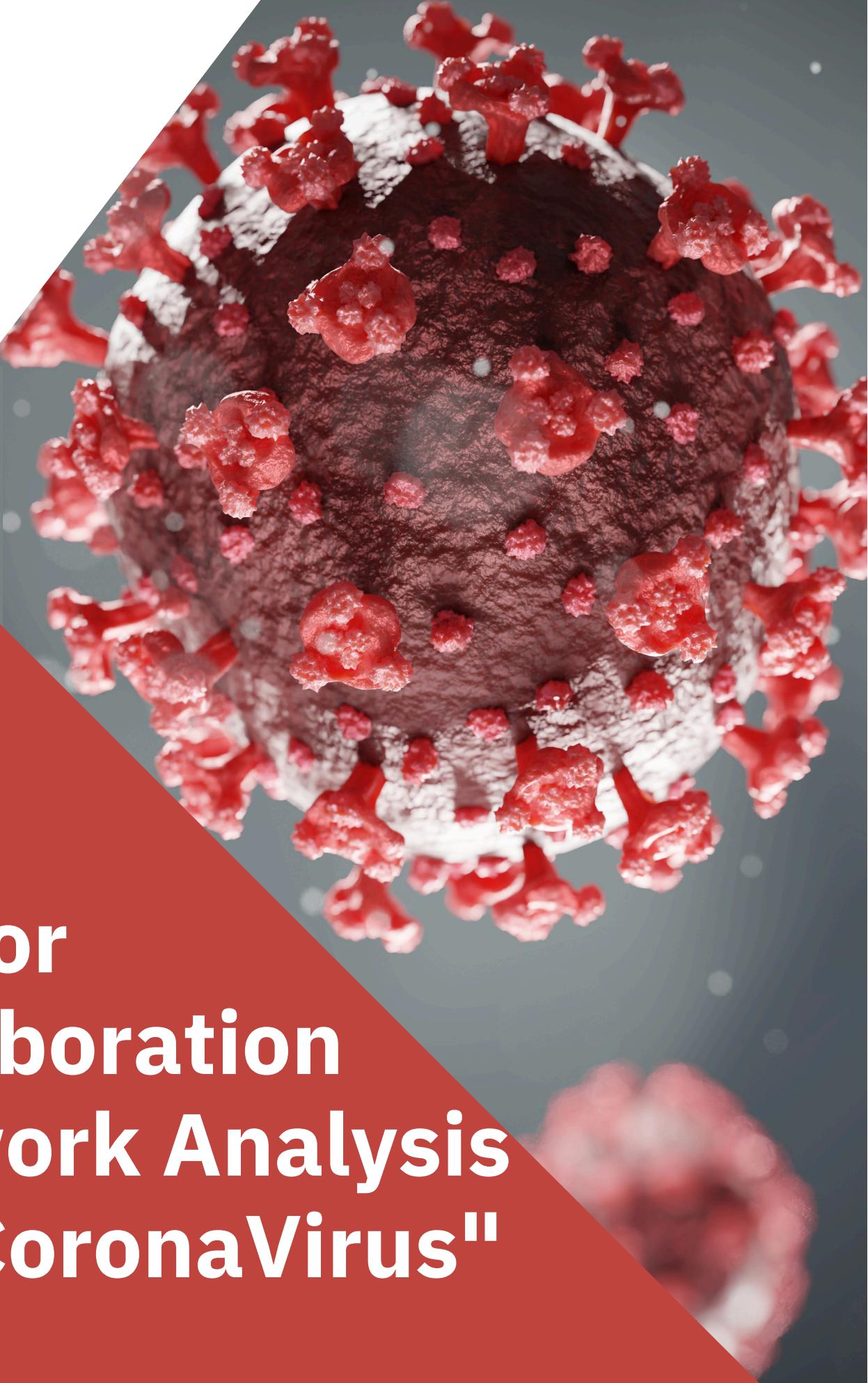


Author Collaboration Network Analysis on "CoronaVirus"





Università degli Studi di Cagliari

Facoltà di Scienze Economiche, Giuridiche e Politiche

Corso di Laurea in Data Science, Business Analytics e Innovazione

PROGETTO SULL' ANALISI DELLE RETI DI COLLABORAZIONE TRA AUTORI NELLA RICERCA SCIENTIFICA IN AMBITO “CORONAVIRUS”

A cura di:

**Francesco Mussetti
Enrico Caria
Ilaria Muru
Alberto Paschina**

Professore:

Stefano Matta

Anno Accademico 22/23

INDICE

1.0) INTRODUZIONE

1.1) QUESITI POSTI

2.0) DATASET

2.1) DESCRIZIONE

2.2) PRE-PROCESSING

3) ANALISI ESPLORATIVA

4.0) ANALISI PER QUESITO

4.1) QUESITO A

4.2) QUESITO B

4.3) QUESITO C

4.4) QUESITO D

4.5) QUESITO E

4.6) QUESITO F

5) CONCLUSIONI

6) BIBLIOGRAFIA

La scienza dei network, anche nota come "analisi delle reti" o "teoria dei network", è una branca della scienza che si occupa di studiare le proprietà e le caratteristiche dei network, ovvero di insiemi di oggetti (ad esempio persone, computer, proteine) interconnessi tra loro.

In particolare, la scienza dei network si concentra sull'analisi dei modelli di connessione tra gli oggetti, cercando di comprendere come si diffondono le informazioni, come emergono comportamenti collettivi e come si propagano eventuali disfunzioni o malattie.

La scienza dei network utilizza una vasta gamma di tecniche matematiche e statistiche per analizzare le proprietà dei network, come la loro struttura, la loro robustezza, la loro dinamica nel tempo e le proprietà emergenti che possono emergere dai comportamenti locali degli oggetti interconnessi.

La scienza dei network ha molte applicazioni pratiche, come la gestione delle reti di telecomunicazioni, la comprensione delle reti sociali, l'identificazione di comunità all'interno di grandi dataset e la valutazione della sicurezza delle infrastrutture critiche.

In questo progetto, attraverso l'utilizzo delle metodologie proprie della scienza dei network si analizzeranno le Reti di collaborazione, come esempio le reti di co-autori in un determinato campo di ricerca, che possono essere rappresentate come grafi, dove i nodi rappresentano i ricercatori e le relazioni rappresentano le collaborazioni tra di loro.

Nel dettaglio, l'obiettivo di questo progetto si concentra sull'analisi delle Reti di Collaborazione tra Autori nella Ricerca Scientifica in ambito "Coronavirus" al fine di riuscire a descrivere tale fenomeno e come esso si sia modificato nel tempo anche a seguito della pandemia Sars-Covid19.

Tramite tale analisi dunque, cerchiamo di rispondere ad alcuni quesiti naturali che ci siamo posti in fase di progettazione riguardanti le possibili dinamiche che avvengono tra Autori o tra Autori e Riviste Scientifiche, che rappresentano il cardine di questa analisi.

L'elaborato si compone di diverse parti:

1. Raccolta dei dati, pre-processing e Analisi Esplorativa del Dataset.
2. Analisi specifica con l'utilizzo dei grafi per ogni singolo quesito posto.
3. Interpretazione dei risultati e Conclusioni.

1.1) QUESITI POSTI

A

COM'È STRUTTURATA LA COMUNITÀ SCIENTIFICA IN AMBITO CORONAVIRUS IN TERMINI DI COLLABORAZIONI?

COME L'AVVENTO DELLA PANDEMIA COVID19 HA INFUITO SU TALI STRUTTURE?

B

GLI AUTORI CON PIÙ COLLABORAZIONI COME SI POSIZIONANO ALL'INTERNO DEL GRAFO?

MANTENGONO LE STESSE POSIZIONI ANCHE IN SEGUITO ALLA PANDEMIA?

SVOLGONO UN RUOLO DI "PONTE" TRA DUE O PIÙ GRUPPI CIRCOSCRITTI DI AUTORI?

C

E' POSSIBILE INDIVIDUARE DELLE SOTTO-COMUNITÀ ALL' INTERNO DELLA COMUNITÀ SCIENTIFICA DOVE GLI AUTORI TENDONO A CREARE RETI DI SCAMBI RICORRENTI PARTICOLARMENTE FITTE TRA LORO?

TALI AUTORI HANNO UNA BETWEENNESS MAGGIORE E UN LIVELLO DI CITAZIONI MAGGIORE?

D

QUALI SONO LE RELAZIONI TRA AUTORI E RIVISTE?

ESISTE CORRISPONDENZA TRA LE MISURE DI CENTRALITÀ E LE RIVISTE PIÙ "FAMOSE"?

E

COME SI POSIZIONANO I "FIRST AUTHOR" ALL'INTERNO DEL GRAFO?

HANNO UNA BETWEENNESS MAGGIORE?

SVOLGONO QUINDI UN RUOLO DI "PONTE" TRA DUE O PIÙ GRUPPI CIRCOSCRITTI DI AUTORI?

F

QUALI SONO I PRINCIPALI TOPIC TRATTATI DURANTE GLI ANNI ANALIZZATI?

QUAL'È STATO IL LORO ANDAMENTO NEL CORSO DEL TEMPO E COME L'AVVENTO DELLA PANDEMIA L'HA MODIFICATO ?

SI PUÒ INDIVIDUARE L'ANDAMENTO DEI TOPIC TRAMITE IL GRAFO?

E' DUNQUE POSSIBILE EVIDENZIARE IL PASSAGGIO DI UN TOPIC DA INTERESSANTE A MENO INTERESSANTE E VICEVERSA TRAMITE LO STRUMENTO DEL GRAFO?

2.1) DESCRIZIONE

Al fine di analizzare le Reti di Collaborazione tra Autori nella Ricerca Scientifica in ambito “Coronavirus” e rispondere ai quesiti posti, abbiamo utilizzato come fonte gli articoli scientifici pubblicati su PUBMED. PubMed è un database bibliografico gratuito che contiene milioni di citazioni e abstract di articoli pubblicati in letteratura medica e biomedica. Viene gestito dal National Center for Biotechnology Information (NCBI) degli Stati Uniti e fornisce un accesso facile a una vasta quantità di informazioni sulle scoperte e le ricerche scientifiche in campo medico e biomedico rappresentando un’importante risorsa per ricercatori, medici, studenti e altri professionisti che lavorano nel settore medico o sono interessati ai progressi delle scienze biomediche.

Inserendo la KEYWORD “Coronavirus” sul motore di ricerca del database di PUBMED si ottengono

227.152,00 risultati che rappresentano le pubblicazioni scientifiche sul tema Coronavirus dal 1949 al 2023. Attraverso l’utilizzo della funzione integrata di PUBMED ci è stato possibile scaricare i risultati in formato CSV, ottenendo così i dati in formato tabellare contenenti le informazioni rispetto a 11 variabili descritte di seguito:

- **PMID (PubMed Identifier):** un numero univoco assegnato a ciascuna pubblicazione presente nel database.
- **Title:** il titolo della pubblicazione.
- **Authors:** l’elenco degli autori che hanno contribuito alla pubblicazione.
- **Citation:** le informazioni sulle citazioni fornite riguardano il volume dell’articolo, il numero della rivista o il numero del fascicolo in cui è stato pubblicato l’articolo, il numero di pagina dell’articolo e il DOI.
- **First Author:** il primo autore della pubblicazione, che spesso è anche il corrispondente autore.
- **Journal/Book:** il nome della rivista o del libro in cui è stata pubblicata la pubblicazione.
- **Publication Year:** l’anno di pubblicazione della pubblicazione.
- **Create Date:** la data in cui la pubblicazione è stata inserita nel database PubMed.
- **PMCID (PubMed Central Identifier):** un numero univoco assegnato a pubblicazioni presenti in PubMed Central, un database gratuito di articoli a pieno testo.
- **NIHMS ID (National Institutes of Health Manuscript System Identifier):** un numero univoco assegnato a pubblicazioni che sono state sottomesse al sistema NIHMS per la revisione e la pubblicazione.
- **DOI (Digital Object Identifier):** un identificatore univoco assegnato a pubblicazioni digitali che permette di individuare univocamente un documento e di accedervi in modo permanente.

2.2) PRE-PROCESSING

Una volta ottenuti i dati grezzi, si è passati alla fase di pre-processing in modo da ripulire i dati e trasformarli in un formato più adatto alla nostra analisi. In tal senso si è proceduto a una prima rielaborazione generale dei dataset e a successive rielaborazioni più specifiche a seconda dell'analisi da effettuare per rispondere ai differenti quesiti.

Con riguardo al pre-processing generale, non essendo possibile scaricare più di 10.000 records alla volta da PubMed, si è reso necessario il Download di più Dataset distinti a seconda della numerosità degli articoli scientifici per anno, che sono stati successivamente concatenati grazie all'utilizzo della libreria Pandas di Python andando a creare un nuovo Dataset completo, contenente 227.737 articoli scientifici in tema "coronavirus" da 1949 al 2023 senza duplicati.

Al fine di migliorare la qualità dei dati di partenza delle nostre analisi, si è deciso di verificare la presenza di valori anomali quali Na e NaN e di variabili non utili all'analisi, procedendo alla loro eliminazione.

Nel dettaglio vista la grande presenza di valori anomali per le Variabili **PMCID (PubMed Central Identifier)**, **NIHMS ID (National Institutes of Health Manuscript System Identifier)** e **DOI (Digital Object Identifier)** e al contempo vista il loro scarso contributo informativo, abbiamo optato per una loro eliminazione così come per **Create Date** e **Citation** assicurandoci infine di eliminare anche ulteriori righe che presentassero valori anomali.

Un ulteriore grado di qualità è stato fornito tramite una seconda analisi, dove si è andato a ripulire il Dataset specialmente in riferimento alla variabile "Authors" che, dovendo essere processata per individuare i vari coautori nelle analisi successive, presentava al suo interno caratteri anomali che, associati al nome di un autore, avevano come effetto quello di fornire duplicati degli stessi, fornendo un quadro più confusionario e con maggiori attori rispetto alla realtà.

Lo stesso discorso vale per quegli autori che erano presentati sia in carattere minuscolo e maiuscolo.

A tale scopo è stata quindi effettuata un'analisi dettagliata che individuasse i caratteri anomali e li eliminasse, fornendo così un'informazione più chiara ed intellegibile.

Al fine di rispondere al quesito E si è optato per l'implementazione di una Topic Modeling, e si è reso dunque necessario un ulteriore fase di pre-processing al fine di ricercare gli Abstract di ciascun articolo analizzato sui quali applicare il modello.

L'abstract, contenuto all'interno dei file di testo scaricabili da PubMed secondo le medesime condizioni valide per il pre-processing generale affrontato in precedenza, è stato estratto grazie al codice contenuto nello stesso file Python "Preprocessing_FINAL" ed associato a ciascun articolo tramite il suo PMID, ottenendo così il "**Dataset_Corona_1949_2023_FINAL.xlsx**" alla base di tutte le analisi effettuate successivamente, avente 224.383 osservazioni e avente le seguenti 7 variabili:

- **PMID (PubMed Identifier)**: un numero univoco assegnato a ciascuna pubblicazione presente nel database.
- **Title**: il titolo della pubblicazione.
- **Authors**: l'elenco degli autori che hanno contribuito alla pubblicazione.
- **First Author**: il primo autore della pubblicazione, che spesso è anche il corrispondente autore.
- **Journal/Book**: il nome della rivista o del libro in cui è stata pubblicata la pubblicazione.
- **Publication Year**: l'anno di pubblicazione della pubblicazione.
- **Abstract**: l'abstract collegato a ciascuna pubblicazione.

