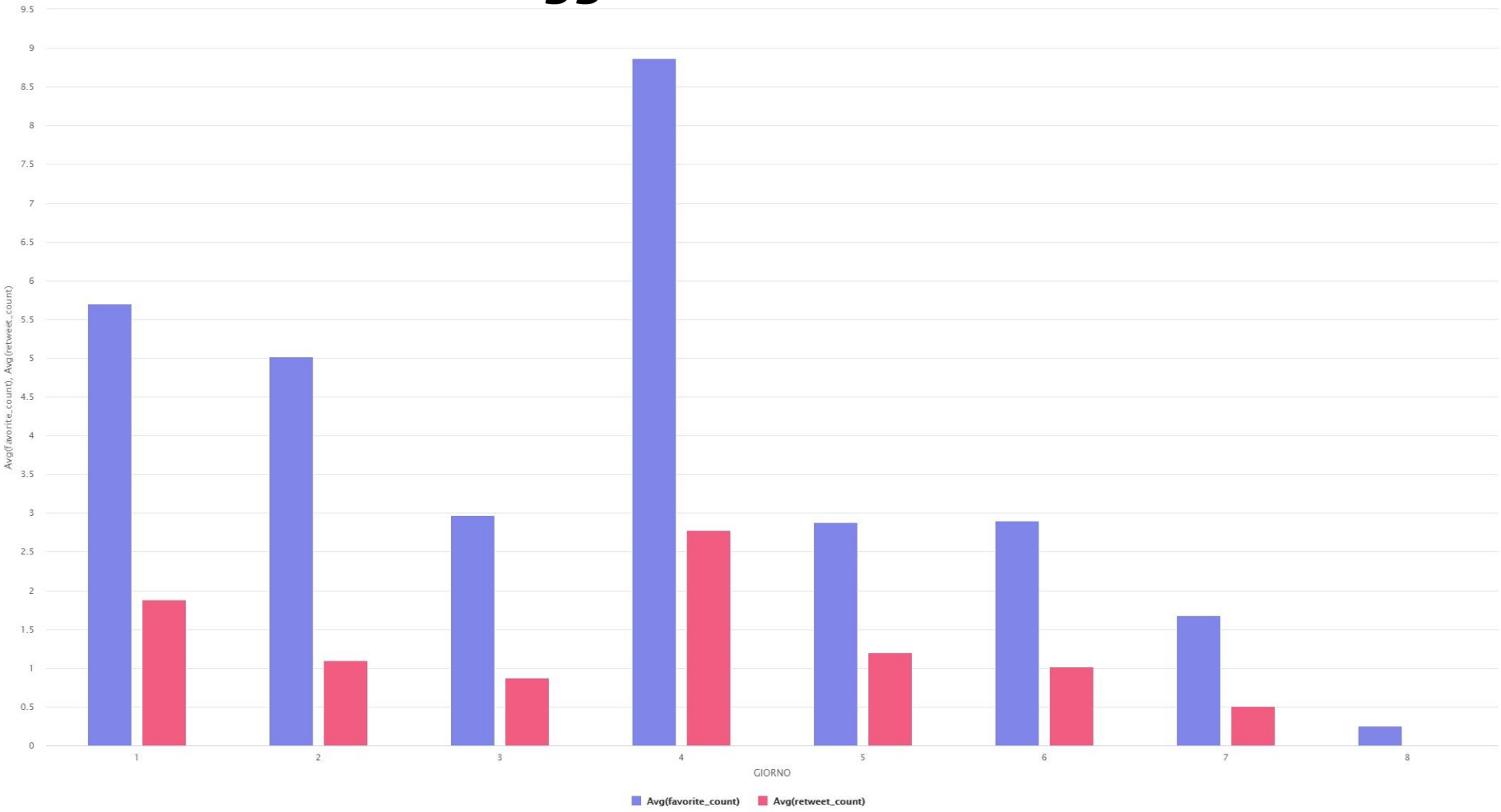
RETWEET_COUNT>=1

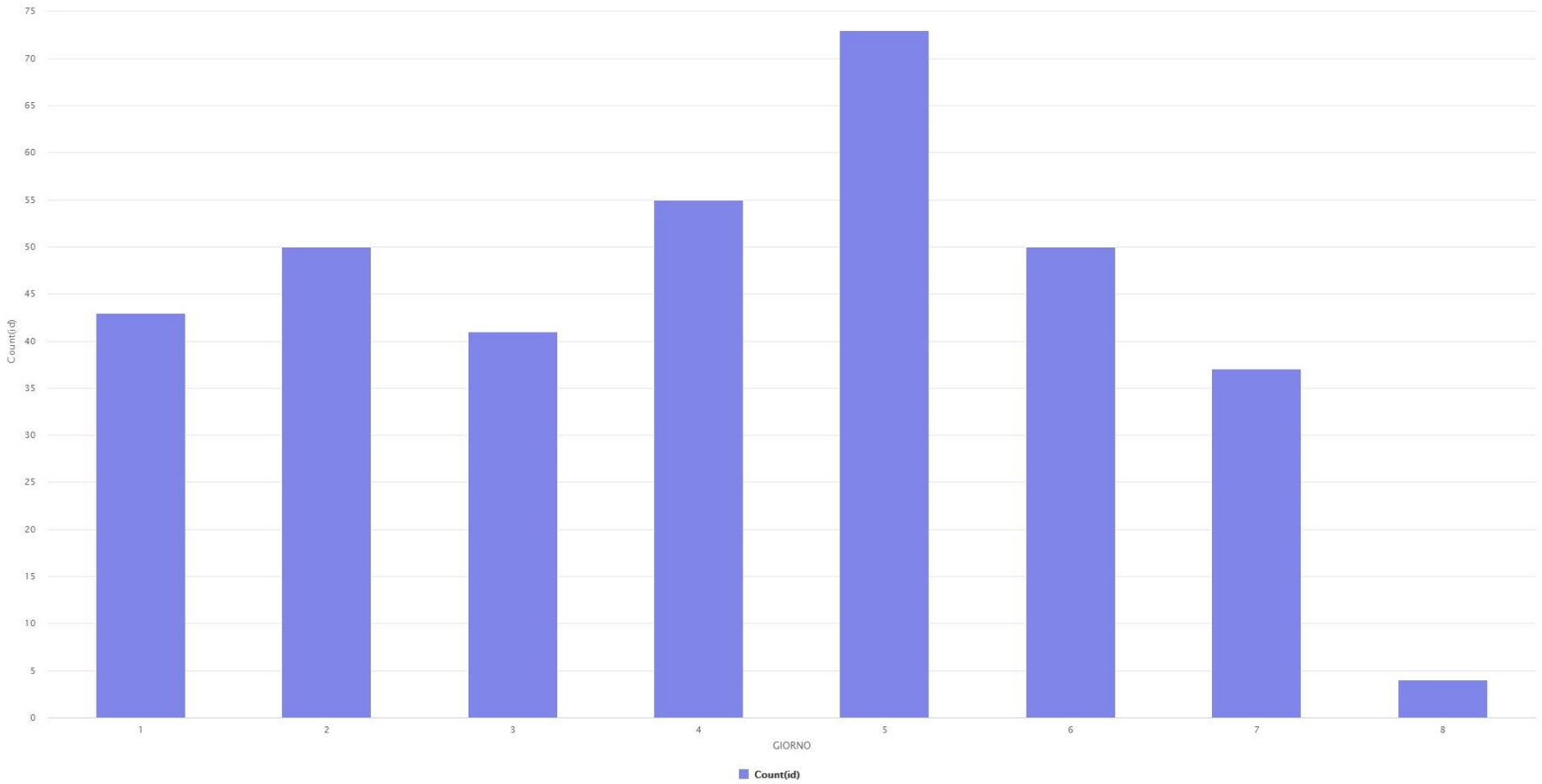


60 tweet

- Receive vaccin (supporto=0.133)
- Govern (0.083)
- Start (0.083)
- Effort (0.067)

*Si considerano ora i tweet della **giornata** che ha ottenuto il **maggior numero di interazioni**.*





*I tweet selezionati (**giorno 4**) sono **55**.*

Cluster 3, 4 gennaio 2021: FREQUENT ITEMSET

ITEMSET	Supporto	Differenza	Differenza %
vaccin	1	0	0%
get	0.182	0.023	12.64%
approve	0.091	0.051	56.04%
day	0.091	0.057	62.64%
dose	0.091	0	0%
health	0.091	-0.005	-5.49%
india	0.091	0.034	37.36%
pfizer	0.091	0.049	53.85%
know	0.091	0.040	43.96%
bharat biotech	0.073	0.048	65.75%