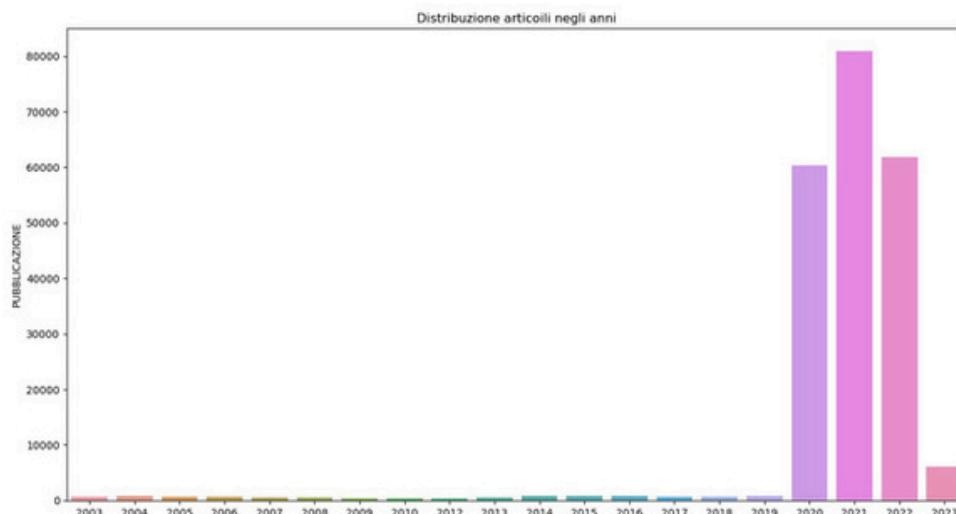
3.0 ANALISI ESPLORATIVA

Per quanto riguarda l'analisi del dataset è stata realizzata in diversi step, un primo passaggio in cui esaminiamo il dataset nel complesso, quindi considerando tutte le osservazioni dal 1949 al 2023. Successivamente focalizzando l'attenzione sugli anni in cui il corona virus si è diffuso maggiormente (2020,2021,2022).

Nel corso di questa analisi analizzeremo principalmente due variabili del dataset ‘First Author’ e ‘Journal/Book’. La prima come si può intuire ci indica il nome del soggetto che ha realizzato la pubblicazione. Mentre la seconda variabile ci indica in quale rivista è stato pubblicato l'articolo.

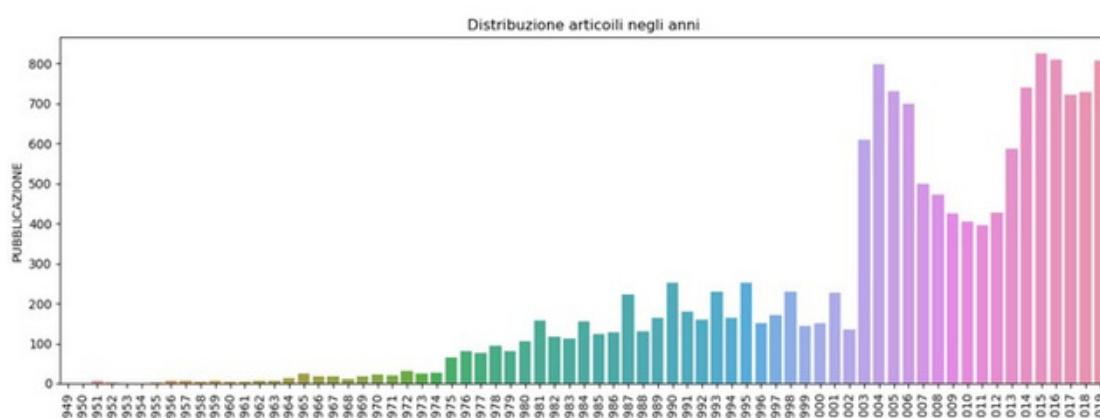
Dataset corona-virus completo (1949/2023)

Il dataset comprende 224383 osservazioni per 7 colonne (di seguito riportate e descritte nel paragrafo precedente: PMID,Title,Authors,Citation,First Author,Journal/Book,Publication Year).



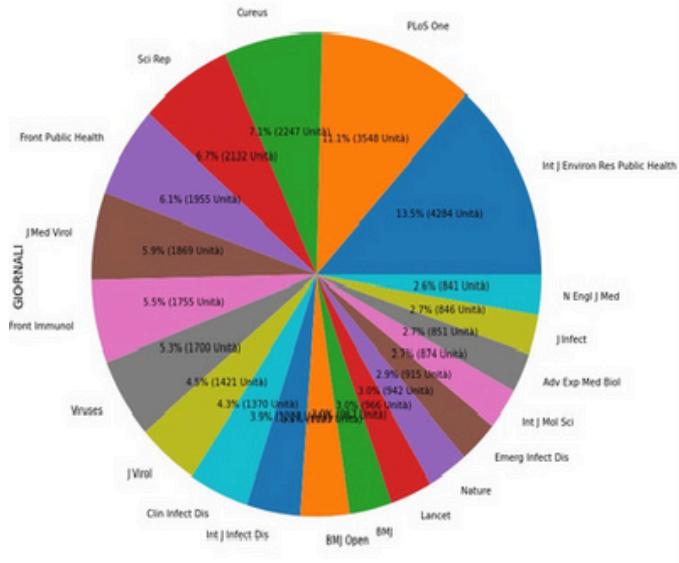
Come possiamo notare dal grafico (in cui sono stati filtrati solamente gli anni con un numero di pubblicazioni superiori a 400), le osservazioni relative al numero di articoli pubblicati in un determinato anno si distribuiscono per lo più negli anni in cui è scoppiata la pandemia dovuta al coronavirus, quindi 2020,2021 e 2022, anni in cui le pubblicazioni superano le 60.000 unità e come già detto in precedenza, andremo ad analizzare singolarmente per capire quali riviste e quali autori hanno avuto maggiore importanza e se ciò sarà poi confermato dalla network analysis.

Abbiamo realizzato inoltre la distribuzione delle osservazioni di tutto l'arco temporale precedente alla pandemia (1949-2018) , in cui notiamo dei picchi in corrispondenza di scoppi di epidemia, il più importante lo abbiamo tra il 2002 e il 2003, ma ci sono vari periodi che hanno portato una frequenza maggiore di pubblicazioni, la maggior parte coincidenti a periodi di diffusione del coronavirus. Di seguito il grafico .

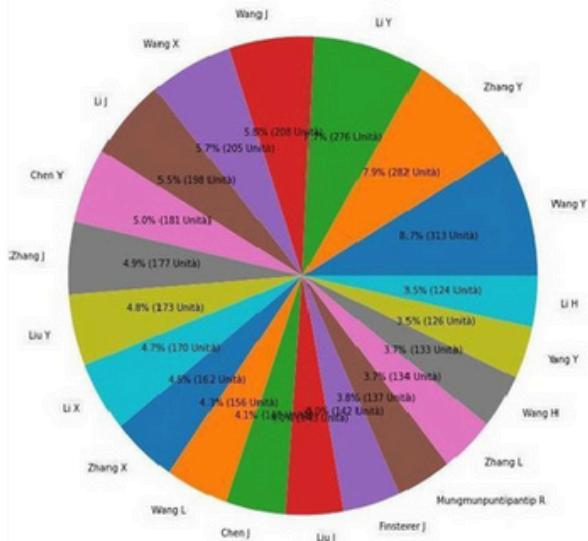


3.0 ANALISI ESPLORATIVA

Distribuzione giornali con pubblicazioni più frequenti



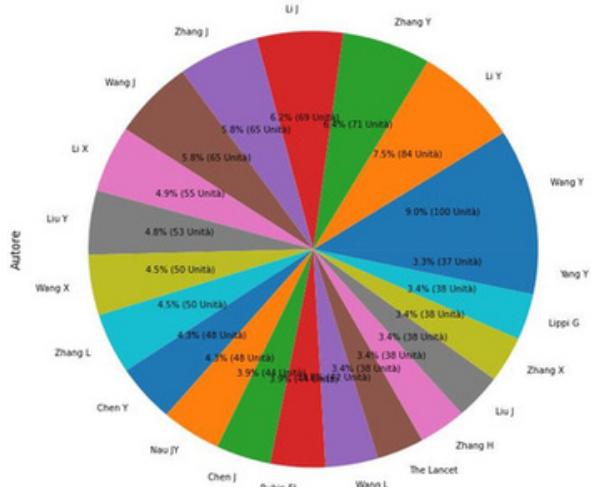
Autori con pubblicazioni più frequenti



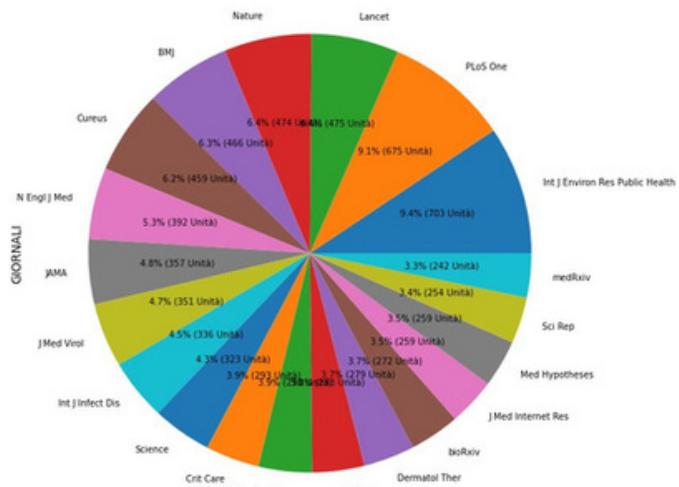
Per realizzare i grafici sono stati filtrati i venti autori con più pubblicazioni, ovvero gli autori con più di 120 pubblicazioni e le venti riviste con più osservazioni, quindi con più di 800 articoli pubblicati, in modo da farci un'idea su quali saranno i nodi di maggiore importanza nelle successive analisi dei network. Notiamo infatti nell'analisi comprendente tutto l'orizzonte temporale come gli autori per la maggior parte siano asiatici, mentre le riviste siano per lo più riviste accademiche americane. Andando più nel dettaglio possiamo andare a vedere in quali anni queste riviste hanno pubblicato la maggior parte delle loro pubblicazioni e se coincidono quindi al periodo della pandemia.

Numero pubblicazioni realizzate dagli autori e dalle riviste nel 2020:

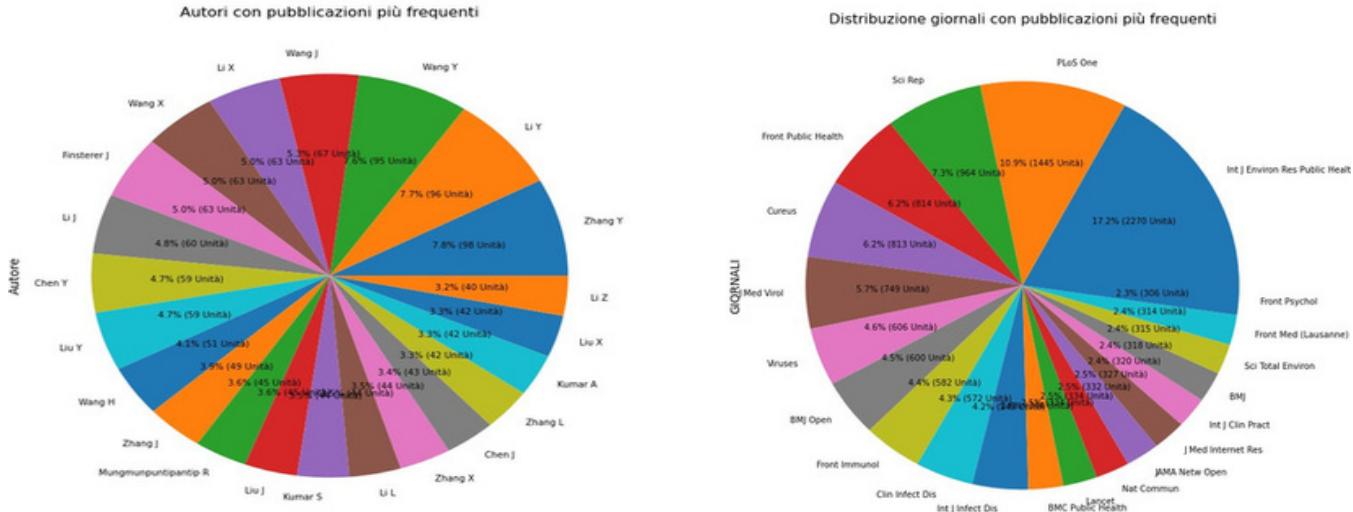
Autori con pubblicazioni più frequenti



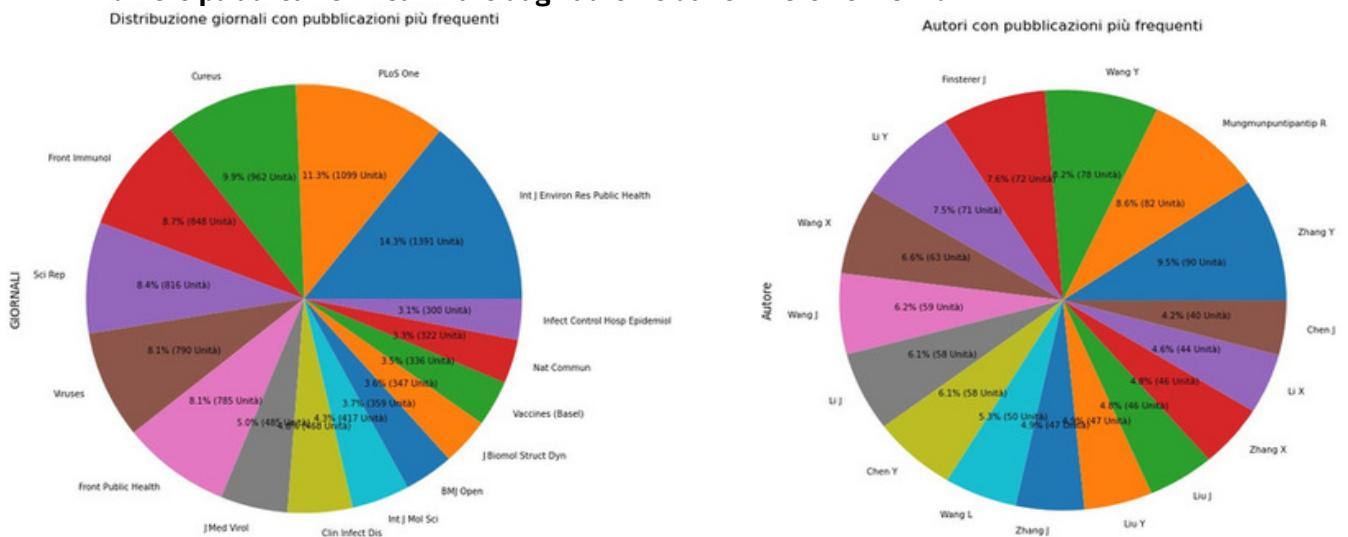
Distribuzione giornali con pubblicazioni più frequenti



Numero pubblicazioni realizzate dagli autori e dalle riviste nel 2021:



Numero pubblicazioni realizzate dagli autori e dalle riviste nel 2022:



Come già si poteva intuire dal grafico generale le riviste che hanno pubblicato maggiormente nel periodo del covid coincidono con le etichette messe in evidenza dai pieplot realizzati anno per anno, in particolare notiamo il ripetersi di alcuni nomi come :

Plos One: Rivista scientifica di tipo open access, pubblicata dalla libreria delle scienze dal 2006, con sede San Francisco

Int J environ Res public Health: Rivista scientifica di tipo open access, pubblicata da MDPI, nata nel 2004 ma che ebbe un grande impatto nel 2021 appunto.

Cureus: Quotidiano di medicina generale open access, sottoposto a revisione paritaria pre e post pubblicazione.

Fondata nel 2009.

Sci Rep: rivista scientifica open access con revisione paritaria pubblicata da Nature Portfolio che si occupa delle scienze naturali. Fondata nel 2011.

Front public Health: rivista multidisciplinare che si occupa appunto di salute (professionale, mentale obesità ecc) open access, pubblica online con sede in svizzera e fondata nel 2013.

Viruses: Rivista open access pubblicata online dal MDPI .Di nazionalità americana.

Come possiamo notare la maggior parte delle riviste con maggiori pubblicazioni sono per lo più open access e di origine Americana, Nessuna di queste riviste è nata prima degli anni 2000 ma possiamo notare come tutte abbiano avuto il loro maggiore exploit durante gli anni del covid 2020-2022

Per quanto riguarda gli autori notiamo come i nomi più ricorrenti siano : Zhang, Chen e Wang , nomi particolarmente comuni in Cina, che si ripetono in maniera importante anno per anno nei grafici, dandoci un'idea di chi siano gli autori che probabilmente avranno maggiori collaborazioni soprattutto nel periodo della pandemia.

4.0 ANALISI PER QUESITO

4.1) QUESITO A

Al fine di poter rispondere al quesito A, ovvero riuscire ad individuare le “alleanze” nonchè collaborazioni tra i diversi autori si è provveduto all’analisi dei grafi in modo da poter schematizzare e rendere evidenti quali siano i gruppi di autori che hanno collaborato nel periodo oggetto di studio.

Per poter realizzare questo studio si è provveduto in prima fase alla suddivisione del dataset in due parti la prima che rispecchia il periodo pre-pandemia dal 1949 al 2018 e un secondo periodo che rappresenta il periodo di pandemia 2019-2023.

NUMERO OSSERVAZIONI DATAFRAME 1949/2018:

14.414

NUMERO OSSERVAZIONI DATAFRAME 2019/2023:

209.969

Successivamente, dopo aver estratto un campione del 10% di tali osservazioni si è provveduto alla scomposizione della colonna ‘Authors’ in modo tale da poter creare un dizionario in cui poter associare ad ogni autore un ID unico, processo importante per la creazione dei dataframe da poter analizzare.

L’analisi è composta dalla realizzazione di due grandi grafi, relativi ai due periodi di suddivisione. I risultati sono rappresentati nella tabella sottostante :

Grafo	1949-2018	2019-2023
L'ordine del grafo: Numero di nodi collegati direttamente ad un altro nodo/vertice	5827	108951
La dimensione del grafo: Numero lati/archi	29918	973722
Densità grafo: Rapporto tra il numero di lati e il numero massimo di lati che il grafo potrebbe avere, misura che varia da 0 a 1: un grafo con densità > 0.5 è considerato denso, un grafo con densità < 0.5 è considerato sparso	0.002	0.0
Grado medio: Rapporto tra la somma del grado di tutti i nodi e il numero dei nodi del grafo	10.269	17.874
N°componenti: Coppia di vertici connessa da cammini	432	6203
Sottografo indotto: Componente più elevata	3603	80123

4.2) QUESITO B

Per rispondere a questa domanda è stato necessario identificare i 24 Autori/nodi più influenti del periodo, quindi gli autori con un maggior numero di collaborazioni
Rappresentati Nella tabella di seguito :

1949/2018	Autore/grado	1949/2018	Autore/grado	2019/2023	Autore/grado	2019/2023	Autore/grado
1	'li y', 159	13	'zhang h', 93	1	'zhang y', 1465	13	'zhang j', 913
2	'baric rs', 137	14	'li w', 85	2	'wang y', 1419	14	'li h', 864
3	'drosten c', 126	15	('zhang y', 84)	3	'wang j', 1313	15	'liu z', 861
4	'wang j', 118	16	'zhang x', 82	4	'liu y', 1194	16	'zhang x', 859
5	'wang y', 111	17	'xu y', 82	5	'li y', 1167	17	'liu x', 854
6	'wang h', 110	18	'li h', 82	6	'wang l', 1097	18	'chen j', 816
7	'perlmans', 106	19	'guan y', 82	7	'wang x', 1064	19	'wang h', 805
8	'yuen ky', 103	20	'yang y', 82	8	'li j', 1031	20	'wang z', 790
9	'liu y', 102	21	'chan kh', 81	9	('liu j', 1010)	21	'li l', 760
10	'zhang l', 102	22	'weiss sr', 77	10	'chen y', 987	22	'chen z', 743
11	'wang m', 100	23	'zhou y', 77	11	'zhang l', 972	23	('li z', 715)
12	'anjuanez l', 95	24	'liu s', 76	12	'li x', 933	24	('zhang h', 687)

Tra le due tabelle è possibile evidenziare come ci siano delle corrispondenze tra i nodi più influenti dei due grafi, in particolare:

L'autore '**li y'** si trovava nella posizione **1** nel 1° grafo, mentre nel 2° grafo si trova nella posizione **5**

L'autore '**wang j'** si trovava nella posizione **4** nel 1° grafo, mentre nel 2° grafo si trova nella posizione **3**

L'autore '**wang y'** si trovava nella posizione **5** nel 1° grafo, mentre nel 2° grafo si trova nella posizione **2**

L'autore '**wang h'** si trovava nella posizione **6** nel 1° grafo, mentre nel 2° grafo si trova nella posizione **19**

L'autore '**liu y'** si trovava nella posizione **9** nel 1° grafo, mentre nel 2° grafo si trova nella posizione **4**

L'autore '**zhang l'** si trovava nella posizione **10** nel 1° grafo, mentre nel 2° grafo si trova nella posizione **11**

L'autore '**zhang h'** si trovava nella posizione **13** nel 1° grafo, mentre nel 2° grafo si trova nella posizione **24**

L'autore '**zhang y'** si trovava nella posizione **15** nel 1° grafo, mentre nel 2° grafo si trova nella posizione **1**

L'autore '**zhang x'** si trovava nella posizione **16** nel 1° grafo, mentre nel 2° grafo si trova nella posizione **16**

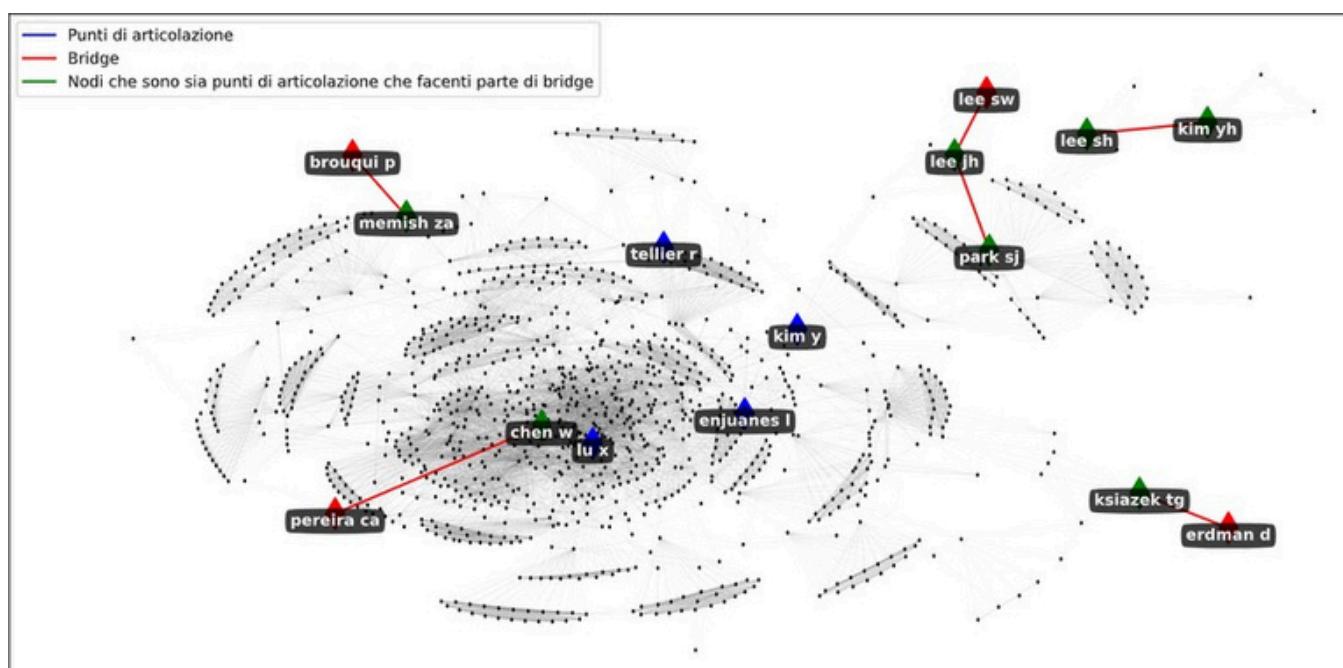
L'autore '**li h'** si trovava nella posizione **18** nel 1° grafo, mentre nel 2° grafo si trova nella posizione **14**

Si può notare come molti degli autori più importanti in termini di pubblicazione prima della scoperta del coronavirus SARS CoV-2 continuino a esserlo anche nel periodo più recente, come a voler significare che i massimi esperti abbiano potuto approfondire la materia partendo da una base di conoscenza già solida.

Ci siamo dunque chiesti se tali autori fungessero da ponte tra gruppi circoscritti di autori (caratterizzati da collaborazioni più fitte).

Per farlo, innanzitutto abbiamo indagato sul campione del 10% sulla struttura generale, filtrando ulteriormente in quanto abbiamo selezionato, ai fini di una maggiore fruibilità di analisi visiva e non solo del grafo, i nodi aventi un grado di un certo livello: per la minore numerosità del campione è bastato scegliere gli autori con un livello di ‘degree’ superiore a 15 nel periodo 1949-2018, mentre per il periodo successivo lo abbiamo impostato pari a 100. In entrambi gli intervalli temporali si può notare come, concentrandosi sugli autori con la maggiore importanza in termini di produzione scientifica, si ottengono dei grafi connessi.

Di seguito il grafico relativo al primo grafo 1949-2018:



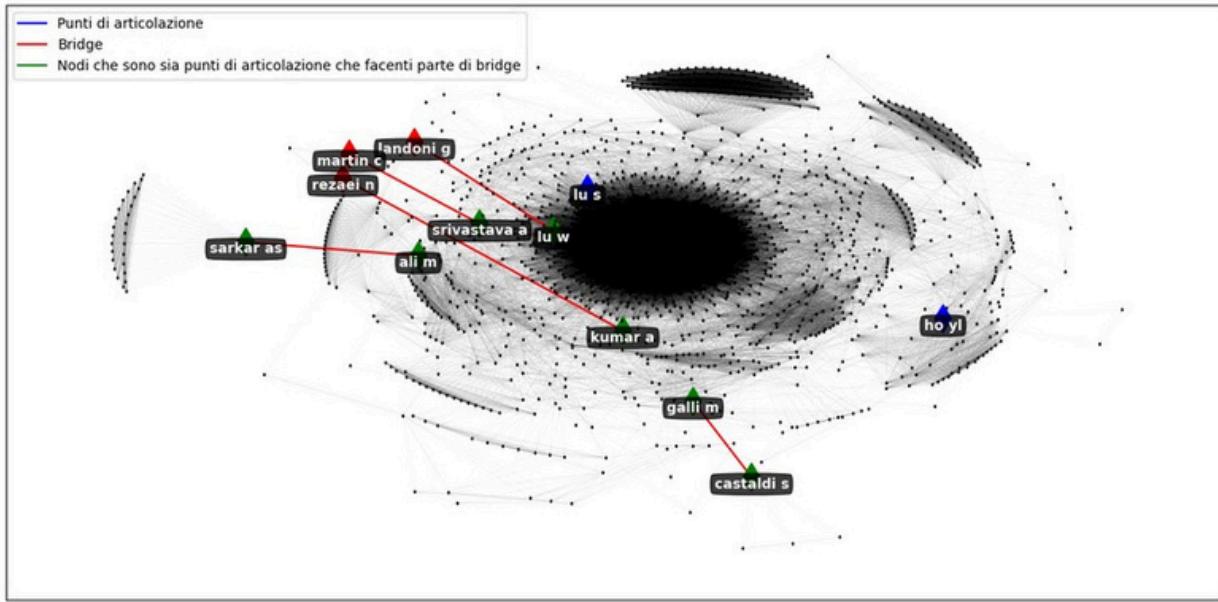
I punti di articolazione del grafo 1949-2018 sono: {'lu x', 'chen w', 'memish za', 'enjuanes l', 'lee jh', 'kim yh', 'tellier r', 'park sj', 'kim y', 'lee sh', 'ksiazek tg'}

I nodi che fanno parte dei bridge del grafo 1949-2018 sono: {'chen w', 'memish za', 'lee jh', 'kim yh', 'pereira ca', 'park sj', 'lee sw', 'lee sh', 'erdman d', 'brouqui p', 'ksiazek tg'}

Nodi che sono sia punti di articolazione che facenti parte di bridge{'chen w', 'memish za', 'lee jh', 'kim yh', 'park sj', 'lee sh', 'ksiazek tg'}

4.2 QUESITO B

Di seguito il grafico relativo al primo grafo 2019-2023:



I punti di articolazione del grafo 2019-2023 sono: {'galli m', 'srivastava a', 'lu s', 'ali m', 'lu w', 'kumar a', 'ho yl', 'castaldi s', 'sarkar as'}

I nodi che fanno parte dei bridge del grafo 2019-2023 sono: {'galli m', 'srivastava a', 'rezaei n', 'ali m', 'martin c', 'lu w', 'kumar a', 'landoni g', 'castaldi s', 'sarkar as'}

Nodi che sono sia punti di articolazione che facenti parte di bridge{'galli m', 'lu w', 'srivastava a', 'ali m', 'kumar a', 'castaldi s', 'sarkar as'}

Nelle immagini sono stati evidenziati appunto **bridge**, quindi archi la cui eliminazione fa sì che il grafo sia suddiviso in più componenti, nodi appartenenti a tali lati e **punti di articolazione**, ossia l'equivalente dei bridge in termini di lati/archi. Tuttavia tali nodi **non** coincidono con i più importanti. Si può interpretare questo risultato in questo senso:**gli autori che più pubblicano tendono anche a collaborare tra loro.**

Indagato in questa direzione. Per prima cosa è possibile notare come nel grafo inerente all'epoca 2019-2023 esista una nuvola più densa di nodi: ci siamo dunque chiesti, al fine di rispondere al quesito, se tale "componente" comprendesse gli autori più importanti; la risposta è che tutti i **cinque nodi con la più alta produzione scientifica si trovano tutti in quella sottocomponente**; In particolare, si tratta del sottografo indotto del grafo filtrato per un grado maggiore di 100, ossia la sua componente massimale che si individuerebbe qualora si eliminassero i punti di articolazione e/o i bridge.

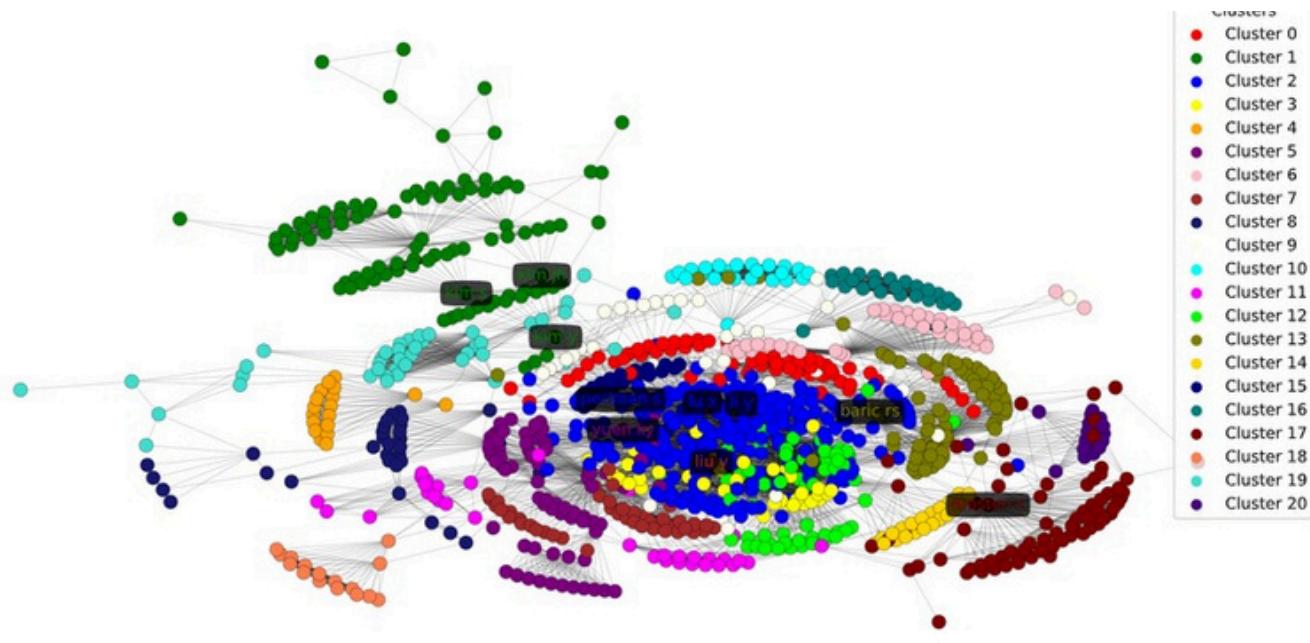
Il passo successivo è stato chiedersi come fosse strutturata tale componente massimale. Quesito posto successivamente

4.3) QUESITO C

Per poter rispondere a tale quesito abbiamo cercato di individuare, attraverso l'utilizzo della **cluster analysis**, dei gruppi di nodi rappresentativi di sottoreti di collaborazione per entrambi i periodi messi a confronto. Le etichette con il nome degli autori dello stesso colore dei diversi cluster a cui appartengono, rappresentanti la **betweenness**, che si può tradurre con “essere tra”, ovvero una misura di centralità che ci indica quali siano i nodi che fanno più da “intermediari” all'interno della rete, quindi ci indica quali autori fossero più influenti in quel determinato cluster.