CLUSTER 4:

BOLLETTINI & NOTIZIE

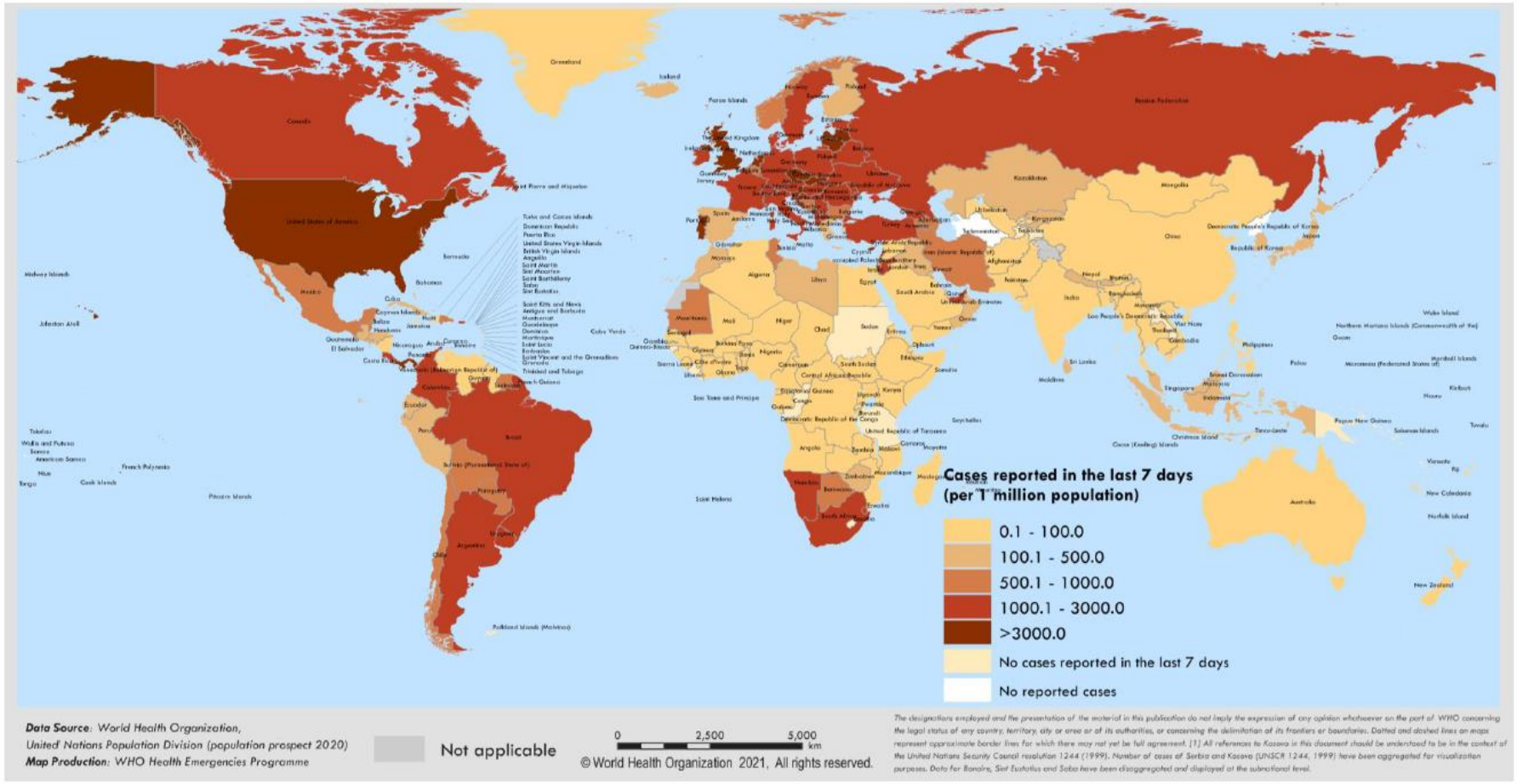
CLUSTER 4: FREQUENT ITEMSET

ITEMSET	Supporto
case	0.783
death	0.443
report	0.197
record	0.172
case, death	0.227
case, report	0.177
death, report	0.118
total	0.113
daily	0.089