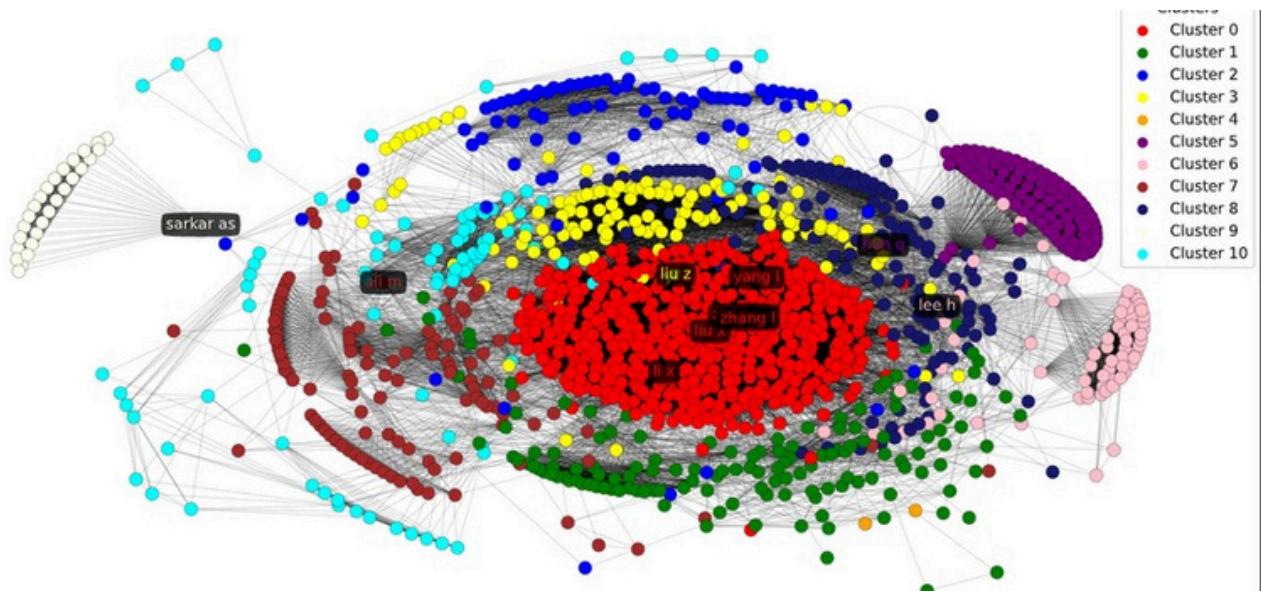
La **betweenness** di un vertice è il numero di cammini minimi che attraversano un nodo diviso per il numero totale di cammini minimi.

Di seguito il grafo con rappresentati i cluster dal 1949 al 2018:



Sono presenti all'interno del grafo 20 cluster che ci indicano perciò come fossero presenti 20 gruppi distinti di autori che collaboravano, ciò può essere dovuto anche al lungo periodo preso come riferimento.

Di seguito il grafo con rappresentati i cluster dal 2019 al 2023 :



Nel secondo periodo di riferimento possiamo notare la presenza di solamente 10 gruppi di alleanze tra gli autori, nonostante il numero molto maggiore di osservazioni. Ciò può essere dovuto sia al periodo minore di riferimento, ma anche dovuto al fatto che la comunità scientifica si sia maggiormente unita durante il periodo della pandemia .

4.4) QUESITO D

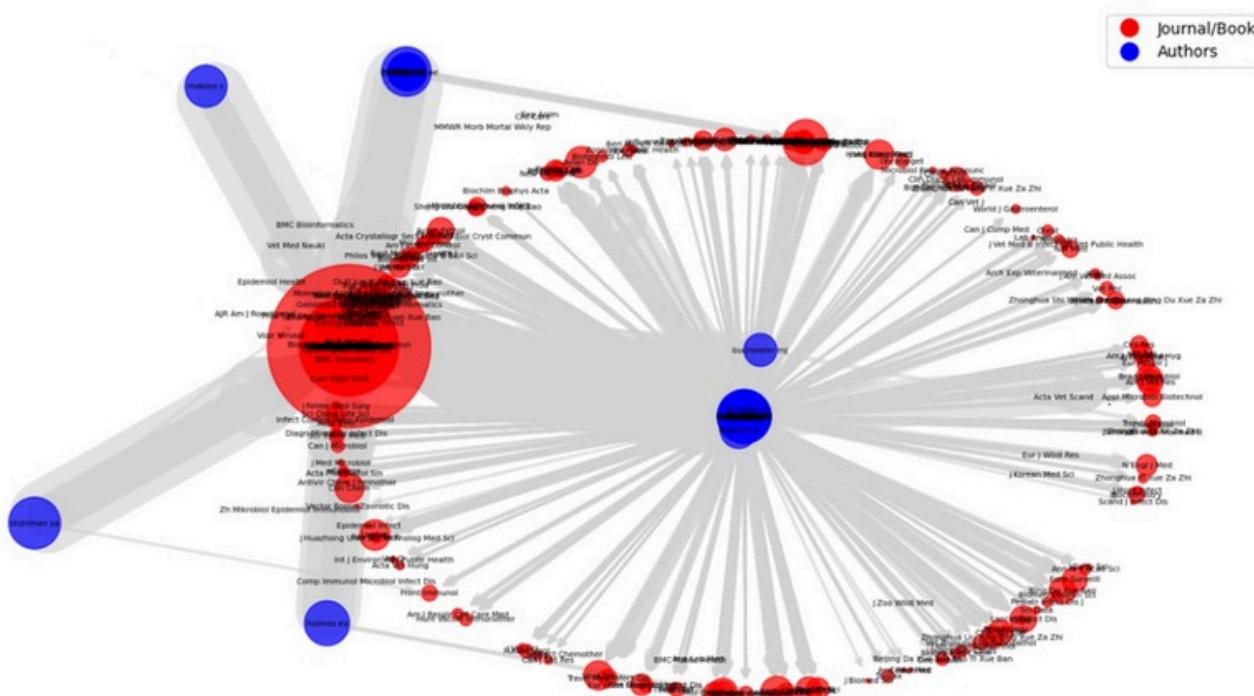
Al fine di rispondere al Quesito D, si è provveduto, in prima istanza, alla divisione del dataset in due macro-gruppi:

- df_authors_49_18, contenente i papers dal 1949 al 2018 (in numero 14)
- df_authors_19_23, contenente i papers dal 2019 al 2023.

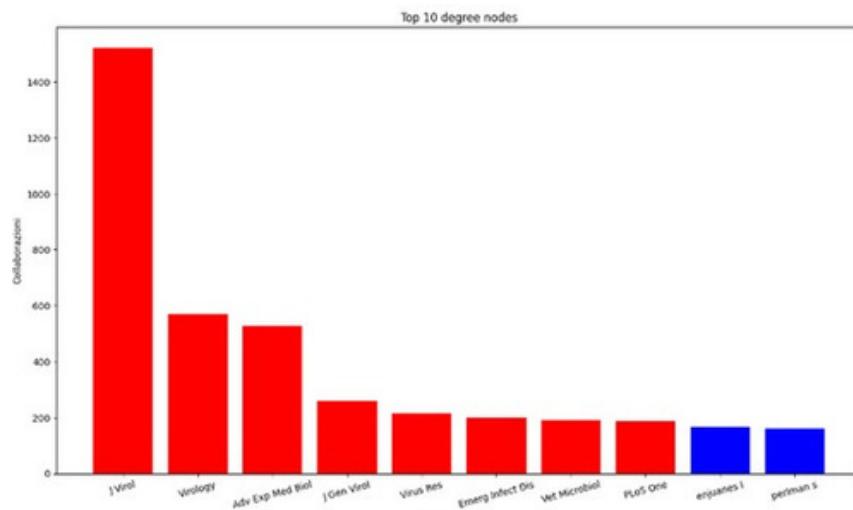
Questa divisione si è resa necessaria sia per una migliore comprensione del fenomeno (ovvero l'incidenza dell'epidemia di SarS-CoV-2 a partire dagli ultimi mesi del 2019, motivo per cui la massima produzione di papers si trova tra il 2020-2023 come si è evinto dall'analisi esplorativa, sia per una questione meramente computazionale. Vediamo ora in dettaglio i grafi ottenuti.

1949 - 2019

Contact between Authors and Journal/Book in 1949-2019

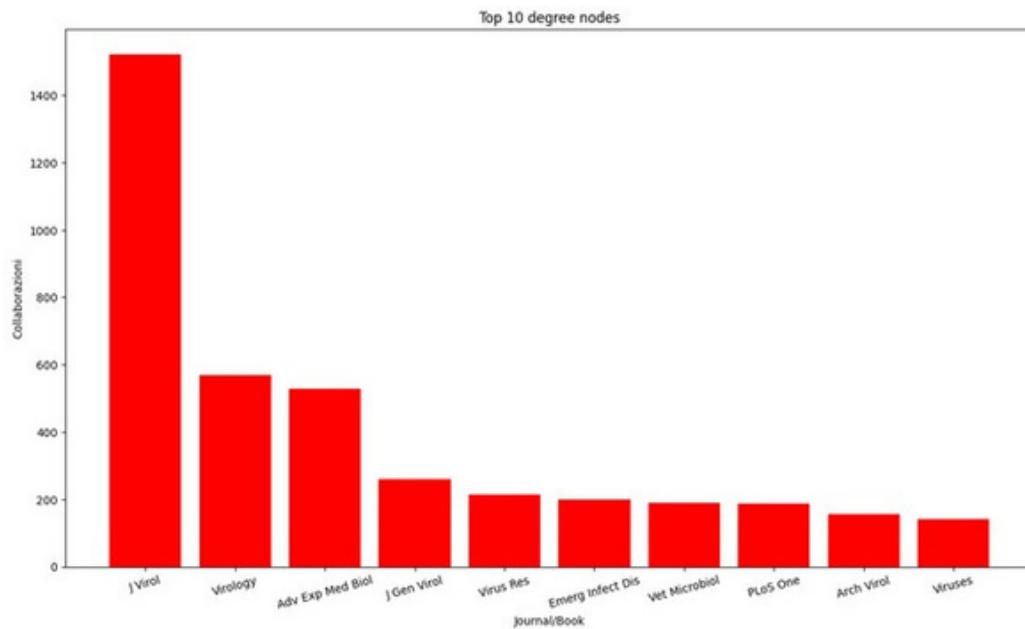


Dal grafo ottenuto per il periodo 1949-2019 (con degree=50) si evince che i 10 nodi dalla degree maggiore sono i Journals, che ricoprono le prime 8 posizioni (rispettivamente *Journal of Virology*, *Virology*, *Advances in Experimental Medicine and Biology*, *Journal of General Virology*, *Virus Research*, *Emerging Infectious Diseases*, *Veterinary Microbiology*, *PLoS One*) , seguono i due autori con il degree maggiore, Enjuanes I con 179 papers pubblicati e Perlman S con 180 papers pubblicati.

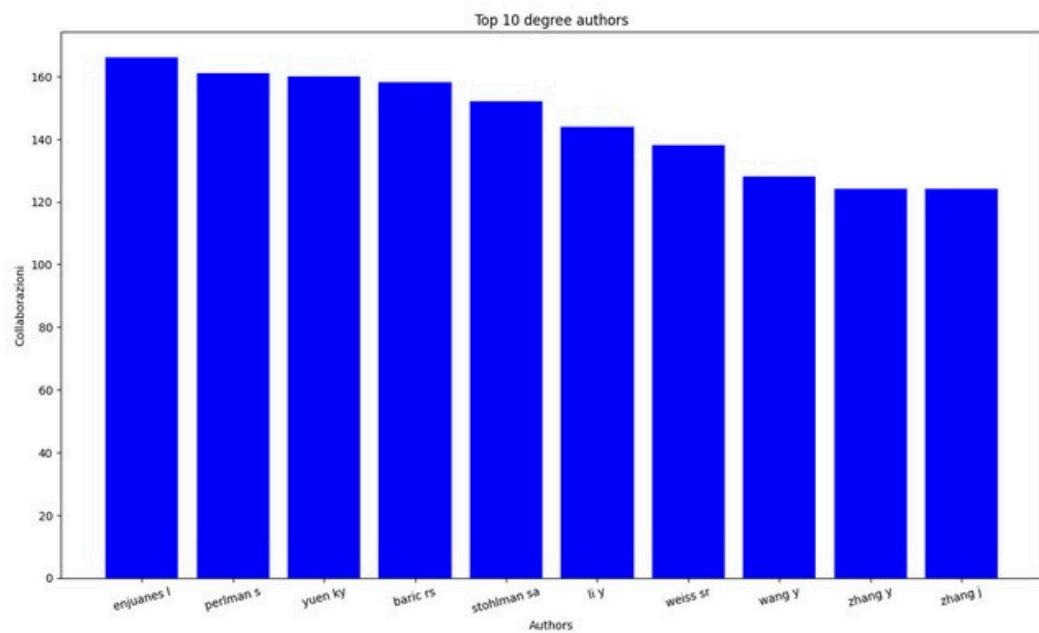


4.4 QUESITO D

Andando nel dettaglio per Journals, si evince che quello con il maggior numero di collaborazioni è il Journal of Virology, che si distacca nettamente.



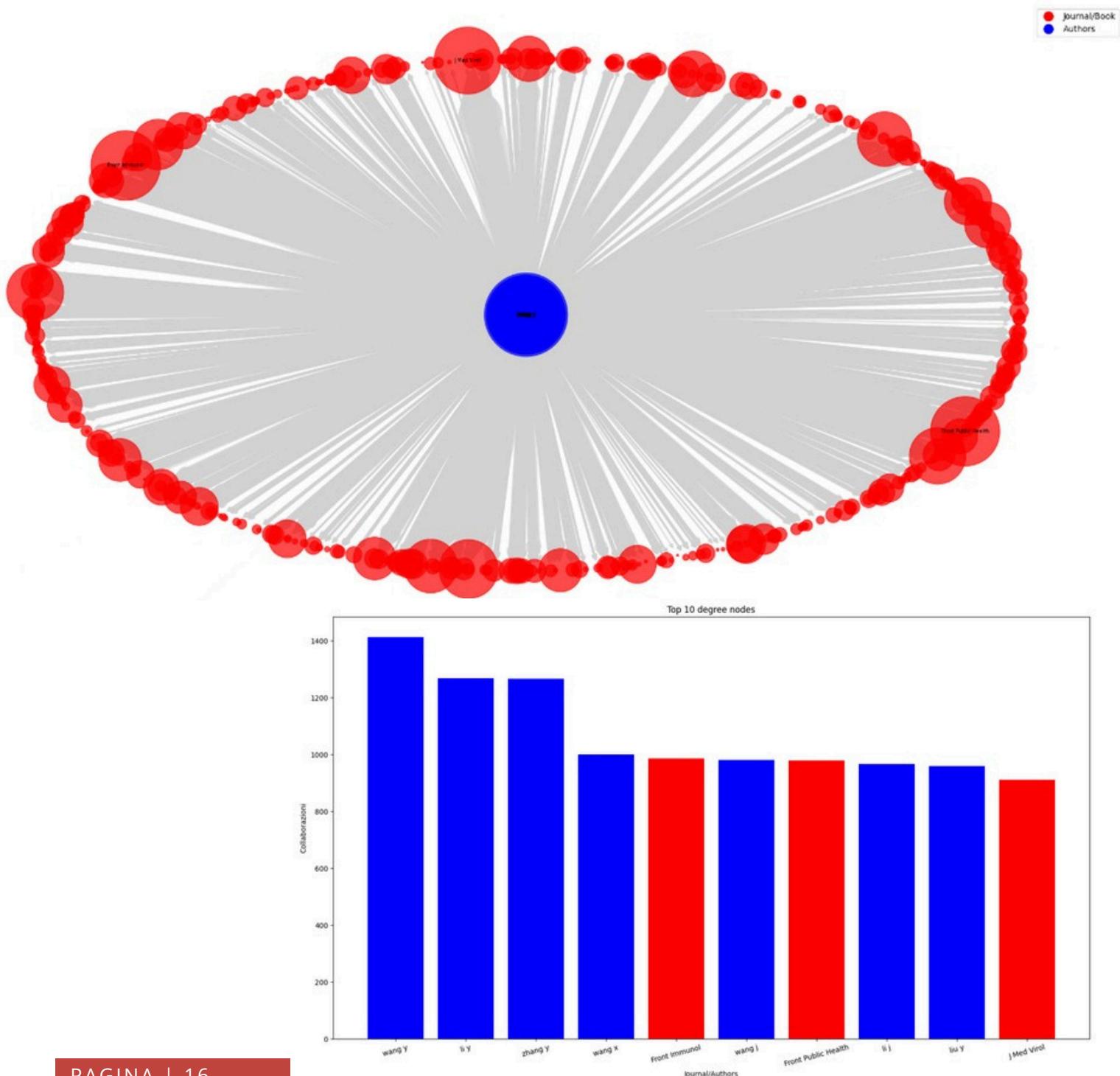
Nel caso invece degli autori, il distacco non è così netto come per i journals, mantenendo nelle prime dieci posizioni un numero di collaborazioni che non scende al di sotto delle 12.



2020- 2023

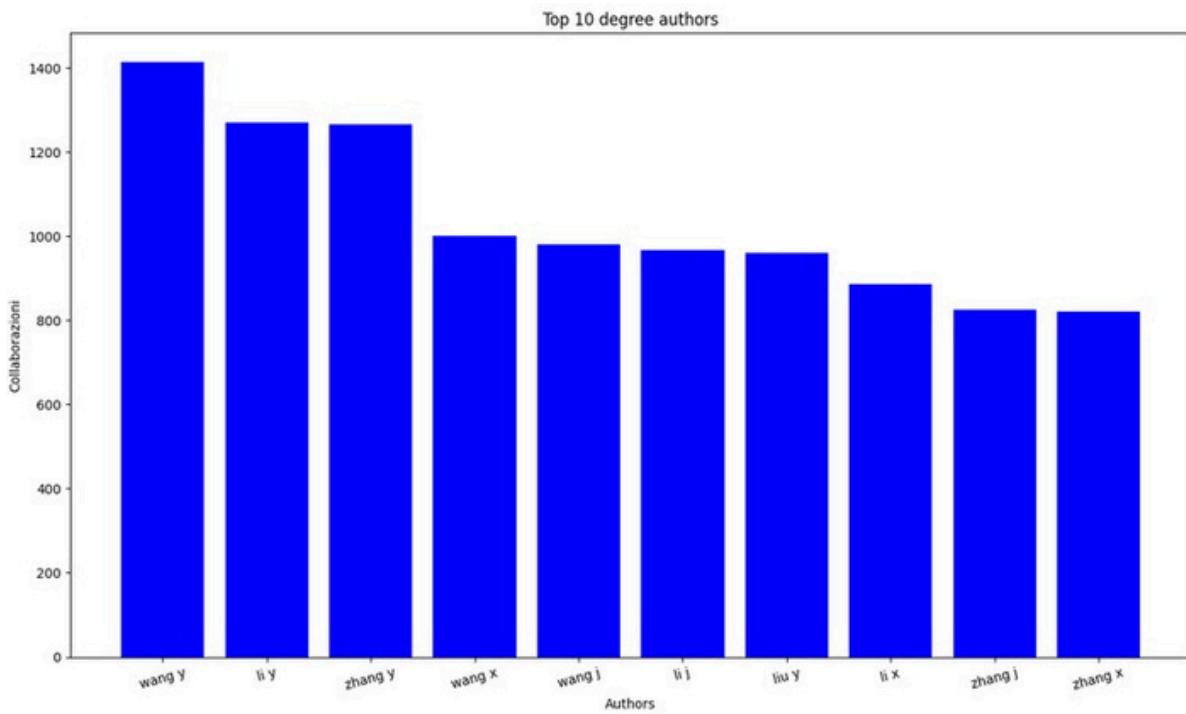
Situazione inversa invece per il triennio 2020-2023: infatti complice un aumento massiccio della produzione di papers, i nodi col più alto degree divengono gli autori. Questa inversione di tendenza è imputabile ad una produzione oltre la media di specifici autori, la cui ricerca si è concentrata per topic (inteso come topic generico Sars-Cov-2) data l'urgenza del tema, dell'analisi delle implicazioni e della ricerca di una soluzione, preminente rispetto a tutti gli altri eventuali temi di ricerca.

Contact between Authors and Journal/Book in 2020-2023

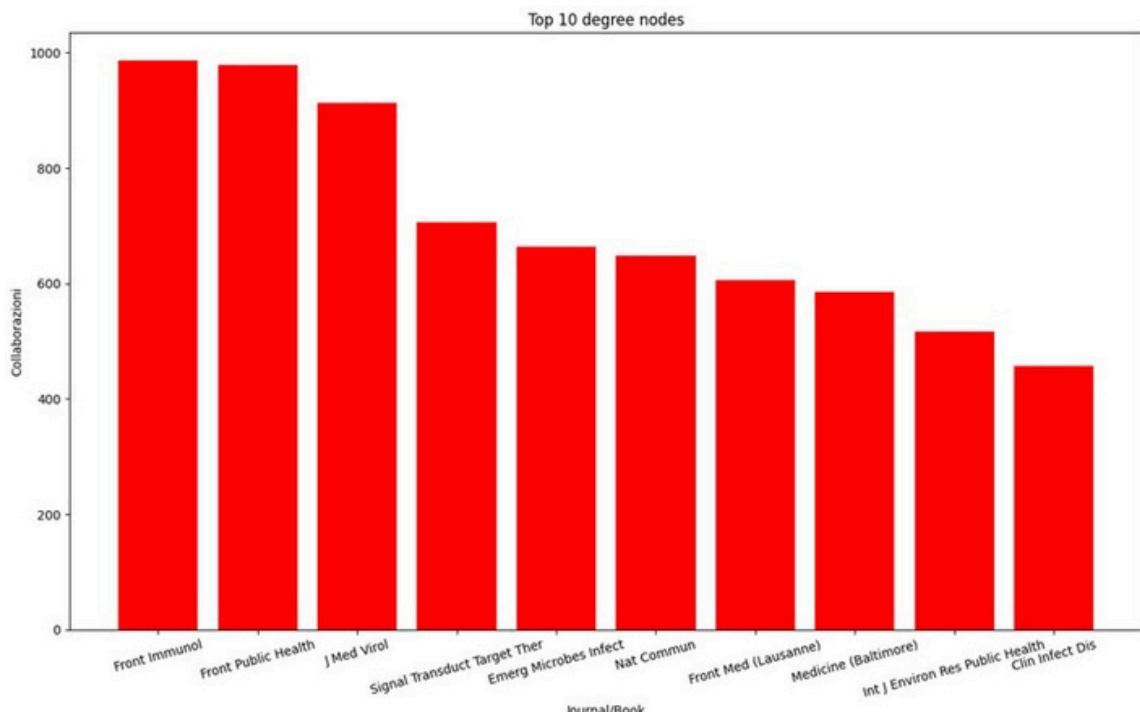


4.4 QUESITO D

Spicca inoltre, nel grafico dei 10 nodi top per autore la presunta origine dei nomi, riconducibile al Sud Est Asiatico[1] , area in cui in primo luogo si è sviluppata l'epidemia e, di conseguenza, si ha avuto una produzione immediata e maggiore di ricerche.



Nel dettaglio per journals invece, vediamo che il numero di collaborazioni maggiori è detenuto dal Frontiers in Immunology, mentre non appare nessuno dei journals con il maggior numero di collaborazioni del periodo pre-SarS2. Risulta evidente uno spostamento verso le riviste di tipo medico piuttosto che di tipo strettamente immunologico.



[1] Non verificato, presunzione

Infine, si è provato a rispondere al sotto-quesito D, ovvero "Esiste corrispondenza tra le misure di centralità e le riviste più "famoso"? .

In termini assoluti, no. Vediamo perché.

Lo SCImago Journal Rank o SJR indicator è un indicatore che misura il grado di influenza scientifica delle riviste accademiche; esso utilizza il numero di citazioni ricevute da una rivista e l'importanza o il prestigio delle riviste da cui tali citazioni provengono.

Prendiamo in considerazione la classificazione per l'H-index[1].

Secondo lo SCImago Journal Rank possiamo individuare come le 5 riviste più autorevoli:

- 1.Nature: Una delle riviste scientifiche più prestigiose al mondo, pubblica ricerche originali e recensioni su una vasta gamma di discipline scientifiche, tra cui la biologia, la fisica, la chimica e le scienze della Terra. Nature è nota per la qualità delle sue pubblicazioni e per il suo alto impatto sulla comunità scientifica.
- 2.Science: Un'altra rivista scientifica prestigiosa che pubblica ricerche originali e recensioni, con particolare attenzione alle scienze naturali, alla biologia molecolare e alla medicina. Anche Science è nota per la sua grande influenza sulla comunità scientifica e per la qualità delle sue pubblicazioni.
- 3.New England Journal of Medicine: Una delle più importanti riviste mediche al mondo, pubblica ricerche originali, recensioni e editoriali su una vasta gamma di argomenti medici, tra cui la ricerca clinica, la medicina interna e la salute pubblica. Il NEJM è noto per la sua rigorosità scientifica e la sua influenza sulla pratica clinica.
- 4.Cell: Una rivista scientifica specializzata in biologia molecolare e cellulare, che pubblica ricerche originali e recensioni su una vasta gamma di argomenti, tra cui la genetica, la biologia dello sviluppo e la neuroscienza. Cell è conosciuta per la qualità delle sue pubblicazioni e la sua influenza sulla ricerca scientifica.
- 5.The Lancet: Una delle riviste mediche più importanti al mondo, pubblica ricerche originali, recensioni e editoriali su una vasta gamma di argomenti medici, tra cui la ricerca clinica, la medicina interna e la salute pubblica. The Lancet è nota per la sua rigorosità scientifica e la sua influenza sulla pratica clinica.

Tuttavia esaminando i grafi, queste riviste non corrispondono mai con le riviste con il degree più alto, né nel caso dell'arco temporale 1949-2019, né 2020-2023. In spiegazione di questa differenza, si può supporre il prestigio, la qualità e l'impatto sulla comunità scientifica (e non) detenuta da queste riviste porti a una selezione accurata e precisa dei papers da pubblicare, prediligendo la qualità sulla quantità.

Se invece si filtra la classifica del SJR per subject areas “Immunology and Microbiology”, e come subject categories “Virology”, si ottengono le seguenti prime 10 posizioni:

Journal of Virology, PLoS Pathogens, Trends in Microbiology, Cell Host and Microbe, Virology, Journal of General Virology, American Journal of Tropical Medicine and Hygiene, Cellular Microbiology, mBio, Journal of Medical Virology.

1949 - 2019

Per il periodo 1949-2019, i primi 10 nodi ottenuti sono:

- 1.Journal of Virology (index-H: 304) - Il Journal of Virology pubblica ricerche su tutti gli aspetti della virologia, compreso il ciclo di replicazione, la patogenesi, la risposta dell'ospite e l'immunità alle infezioni virali.
- 2.Virology (index-H: 183) - Virology è una rivista che pubblica articoli di ricerca originali, recensioni e comunicazioni brevi su tutti gli aspetti della virologia, comprese le interazioni virus-ospite, la virologia molecolare, la patogenesi virale e i vaccini.
- 3.Advances in Experimental Medicine and Biology (index-H: 130) - Advances in Experimental Medicine and Biology è una rivista multidisciplinare che pubblica articoli su vari argomenti legati alla ricerca biomedica, comprese le malattie infettive, il cancro, l'immunologia e le neuroscienze.
- 4.Journal of General Virology (index-H: 173) - Il Journal of General Virology pubblica articoli di ricerca originali, recensioni e comunicazioni brevi su tutti gli aspetti della virologia, comprese le interazioni virus-ospite, la replicazione virale, la patogenesi e l'immunità alle infezioni virali.
- 5.Virus Research (index-H: 129) - Virus Research è una rivista che pubblica articoli su vari aspetti della virologia, comprese la replicazione virale, la patogenesi, le interazioni virus-ospite e lo sviluppo di vaccini.
- 6.Emerging Infectious Diseases (index-H: 240) - Emerging Infectious Diseases è una rivista che pubblica articoli su vari aspetti delle malattie infettive emergenti e riemergenti, compresa l'epidemiologia, le manifestazioni cliniche, la diagnosi, il trattamento e la prevenzione.
- 7.Veterinary Microbiology (index-H: 135) - Veterinary Microbiology è una rivista che pubblica articoli su vari aspetti della microbiologia legati alla salute degli animali, comprese le malattie infettive, le interazioni ospite-patogeno e la resistenza agli antimicrobici.
- 8.PLOS ONE (index-H: 367) - PLOS ONE è una rivista ad accesso aperto che pubblica articoli di ricerca, recensioni e altri tipi di contenuti scientifici in tutti i campi della scienza e della medicina.
- 9.Archives of Virology (index-H: 117) - Archives of Virology è una rivista che pubblica articoli su vari aspetti della virologia, compresa la replicazione virale, la patogenesi, l'epidemiologia e lo sviluppo di vaccini.
- 10.Viruses (index-H: 101) - Viruses è una rivista che pubblica articoli di ricerca, recensioni e comunicazioni brevi su tutti gli aspetti della virologia, comprese le interazioni virus-ospite, la replicazione virale, la patogenesi e lo sviluppo di vaccini.

Troviamo quindi analogia fra la classifica delle top 10 secondo SJR ed I journals ottenuti per il periodo 49-19 per Journal of Virology, Virology, Journal of General Virology.

4.4 QUESITO D

Si rende in formato tabellare l'elenco dei journals ottenuti dal grafo, ad esclusione di Advances in Experimental Medicine and Biology, Veterinary Microbiology ed Emerging Infectious Diseases, non presenti nella classifica specifica di settore.

Rank	Title	SJR	SJR Quartile	H index	Total Docs. (2021)	Total Docs. (3years)
1	Journal of Virology	2,049	Q1	304	704	2361
5	Virology	1,112	Q2	183	188	745
6	Journal of General Virology	1,317	Q2	173	197	482
12	Virus Research	1,154	Q2	129	279	746
13	Archives of Virology	0,602	Q3	117	412	1237
18	Viruses	1,463	Q2	101	2550	3375
71	PLoS ONE	0,852	Q1	367	16626	49727

Se riordiniamo lo stesso elenco in base al H index, otteniamo un risultato con un solo journal in posizione differente: PloS One compie un balzo in avanti; per le restanti, viene tendenzialmente rispettato l'ordine ottenuto nel grafo.

Title	SJR	SJR Quartile	H index	Total Docs. (2021)
PLoS ONE	0,852	Q1	367	16626
Journal of Virology	2,049	Q1	304	704
Virology	1,112	Q2	183	188
Journal of General Virology	1,317	Q2	173	197
Virus Research	1,154	Q2	129	279
Archives of Virology	0,602	Q3	117	412
Viruses	1,463	Q2	101	2550

Viene quindi rispettata, per il periodo 49-19, la relazione fra la classifica secondo l'SJR e i nodi ottenuti dal grafo.