ITEMSET	Supporto
activ, case	0.074
posit, case	0.074
daili, case	0.074
death, toll	0.064
hospit	0.069
case, confirm	0.059
test	0.054
case, hour	0.044
posit, test	0.034

CLUSTER 4: ASSOCIATION RULES

Premessa → Conclusione	Supporto	Confidenza	Lift
report→ case, death	0.099	0.500	2.21
uk→ death	0.044	0.900	2.03
india → total	0.030	1	8.83
nigeria → case, record	0.025	1	6.55
florida → case, report	0.023	0.889	6.64



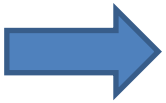
*Casi **COVID-19** su 1 milione di abitanti registrati nella settimana dal **28/12/2020** al **03/01/2021***

Filtro per numero di like e retweet

Alcuni degli itemset più significativi per i tweet che hanno un numero di **like** e **retweet** maggiori della **media** del cluster:

FAVOURITE_COUNT>=4

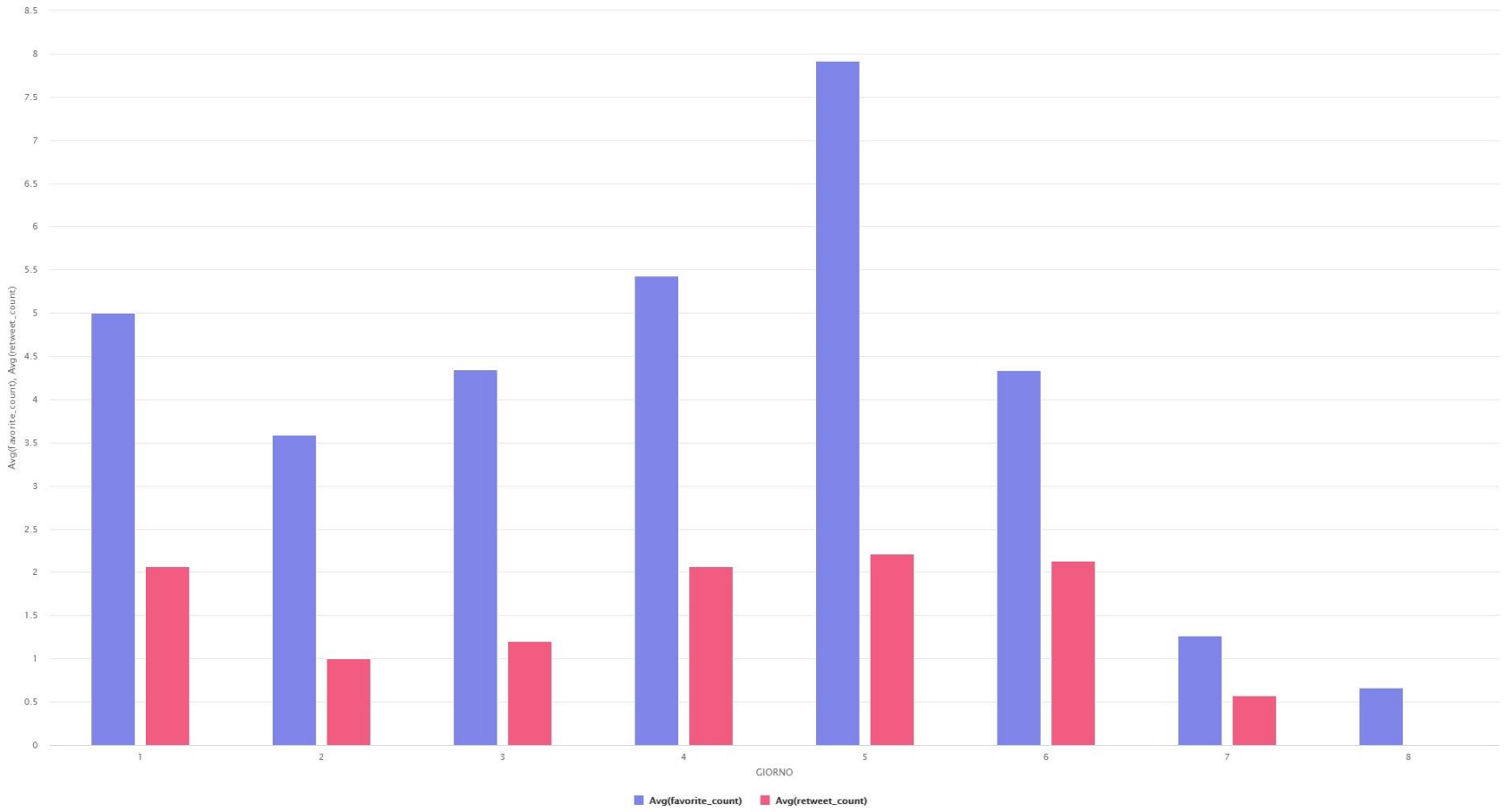
RETWEET_COUNT>=1

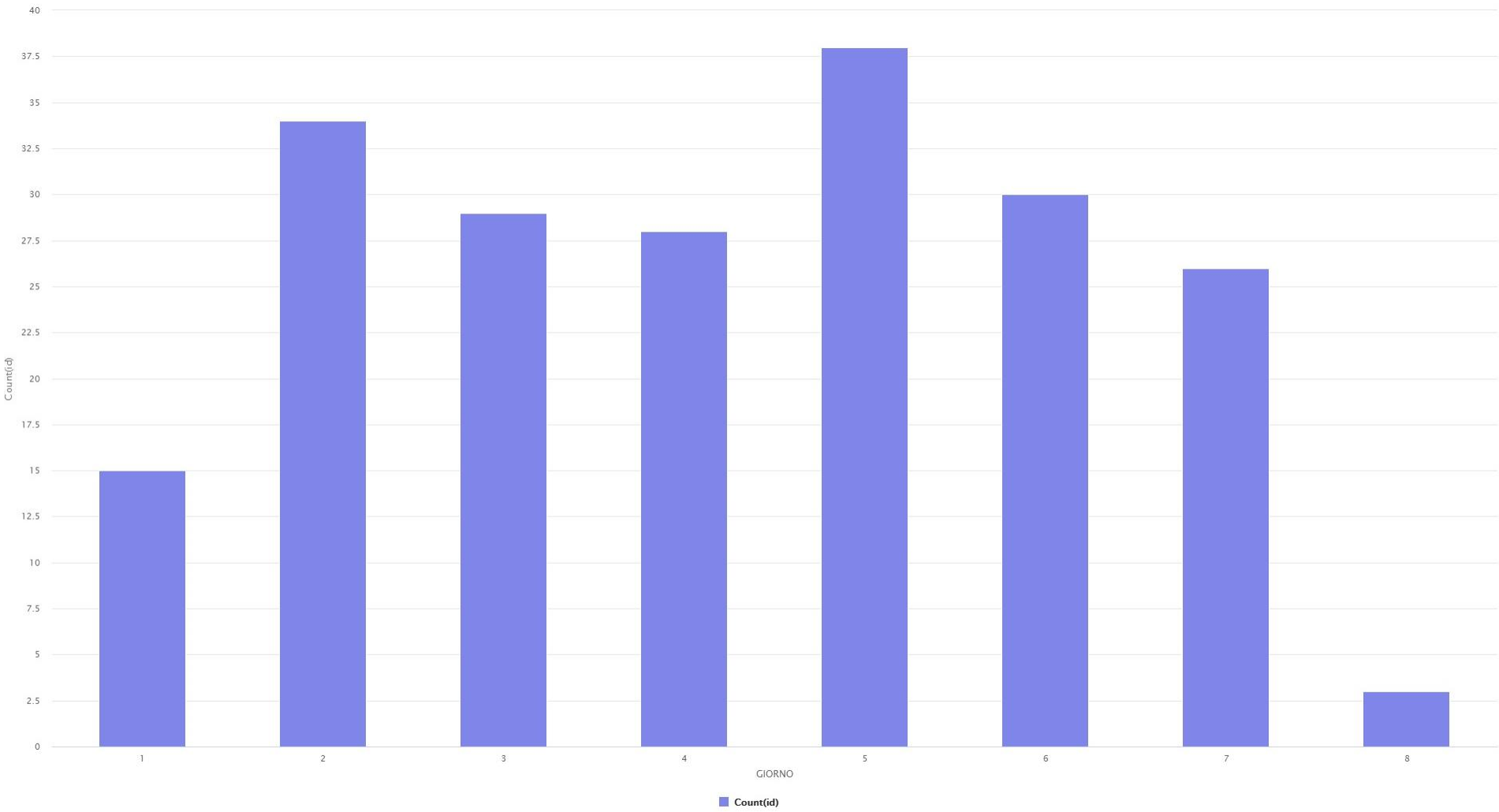


34 tweet

- Case death report (supporto=0.235)
- Health (0.206)
- Break (0.176)
- Additional (0.118)