2020 - 2023

Per quanto riguarda il triennio 2020-2023 le riviste ottenute sono le seguenti:

- 1.Frontiers in Immunology: una rivista peer-reviewed che pubblica ricerche originali e recensioni su tutti gli aspetti dell'immunologia, tra cui la biologia delle cellule immunitarie, la patologia immunitaria e l'immunoterapia.
- 2.Frontiers in Public Health: una rivista peer-reviewed che pubblica ricerche originali e recensioni su una vasta gamma di argomenti relativi alla salute pubblica, tra cui la prevenzione delle malattie, l'epidemiologia e la promozione della salute.
- 3.Journal of Medical Virology: una rivista peer-reviewed che pubblica ricerche originali e recensioni sulla virologia medica, con particolare attenzione alle malattie virali umane, alla diagnosi e alla terapia.
- 4.Signal Transduction and Targeted Therapy: una rivista peer-reviewed che pubblica ricerche originali e recensioni su segnali di trasduzione intracellulari, farmacologia molecolare e terapie mirate.
- 5.Emerging Microbes & Infections: una rivista peer-reviewed che pubblica ricerche originali e recensioni sulla microbiologia, con particolare attenzione ai batteri e ai virus emergenti e alle malattie infettive.
- 6.Nature Communications: una delle riviste scientifiche più prestigiose al mondo, pubblica ricerche originali e recensioni su una vasta gamma di discipline scientifiche, tra cui la biologia, la fisica, la chimica e le scienze della Terra. Nature Communications è nota per la qualità delle sue pubblicazioni e per il suo alto impatto sulla comunità scientifica.
- 7.Frontiers in Medicine: una rivista peer-reviewed che pubblica ricerche originali e recensioni su una vasta gamma di argomenti medici, tra cui la biologia molecolare, la medicina personalizzata e la salute pubblica.
- 8.Medicine (Baltimore): una delle più antiche riviste mediche al mondo, pubblica ricerche originali e recensioni su una vasta gamma di argomenti medici, tra cui la ricerca clinica, la medicina interna e la salute pubblica.
- 9.International Journal of Environmental Research and Public Health: una rivista peer-reviewed che pubblica ricerche originali e recensioni sulla salute ambientale e sulla salute pubblica, tra cui gli effetti degli inquinanti ambientali sulla salute umana e le strategie di prevenzione.
- 10.Clinical Infectious Diseases: una delle riviste mediche più importanti al mondo, pubblica ricerche originali, recensioni ed editoriali sulla ricerca clinica, medicina interna e salute pubblica, con particolare attenzione alle malattie infettive e alle infezioni nosocomiali.

Come è evidente, è avvenuto uno spostamento del topic trattato dallo specifico delle riviste strettamente immunologiche, al generale. Questo non risponde in termini assoluti alle riviste maggiormente famose, ma vediamo se in termini relativi, e rispetto all'H-index, viene rispettata la classifica delle riviste.

Rank	Title	SJR	SJR Best Quartile	H index	Total Docs. (2021)
1032	Frontiers in Immunology	2,331	Q1	155	6058
5165	Frontiers in Public Health	1,298	Q1	64	2245
1366	Journal of Medical Virology	2,656	Q1	137	1076
6556	Signal Transduction and Targeted Therapy	5,157	Q1	53	427
4751	Emerging Microbes and Infections	3,322	Q1	68	225
48	Nature Communications	4,846	Q1	410	7126
6347	Frontiers in Medicine	1,179	Q1	54	2825
1042	Medicine (United States)	0,47	Q3	155	4445
1344	International Journal of Environmental Research and Public Health	0,814	Q1	138	13423
83	Clinical Infectious Diseases	4,394	Q1	353	1947

Rank	Title	SJR	SJR Best Quartile	H index	Total Docs. (2021)
48	Nature Communications	4,846	Q1	410	7126
83	Clinical Infectious Diseases	4,394	Q1	353	1947
1032	Frontiers in Immunology	2,331	Q1	155	6058
1042	Medicine (United States)	0,47	Q3	155	4445
1344	International Journal of Environmental Research and Public Health	0,814	Q1	138	13423
1366	Journal of Medical Virology	2,656	Q1	137	1076
4751	Emerging Microbes and Infections	3,322	Q1	68	225
5165	Frontiers in Public Health	1,298	Q1	64	2245
6347	Frontiers in Medicine	1,179	Q1	54	2825
6556	Signal Transduction and Targeted Therapy	5,157	Q1	53	427

Viene quindi rispettata, per il periodo 49-19, la relazione fra la classifica secondo l'SJR e i nodi ottenuti dal grafo.

4.5 QUESITO E

Al fine di rispondere al quesito E, si è effettuata un'ulteriore analisi della degree e della betweenness dei nodi all'interno del grafo che rappresenta le collaborazioni tra autori di pubblicazioni scientifiche sull'argomento del Coronavirus.

In particolare, si è valutata la posizione dei "First Author" (ovvero gli autori che appaiono come primi autori in una pubblicazione) all'interno del grafo in termini di degree e si è evidenziato se gli stessi presentassero una betweenness maggiore rispetto agli altri autori nel grafo.

La "degree" si riferisce al numero di collaborazioni che un autore ha con altri autori.

Un'autore con una "degree" elevata potrebbe indicare una persona con una vasta rete di collaboratori, indicando una significativa partecipazione alla comunità scientifica.

La betweenness di un nodo in un grafo è una misura di quanto il nodo agisca da intermediario nel passaggio delle informazioni tra altri nodi, dunque un nodo con una betweenness alta viene considerato come un "ponte" tra gruppi di nodi e può avere un ruolo importante nella facilitazione della comunicazione o della collaborazione tra questi gruppi.

L'analisi effettuata mira quindi ad indagare se gli autori con una posizione di intermediazione elevata, cioè quelli che hanno un ruolo chiave nel collegamento tra gruppi di autori o discipline diverse e che possono essere importanti nel facilitare la diffusione delle conoscenze scientifiche e nel promuovere la collaborazione tra diversi settori di ricerca siano principalmente First Author.

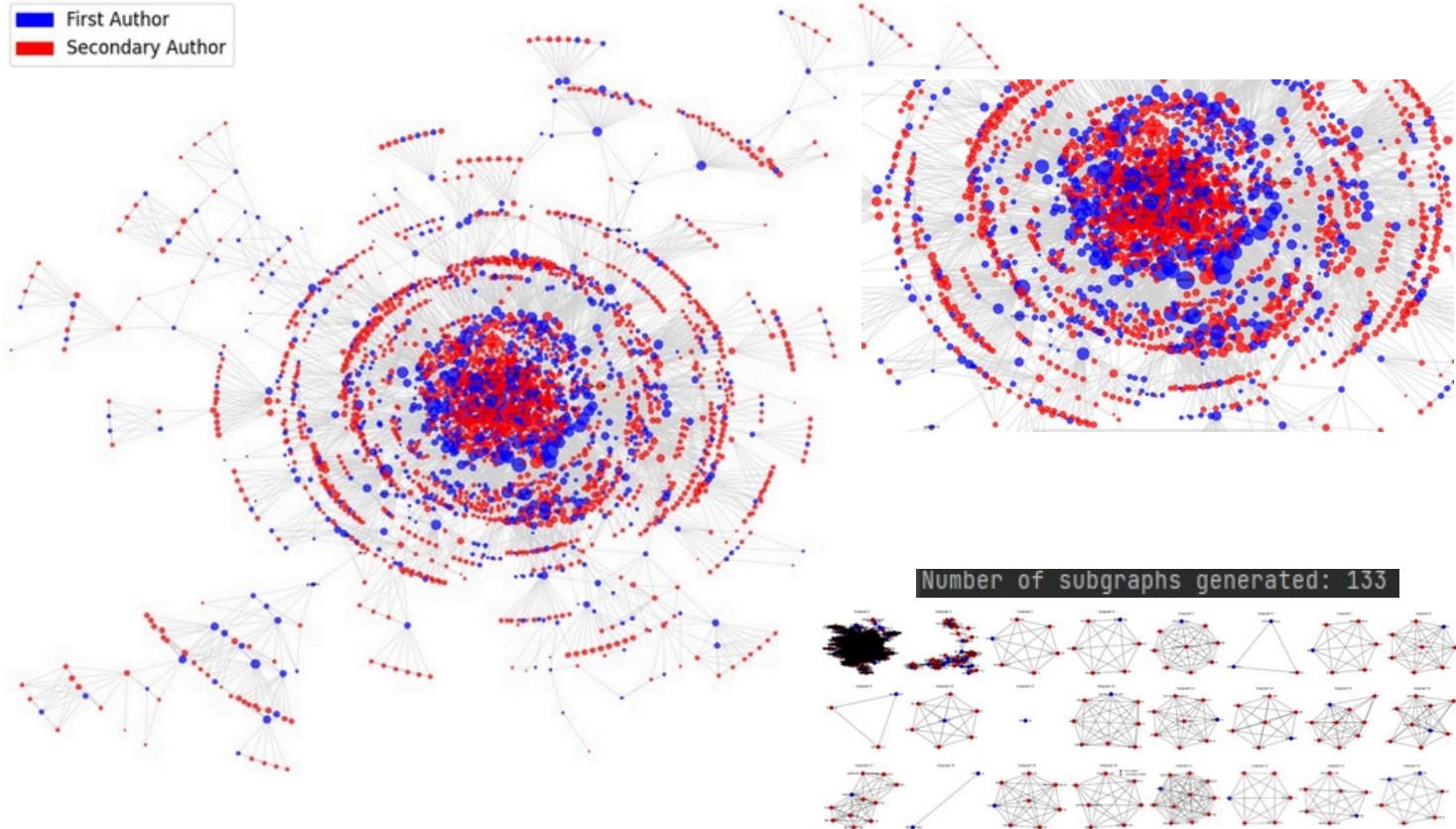
Nell'analisi effettuata, viene calcolata la degree e la betweenness di ciascun nodo nel grafo e vengono identificati i primi 20 nodi con la maggior degree e la maggior betweenness visualizzandoli in un grafico a barre.

Il grafico a barre della "degree" degli autori mostra il numero di collaborazioni di ciascun autore con altri autori. Le barre più alte indicano gli autori con un maggior numero di collaborazioni, che potrebbero essere considerati autori molto attivi o con una vasta rete di collaboratori. Il grafico a barre della "betweenness centrality" degli autori, invece, mostra il grado di intermediarietà di ciascun autore nella rete di coautori. Le barre più alte indicano gli autori con un maggior grado di intermediarietà, cioè quelli che hanno un ruolo chiave nel collegamento tra altri autori o gruppi di autori.

Inoltre, vengono generati dei sottografi eliminando questi 20 nodi con maggior betweenness, al fine di visualizzare l'effetto della loro rimozione sul grafo originale.

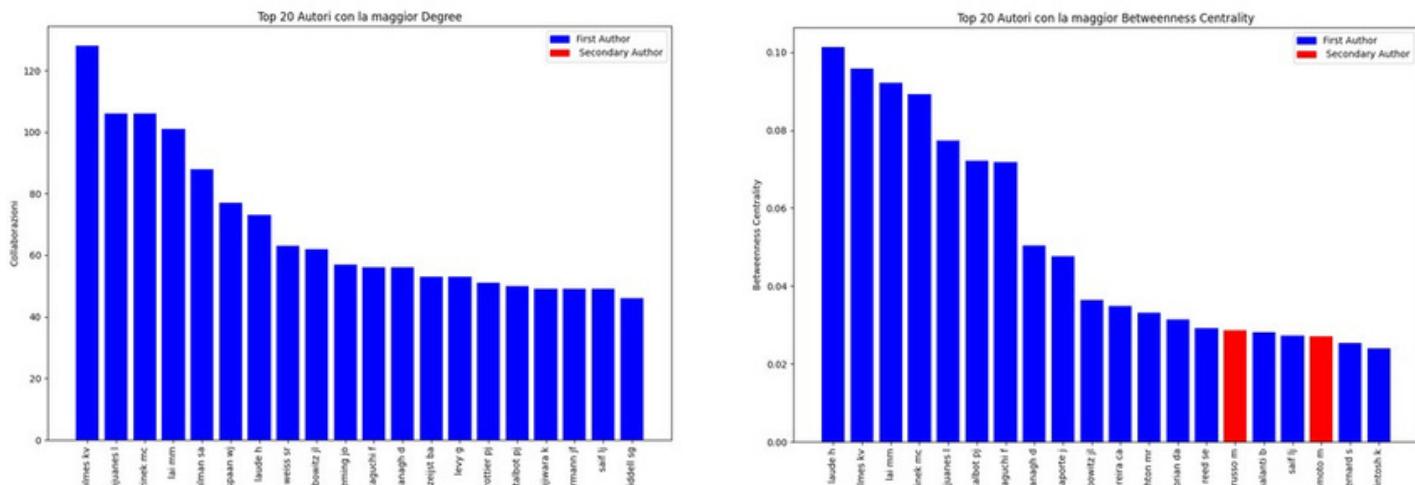
Al fine di avere una visualizzazione più nitida dei nodi del grafo, sono stati generati più grafi e grafici a seconda dei vari periodi di tempo presi in analisi, filtrando di volta in volta il grafo secondo determinati livelli di degree e di numero di pubblicazioni all'aumentare del numero di pubblicazioni nel corso del tempo.

Grafo del Coautorato evidenziato per "First Author" nel 1949_1999 (Degree=5; Pubblicazioni >= 1):



Per il periodo 1949-1999 il grafo si presenta abbastanza sparso in termini di Autori con alti livelli di betwenness ed è possibile infatti generare 133 sottografi a seguito dell'eliminazione dei 20 nodi con i valori maggiori.

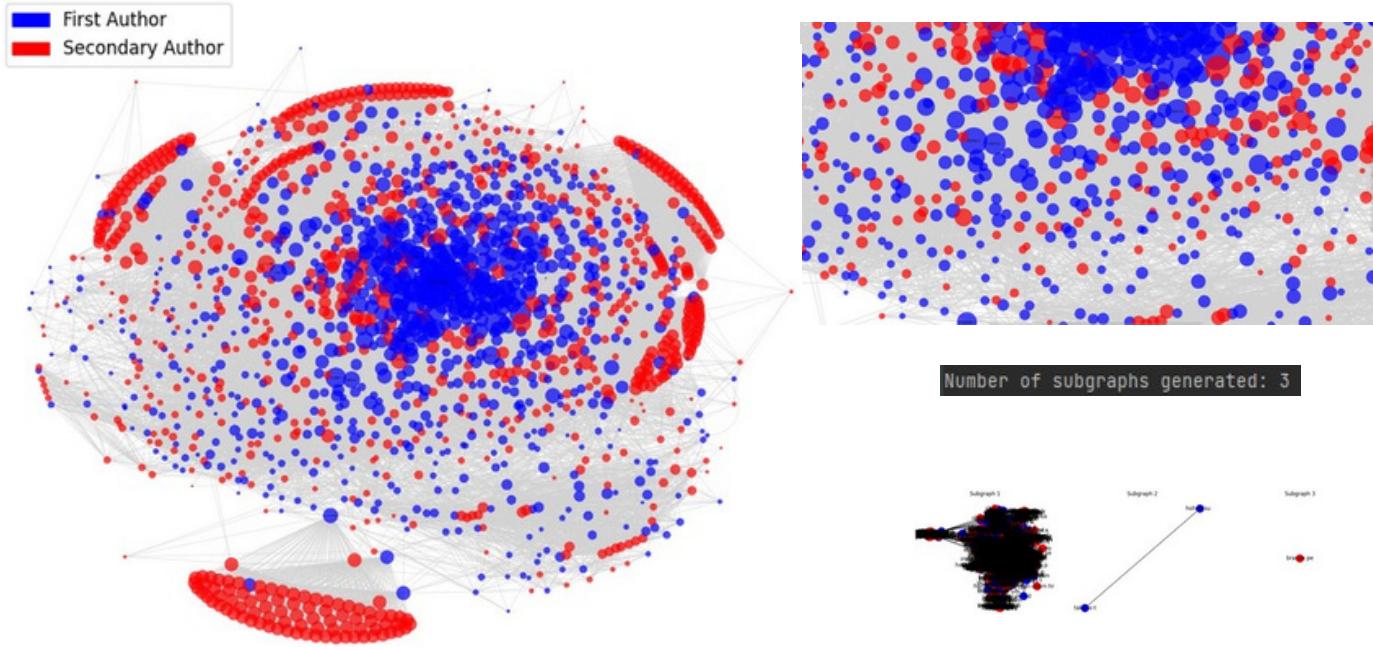
Ciò significa che questi 20 nodi agivano come intermediari chiave nella facilitazione della comunicazione o collaborazione tra altri autori o gruppi di autori. L'eliminazione di questi nodi ha causato la separazione del grafo in 133 sottografi separati, ognuno dei quali potrebbe rappresentare una comunità o una parte isolata del grafo.



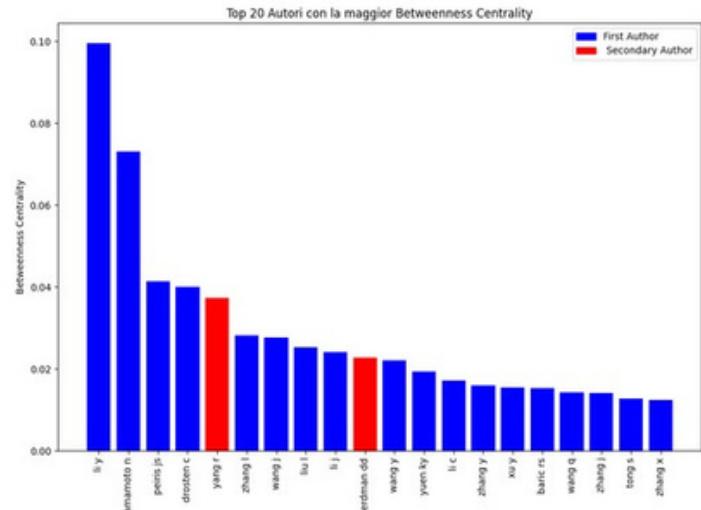
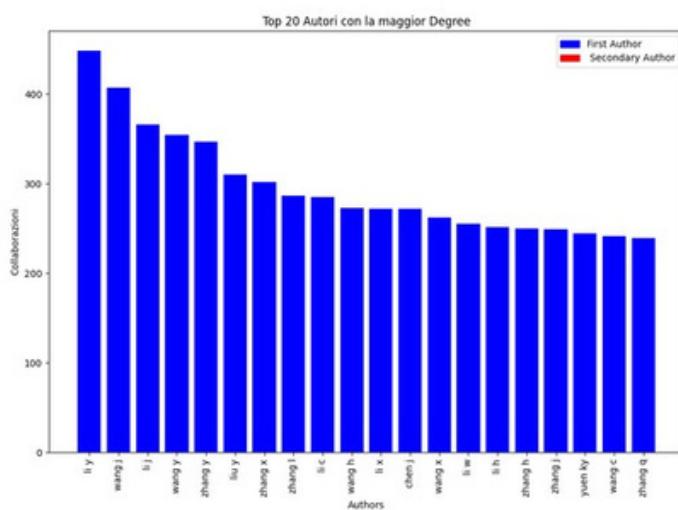
Dalla comparazione tra i grafici a barre possiamo vedere che i nomi degli autori differiscono in parte tra top degree e top betweenness, confermando le posizioni "non centrali" di alcuni nodi non presenti in top degree, come ad esempio l'evidente presenza di due Autori secondari Russo M. e Bernard S.

Con riguardo al quesito principale, possiamo evidenziare come nella maggior parte dei casi analizzati gli autori con più alto livello di intermediazione risultano essere First Author.

**Grafo del Coautoreato evidenziato per "First Author" nel 2000_2018
(Degree=50; Pubblicazioni >= 1):**



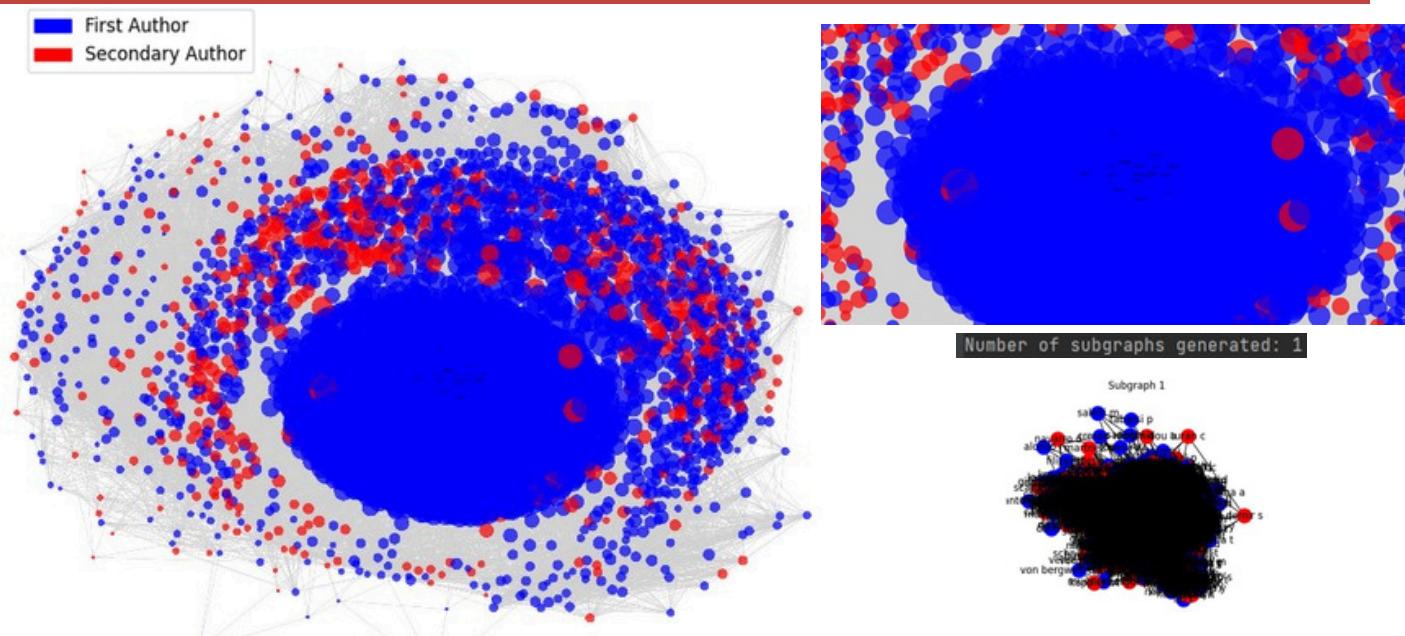
Per il periodo 2000-2018 il grafo si presenta meno sparso rispetto al periodo precedente in termini di Autori con alti livelli di betweenness ed è possibile infatti generare solo 3 sottografi a seguito dell'eliminazione dei 20 nodi con i valori maggiori. Interessanti risultano essere le comunità di autori secondari evidenziate ai lati esterni del grafo.



Dalla comparazione tra i grafici a barre anche in questo caso possiamo vedere che i nomi degli autori differiscono in parte tra top degree e top betweenness, confermando le posizioni "non centrali" di alcuni nodi non presenti in top degree, come ad esempio l'evidente presenza di due Autori secondari Yang R. e Wang Y.

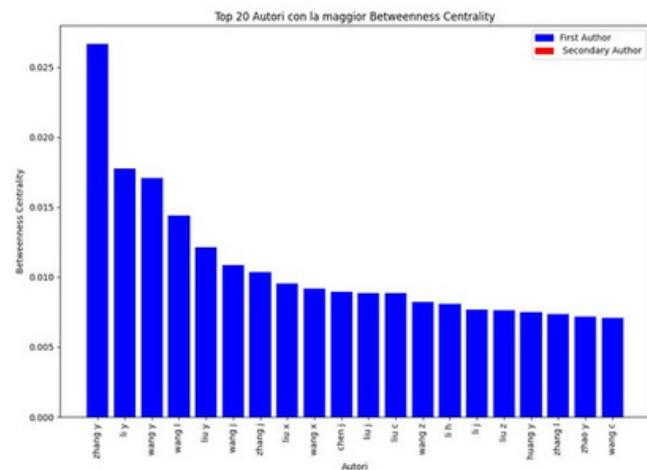
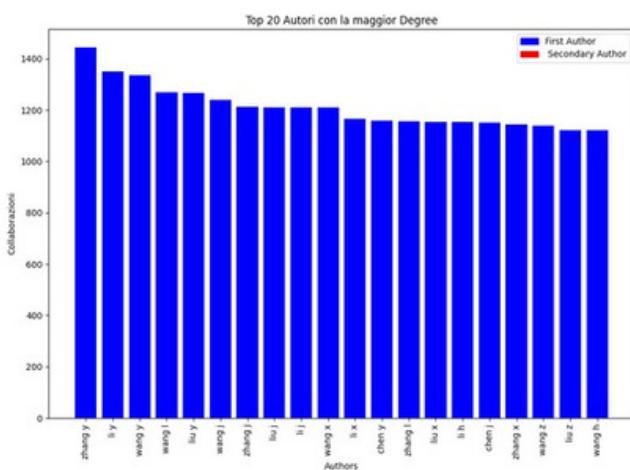
Con riguardo al quesito principale, possiamo evidenziare come nella maggior parte dei casi analizzati gli autori con più alto livello di intermediazione risultano essere First Author.

**Grafo del Coauttorato evidenziato per "First Author" nel 2019_2023
(Degree=250; Pubblicazioni >= 2):**



Per il periodo 2019-2023 il grafo si presenta molto più connesso rispetto al periodo precedente in termini di Autori con alti livelli di betweenness ed è possibile infatti generare solo 1 sottografo a seguito dell'eliminazione dei 20 nodi con i valori maggiori.

Il numero di sottografi generati può dipendere dal fatto che i 20 nodi con la maggior betweenness centrality siano connessi tra loro e formino un unico componente连通的 e allora verrà generato solo un sottografo.



Dalla comparazione tra i grafici a barre in questo caso possiamo vedere che i nomi degli autori non differiscono molto tra top degree e top betweenness, confermando le posizioni "centrali" che assumono i nodi con livelli elevati di betweenness.

Con riguardo al quesito principale, possiamo evidenziare come nella totalità dei casi analizzati gli autori con più alto livello di intermediazione risultano essere First Author.

In conclusione possiamo dire che i First Author svolgono nella maggior parte dei casi un ruolo di connessione tra due o più autori o comunità come sospettavamo, e questo è valido per tutti i periodi analizzati.

Rispetto all'analisi temporale si è evidenziato come con il passare del tempo abbia aumentato il grado di densità del grafo, passando da un grafo abbastanza sparso ad uno maggiormente connesso.

4.6 QUESITO F

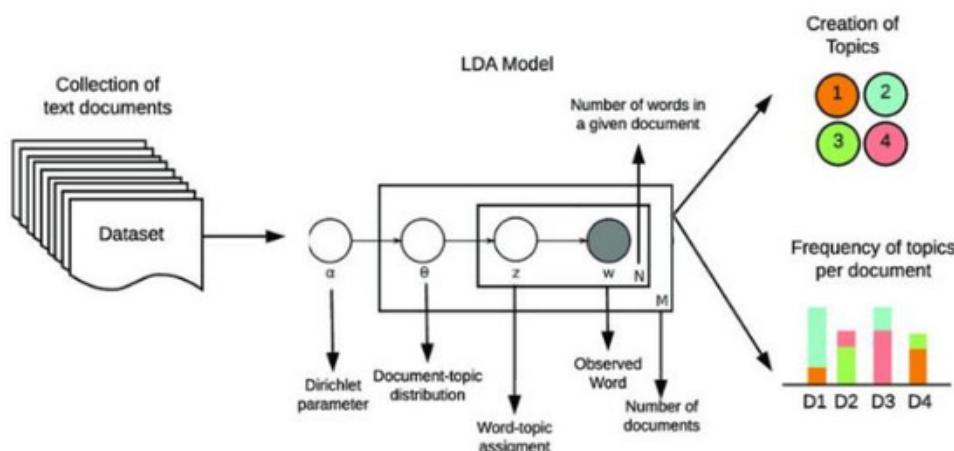
Al fine di poter rispondere al Quesito F , composto dai vari sotto-quesiti sopra citati si è optato per l'implementazione di un algoritmo di Topic Modeling.

Gli algoritmi di Topic Modeling sono dei modelli statistici non supervisionati utilizzati nel Natural Language Processing (NLP) e più in generale nell'apprendimento statistico che consentono di associare un argomento o topic ad un documento presente in una collezione di documenti. Questi algoritmi si basano sul concetto che documenti che trattano lo stesso argomento hanno una probabilità maggiore di contenere termini simili rispetto a documenti nei quali sono trattati argomenti differenti; ovviamente bisogna tenere presente che termini quali articoli, congiunzioni e così via, non possono essere considerati come esplicativi di specifici argomenti in quanto parole troppo comuni presenti in tutti i documenti; per questo motivo, tutti i testi prima dell'applicazione di questi algoritmi di Topic Modeling vengono sottoposti ad una fase di pre-processamento nella quale vengono “ripuliti” da questi termini non decisivi per l'individuazione dei topics.

Dunque, l'obiettivo della Topic Modeling è quello di individuare i termini che compongono un particolare Topic e quindi di poter raggruppare in maniera automatica documenti che trattano uno stesso Topic. Attraverso l'utilizzo di Python, nel nostro caso si è optato per l'utilizzo dell' algoritmo di Latent Dirichlet Allocation (LDA) che è stato applicato ai documenti estratti dal DataBase “PUBMED”, contenente articoli,tesi di ricerca e pubblicazioni scientifiche in ambito “Coronavirus” che formano il Database alla base di questo progetto.

Latent Dirichlet Allocation è un modello probabilistico che consente di estrarre argomenti da un insieme di documenti e si basa sul presupposto che se due termini si trovano spesso in più documenti, probabilmente questi costituiscono il seme di un topic. Più precisamente, nel modello LDA ogni documento è considerato come un insieme di parole che, combinate linearmente tra loro, formano uno o più sottoinsiemi di argomenti latenti. In ciascun documento dunque, potranno essere affrontati diversi topics in base alla presenza e alla frequenza delle parole che compongono ciascun topic; ciò significa che, una volta individuati i differenti i topics trattati nel corpus di documenti, è possibile affermare che un particolare documento tratta uno specifico argomento poiché quello è il topic dominante ma non esclusivo nel documento stesso. Quest'ultimo punto è proprio caratteristico dell'LDA, in quanto in questo modello ogni documento viene visto come una mixture di diversi topics, uno dei quali, generalmente, sarà predominante sugli altri. Di seguito un piccolo grafico che riassume la spiegazione appena esposta.

Latent Dirichlet Allocation (LDA)



Di seguito viene invece fornita una descrizione dei passaggi resisi necessari all'implementazione della LDA e dei relativi risultati ottenuti.

La normalizzazione dei documenti consiste in tutta una serie di elaborazioni dei testi che vengono fatte per rendere quest'ultimi utilizzabili da un sistema di information retrieval. Tra le varie azioni che si intraprendono in questo processo abbiamo lo stemming e/o la lemmatization, l'eliminazione delle stopwords, della punteggiatura e/o dei numeri. Tutto ciò viene eseguito successivamente alla tokenizzazione del testo, cioè la segmentazione del testo completo in singole clausole o in singole parole.

Nel caso in esame nel branch *GridSearch_NWE* abbiamo tokenizzato i documenti per singole word attraverso l'utilizzo del metodo `word_tokenize()` di `nltk`, il quale rappresenta un gruppo di librerie e moduli Python utilizzati nel Machine Learning, nell'Artificial Intelligence e nell'Information Retrieval. Precedentemente a questa operazione abbiamo utilizzato le regular expression per una pre-elaborazione del testo, ovvero abbiamo sostituito la punteggiatura e i numeri con uno spazio. Le RegEx o regular expression rappresentano delle nozioni algebriche che descrivono dei pattern di stringhe; infatti, tali funzioni sono utilizzate per filtrare e confrontare stringhe testuali tra loro, come espresso precedentemente.

Mentre nel branch *TopicModeling_Gensim_NWE* la tokenizzazione è stata effettuata sia per singole parole che per bigrammi e trigrammi, in modo tale da riuscire a costruire dei topics maggiormente esplicativi delle argomentazioni trattate nei vari abstracts. Infatti, l'utilizzo dei bigrammi e trigrammi consente di comprendere meglio il senso delle frasi e di conseguenza, il senso dell'intero documento in quanto, per esempio, parole positive possono essere precedute o seguite da parole negative che alterano totalmente il significato (per esempio: like e not like).

Per quanto riguarda la lemmatizzazione e lo stemming in entrambi i branch abbiamo preferito optare per la sola lemmatizzazione in quanto quest'ultima considera il contesto e converte la parola nella sua forma base significativa, ovvero nel suo lemma. Lo stemming, invece, rimuove i suffissi e ciò molto spesso porta a restituire parole prive di significato a differenza della lemmatization che invece restituisce sempre parole che sono presenti nel dizionario preso come riferimento dal metodo.