*Si considerano ora i tweet della **giornata** che ha ottenuto il **maggior numero di interazioni**.*





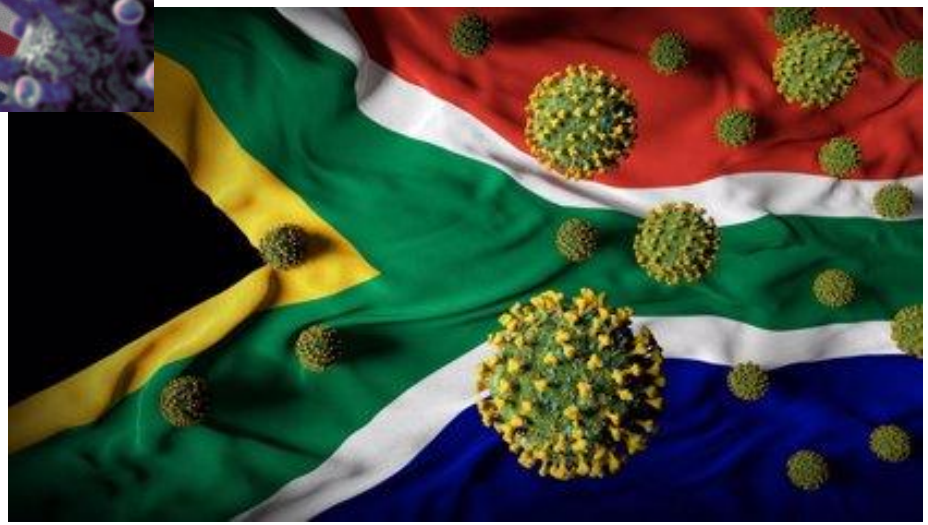
*I tweet selezionati (**giorno 5**) sono **38**.*

Cluster 4, 5 gennaio 2021: FREQUENT ITEMSET

ITEMSET	Supporto	Differenza	Differenza %
case	0.921	0.138	14.98%
death	0.368	-0.075	-20.38%
case, death	0.289	0.062	21.45%
Report	0.263	0.066	25.10%
january	0.132	0.088	66.67%
posit	0.132	0.058	43.94%
activ	0.105	0.031	29.52%
confirm	0.105	0.046	43.81%
pandem	0.105	0.051	48.57%
nigeria	0.079	0.054	68.35%

CLUSTER 5:

VARIANTI



CLUSTER 5: FREQUENT ITEMSET

ITEMSET	Supporto
uk	0.707
variant	0.407
vaccin	0.171
case	0.138
lockdown	0.138
strain	0.114
uk, variant	0.114
news	0.106
uk, vaccin	0.106

ITEMSET	Supporto
variant, vaccin	0.106
uk lockdown	0.089
uk, news	0.065
uk, travel	0.065
variant, south	0.065
variant, spread	0.049
scientist	0.049
transmiss	0.049
variant, south, africa	0.041

CLUSTER 5: ASSOCIATION RULES

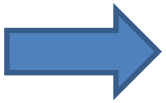
Premessa → Conclusione	Supporto	Confidenza	Lift
vaccin → variant	0.106	0.619	1.523
test → uk	0.057	0.875	1.237
spread → variant	0.049	0.667	1.640
transmiss → variant	0.049	1	2.460
worri → variant	0.033	0.800	1.968
infecti → variant	0.033	1	2.460
vaccin, scientist → variant	0.033	1	2.460
vaccin, variant → scientist	0.033	0.308	6.308
variant, uk → health	0.033	0.286	3.905

Filtro per numero di like e retweet

Alcuni degli itemset più significativi per i tweet che hanno un numero di **like** e **retweet** maggiori della **media** del cluster:

FAVOURITE_COUNT>=6

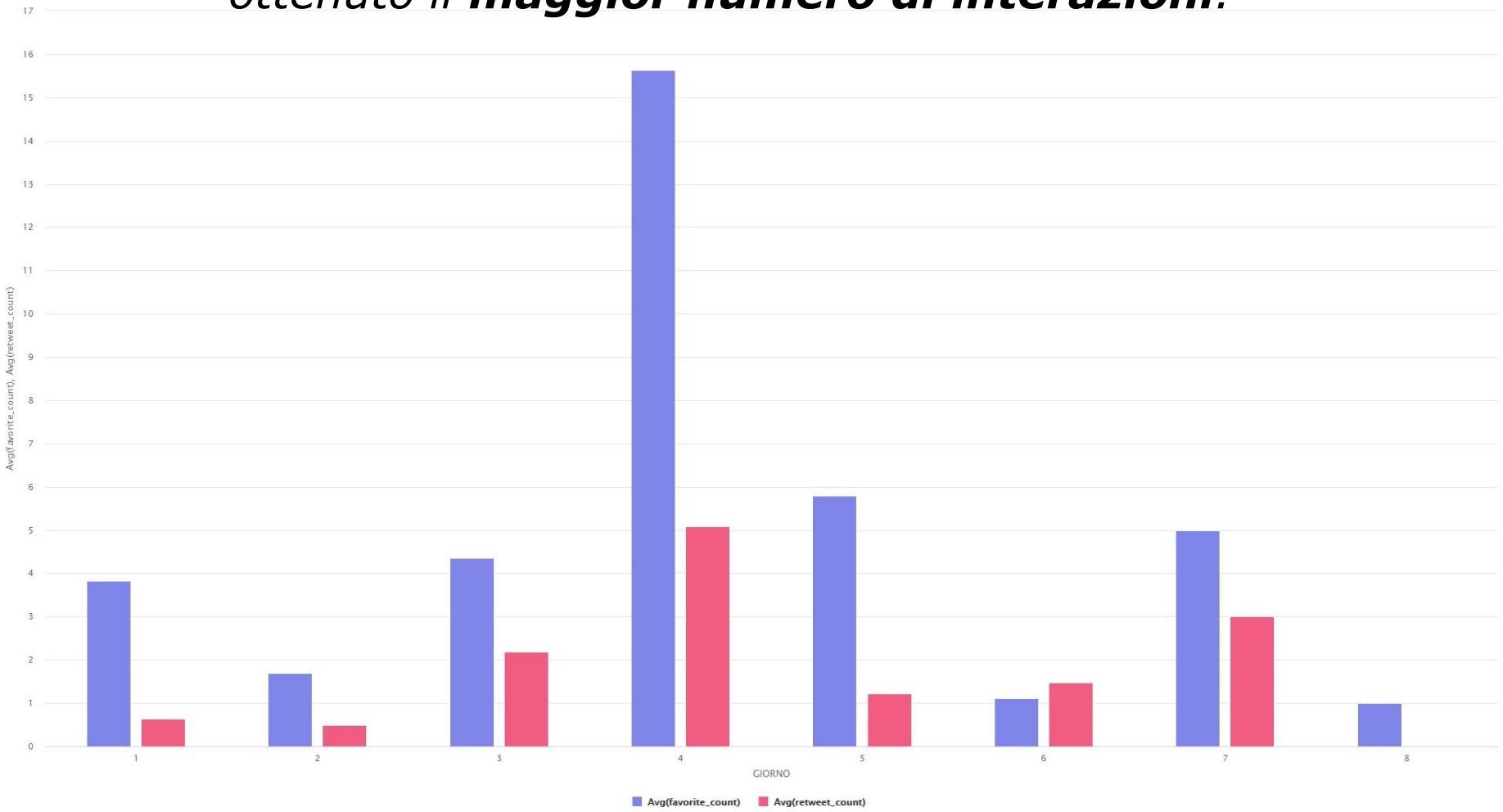
RETWEET_COUNT>=2

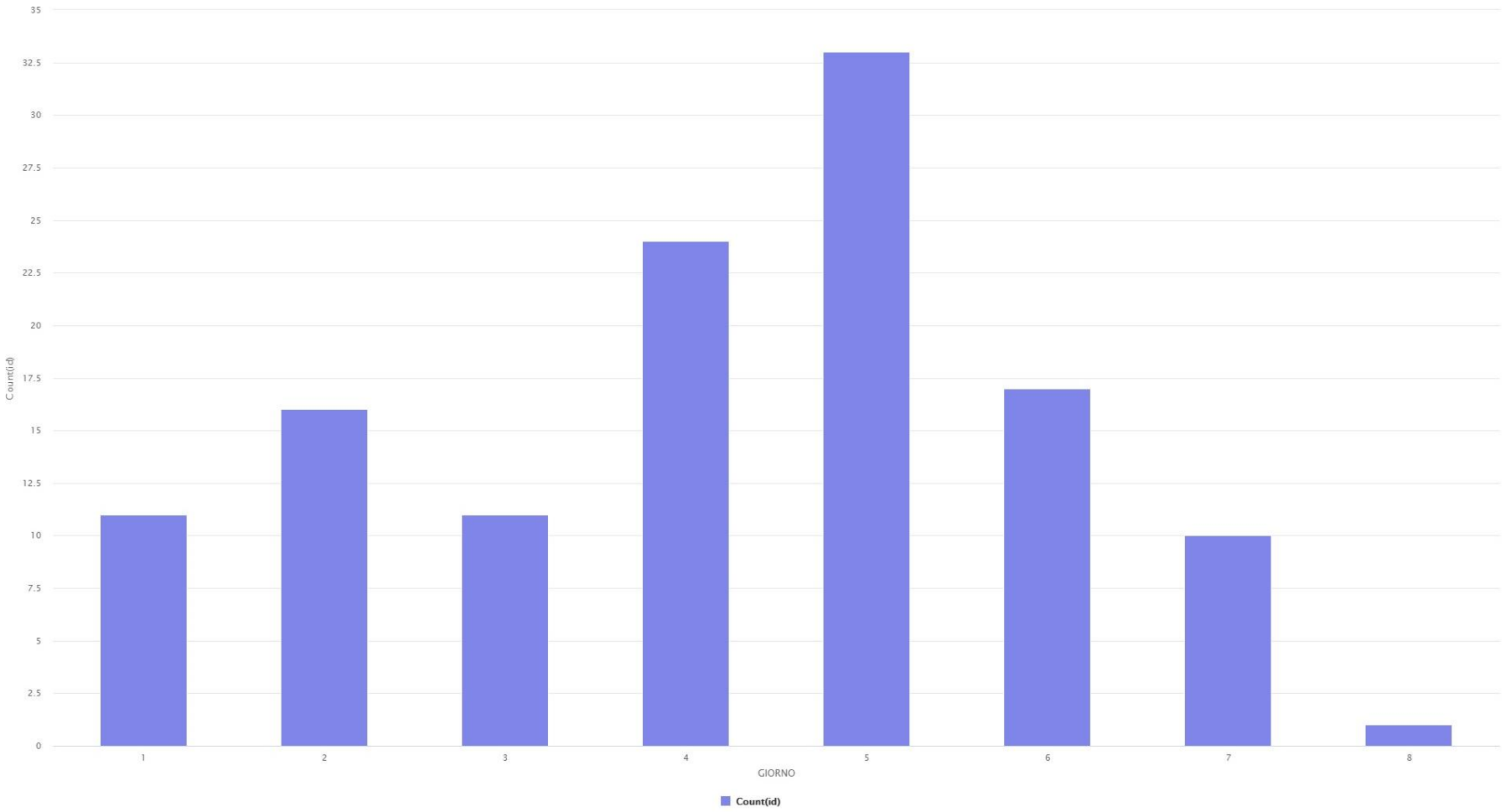


13 tweet

- Boris Johnson (supporto=0.231)
- Mean (supporto=0.231)
- Test (supporto=0.231)
- Clear (supporto=0.154)

*Si considerano ora i tweet della **giornata** che ha ottenuto il **maggior numero di interazioni**.*





*I tweet selezionati (**giorno 4**) sono **24**.*

Cluster 5, 4 gennaio 2021: FREQUENT ITEMSET

ITEMSET	Supporto	Differenza	Differenza %
uk	0.833	0.126	15.13%
vaccin	0.292	0.121	41.44%
variant	0.250	-0.157	-62.8%
strain	0.167	0.053	31.74%
boris (Johnson)	0.125	0.068	54.40%
countri	0.125	0.044	35.20%
govern	0.125	0.060	48%
lockdown	0.125	-0.013	-10.40%
christmas	0.083	0.059	71.08%
alert	0.083	0.059	71.08%

OBIETTIVI DI BUSINESS