2)Applicazione dell' Algoritmo di Topic Modelling e Ottimizzazione dei parametri

Una volta effettuata la normalizzazione del testo, è stato possibile costruire la cosiddetta Bag of Words (BoW). La Bag of Words è un modello utilizzato nell'Information Retrieval che non tiene conto dell'ordine delle parole nel testo ma considera solo le occorrenze delle parole; più specificatamente, il modello BoW costruisce una sorta di vocabolario in cui il testo viene tokenizzato in numeri, e in cui ogni numero rappresenta un identificatore univoco di ciascuna parola e ad ogni identificatore è associato un secondo numero che invece rappresenta la frequenza di quella parola nel documento codificato. Il risultato che si ottiene è unalista di tuple in cui il primo numero rappresenta appunto l'identificatore della parola e il secondo il numero di occorrenze se si utilizza il CountVectorizer, oppure la sua TF-IDF se si utilizza il TfIdfVectorizer. Nel nostro progetto è stato utilizzato esclusivamente il CountVectorizer poichè il modello LDA utilizza la word count.

Generalmente l'utilizzo della TF-IDF è da preferire in quanto il conteggio delle parole porta a considerare molto importanti termini come articoli e congiunzioni che, pur essendo molto frequenti, in realtà risultano poco significative nei vettori codificati. L'alternativa è, appunto, quella di calcolare la frequenza inversa delle parole, e il metodo di gran lunga più popolare è chiamato TF-IDF, che sostanzialmente si adegua al fatto che, come precedentemente detto, alcune parole appaiono più frequentemente in generale ma non per questo sono più significative. TF-IDF è un acronimo che significa "Term Frequency - Inverse Document Frequency" le quali rappresentano le metriche risultanti assegnate a ciascun token.

In particolare, la Term Frequency riassume la frequenza con cui una determinata parola appare all'interno di un documento, quindi sarà tanto più elevata tanto più il termine è frequente; mentre l'Inverse Document Frequency misura la frequenza inversa di una parola in tutti i documenti, ciò significa che l>IDF sarà molto alta nei termini specifici per uno specifico documento e molto bassa nei termini molto comuni.

Per poter usufruire del vantaggio offerto dall'utilizzo della TF-IDF anche con il modello LDA, sono stati successivamente implementati degli algoritmi in grado di generare dei grafici interattivi, che permettono, attraverso la modifica in tempo reale di un determinato parametro LAMBDA, di poter filtrare i termini più "salienti" per un Topic proprio in base al meccanismo di TF-IDF.

$$w_{x,y} = tf_{x,y} \times \log \left(\frac{N}{df_x} \right)$$

TF-IDF

Term x within document y

$tf_{x,y}$ = frequency of x in y
 df_x = number of documents containing x
 N = total number of documents

Inoltre nel branch GridSearch_NWE, abbiamo invece utilizzato il Vectorizer offerto dalla libreria Scikit-learn, la quale rappresenta una libreria professionale utilizzata nel Machine Learning, ovvero CountVectorizer a cui abbiamo passato come parametri:

- il tokenizer da noi costruito;
- un max_df pari a 0,80, il quale permette di ignorare tutti quei termini che sono presenti in più del 80% degli abstracts;
- un min_df pari a 0,20, il quale permette di ignorare tutti quei termini che sono presenti in meno del 20% degli abstracts, poiché una parola con una bassissima frequenza potrebbe indurre solo rumore nel modello;
- infine, max_features è stato impostato a 1000, e ciò indica che nella costruzione del vocabolario verranno considerate solo le prime 1000 parole in ordine di term frequency.

La scelta di settare il parametro precedente max_features, n_components a 6 e learning_decay a 0.5 del modello di Topic modeling è dovuta alla media dei risultati ottenuti tramite l'utilizzo di una Gridsearch applicata ai vari periodi di tempo analizzati, così suddivisi:

1949 → 1999

```
Choosing Optimal Hyperparameter
Best Score Likelihood: -24011.759
Best parameters set:
  model__learning_decay: 0.5
  model__n_components: 7
  vect__max_features: 1000
Tn [3]:
```

2000 → 2019

```
Choosing Optimal Hyperparameter
Best Score Likelihood: -130827.656
Best parameters set:
  model__learning_decay: 0.5
  model__n_components: 8
  vect__max_features: 1000
  vect__ngram_range: [1, 2]
```

2020

```
Choosing Optimal Hyperparameter
Best Score Likelihood: -558442.999
Best parameters set:
  model__learning_decay: 0.5
  model__n_components: 6
  vect__max_features: 1000
  vect__ngram_range: [1, 2]
```

2021

```
Choosing Optimal Hyperparameter
Best Score Likelihood: -1077982.945
Best parameters set:
  model__learning_decay: 0.5
  model__n_components: 7
  vect__max_features: 1000
  vect__ngram_range: [1, 2]
```

2022

```
Choosing Optimal Hyperparameter
Best Score Likelihood: -1098100.042
Best parameters set:
  model__learning_decay: 0.5
  model__n_components: 4
  vect__max_features: 1000
  vect__ngram_range: [1, 2]
```

2023

```
Choosing Optimal Hyperparameter
Best Score Likelihood: -139631.678
Best parameters set:
  model__learning_decay: 0.5
  model__n_components: 7
  vect__max_features: 1000
  vect__ngram_range: [1, 2]
```

La Gridsearch rappresenta una metodologia utilizzata nel processo di ottimizzazione dei parametri che permette appunto di individuare i coefficienti ottimali relativi a tutti i possibili iperparametri dell'algoritmo di Machine Learning. Alla luce delle molteplici prove effettuate è da sottolineare che, nel caso della Topic Modeling da noi effettuata, il numero ottimale di Topics da utilizzare nella stessa è risultato essere 6, valore che verrà utilizzato dunque nelle successive applicazioni del modello LDA. Il branch GridSearch_NWE è stato dedicato proprio all'implementazione di tale algoritmo. In tale branch, alla Gridsearch è stata passata una Pipeline contenente come primo step un CountVectorizer e come secondo step il modello di Topic Modeling, ovvero l'oggetto LatentDirichletAllocation. Una Pipeline può essere vista come una sorta di contenitore di tutti gli step di processamento che permette di ottimizzare e semplificare notevolmente la scrittura del codice in quanto, una volta costruito questo "contenitore", è lui che si occupa di eseguire tutti gli step di processamento sulla base dei dati consegnatogli; senza l'ausilio della Pipeline dovremmo eseguire ogni step separatamente e poi successivamente mettere assieme i risultati ottenuti nei vari step per ottenere il risultato del modello scelto. L'impostazione dei parametri, come precedentemente indicato, è stato frutto di molteplici tentativi in cui i parametri min_df e max_df sono stati settati in maniera differente; la scelta è stata fatta andando a cercare un compromesso tra un buon livello dell'indice Likelihood e un valore dei parametri max_df e min_df che non andasse a scartare troppi termini.

Di seguito si mostrano i risultati ottenuti con il modello LDA che è stato eseguito impostando un numero di topics pari a 6, numero ottimale indicato precedentemente dalla GridSearch.

Il codice dei seguenti risultati si trova nel branch TopicModeling_Gensim_NWE, dove, per visualizzare il contenuto di ciascun topic, abbiamo optato per l'utilizzo della tecnica delle WordCloud, le quali rappresentano i termini secondo un font di grandezza differente in base al peso associato a ciascun termine, peso che in questo caso è rappresentato dalla Term Frequency.

```
#####
# LATENT DIRICHLET ALLOCATION: #####
#####

^^^^ Contributo delle 10 parole più importanti per 6 Topics: ^^^^
[(0, '0.044*"group" + 0.040*"test" + 0.037*"sample" + 0.033*"age" + 0.026*"old" + 0.025*"positive" + 0.021*"high" + 0.018*"serum" + 0.018*"rotavirus" + 0.017*"study"),
(1, '0.088*"virus" + 0.030*"antibody" + 0.029*"viral" + 0.024*"infection" + 0.020*"infect" + 0.017*"isolate" + 0.016*"specific" + 0.014*"coronavirus" + 0.013*"titer" + 0.013*"contain"),
(2, '0.012*"test" + 0.010*"surface" + 0.010*"level" + 0.009*"resistance" + 0.009*"virus" + 0.009*"large" + 0.008*"degree" + 0.008*"sensitive" + 0.008*"end" + 0.008*"growth"),
(3, '0.054*"leader" + 0.026*"expose" + 0.025*"exposure" + 0.021*"outbreak" + 0.020*"survey" + 0.018*"turkey" + 0.018*"illness" + 0.013*"volunteer" + 0.012*"health" + 0.012*"risk"),
(4, '0.024*"day" + 0.023*"infection" + 0.018*"disease" + 0.015*"detect" + 0.014*"induce" + 0.011*"activity" + 0.009*"challenge" + 0.008*"infect" + 0.008*"associate" + 0.008*"occur"),
(5, '0.074*"cell" + 0.042*"protein" + 0.023*"mouse" + 0.022*"virus" + 0.021*"strain" + 0.016*"sequence" + 0.015*"gene" + 0.012*"show" + 0.010*"coronaviruse" + 0.010*"tgev")]
```

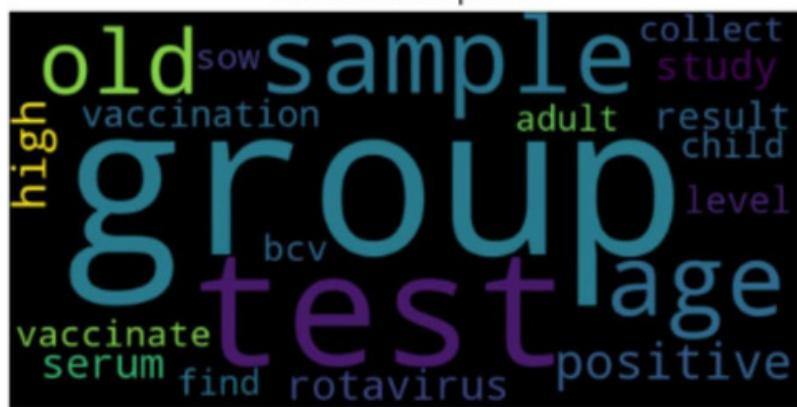
WordCloud associate a ciascuno dei 6 Topics del modello LDA:

-TOPIC #0: Studio epidemiologico del rotavirus.

I termini potrebbero suggerire l'importanza dello studio dei campioni di gruppi di età diversi, compresi i campioni sierologici, per valutare la diffusione e l'incidenza di malattie come il rotavirus.

La parola "positive" potrebbe riferirsi ai test che possono rilevare la presenza di anticorpi specifici e/o del virus stesso nei campioni. La parola "high" potrebbe suggerire la valutazione di tassi elevati di incidenza della malattia. In sintesi, i termini potrebbero far riferimento alla valutazione epidemiologica della diffusione del rotavirus attraverso l'analisi di campioni di gruppi di età diversi, utilizzando test sierologici per rilevare la presenza di anticorpi specifici e del virus stesso.

WordCloud Topic#0



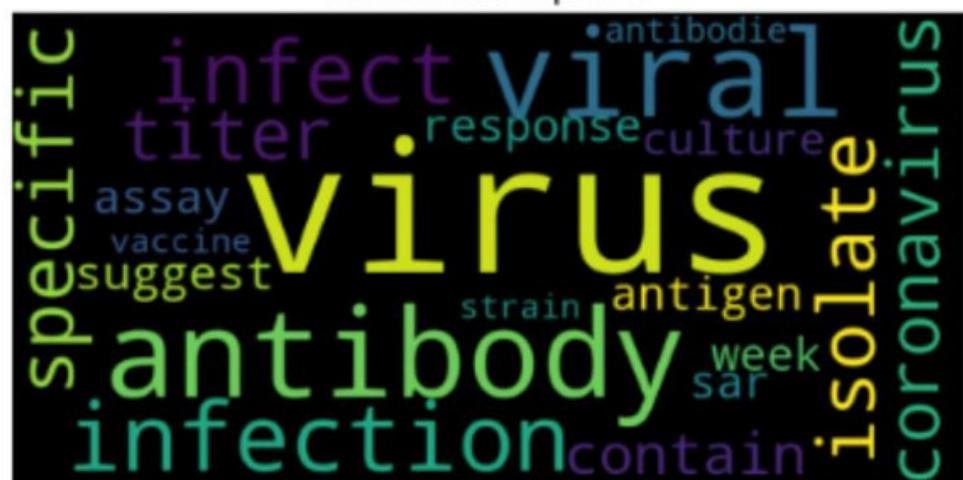
TOPIC °1: Comprensione, controllo e risposta delle infezioni virali.

I termini potrebbero suggerire l'importanza della comprensione dei virus, delle infezioni virali e delle risposte del sistema immunitario, inclusa la produzione di anticorpi specifici per isolare e contenere i virus.

Il termine "coronavirus" indica la specificità della ricerca su virus specifici, mentre "titer" può essere utilizzato per quantificare la risposta anticorpale del sistema immunitario.

In sintesi, i termini si riferiscono all'importanza della comprensione e del controllo delle infezioni virali attraverso la produzione di anticorpi specifici e la ricerca specifica sui virus.

WordCloud Topic#1

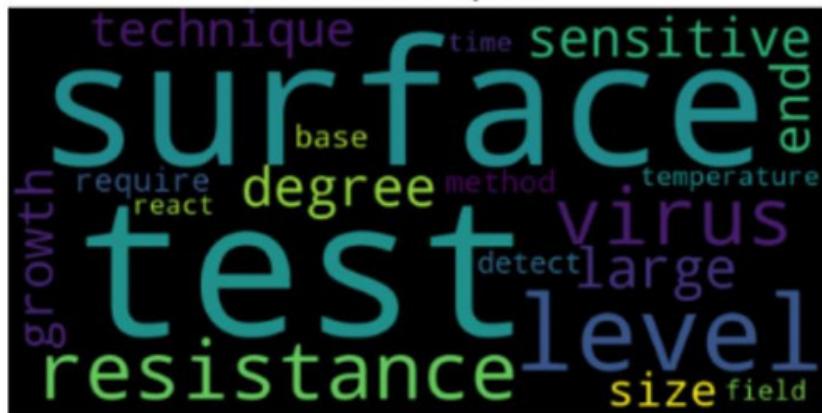


TOPIC#2:Valutazione resistenza virus e test.

I termini potrebbero suggerire l'importanza della valutazione dei livelli di resistenza del virus e della sensibilità dei test diagnostici. Il termine "surface" suggerisce la valutazione dell'effetto dell'ambiente sul virus, mentre "large" può suggerire la valutazione dell'impatto su popolazioni numerose.

La parola "degree" può essere utilizzata per quantificare i livelli di resistenza e sensibilità, mentre "growth" può indicare l'importanza della valutazione della capacità di replicazione del virus. In sintesi, i termini potrebbero riferirsi all'importanza della valutazione della resistenza e della sensibilità del virus e dei test diagnostici, nonché dell'impatto dell'ambiente sulla diffusione del virus.

WordCloud Topic#2

**TOPIC#3:Salute pubblica ed controllo epidemie** I termini indicati sembrano riferirsi all'ambito della salute pubblica e della gestione delle epidemie.

"Leader" potrebbe riferirsi alla figura di un responsabile o di un coordinatore nella gestione di un'epidemia, mentre "expose" e "exposure" potrebbero indicare l'esposizione a una malattia infettiva e il rischio ad essa associato. "Outbreak" indica la diffusione di una malattia infettiva in una determinata popolazione, mentre "survey" potrebbe indicare l'utilizzo di rilevazioni e indagini per mappare la diffusione dell'epidemia. "turkey" e "illness" potrebbero essere indicativi del tipo di malattia in questione, mentre "volunteer" potrebbe riferirsi a coloro che si offrono volontari per partecipare alle indagini o alle sperimentazioni di un vaccino. Infine, "health" e "risk" indicano l'attenzione alla salute pubblica e alla gestione del rischio legato all'epidemia.

In sintesi, i termini elencati sembrano essere riferiti all'ambito della salute pubblica e della gestione delle epidemie, con particolare attenzione alla figura del leader, all'esposizione e al rischio, alla diffusione dell'epidemia e alla salute pubblica.

WordCloud Topic#3



TOPIC#4: Tempistiche dell'infezione ed effetti sulla salute:

I termini elencati sembrano essere tutti legati al tema delle malattie infettive e della loro rilevazione e diffusione. In particolare si riferiscono alla durata e alla diffusione dell'infezione, ai sintomi, alla diagnosi e alla risposta immunitaria. Essi includono anche riferimenti alle sfide nella ricerca di trattamenti efficaci, all'attività fisica in relazione alla salute e alla trasmissione del virus da una persona infetta a una persona sana.

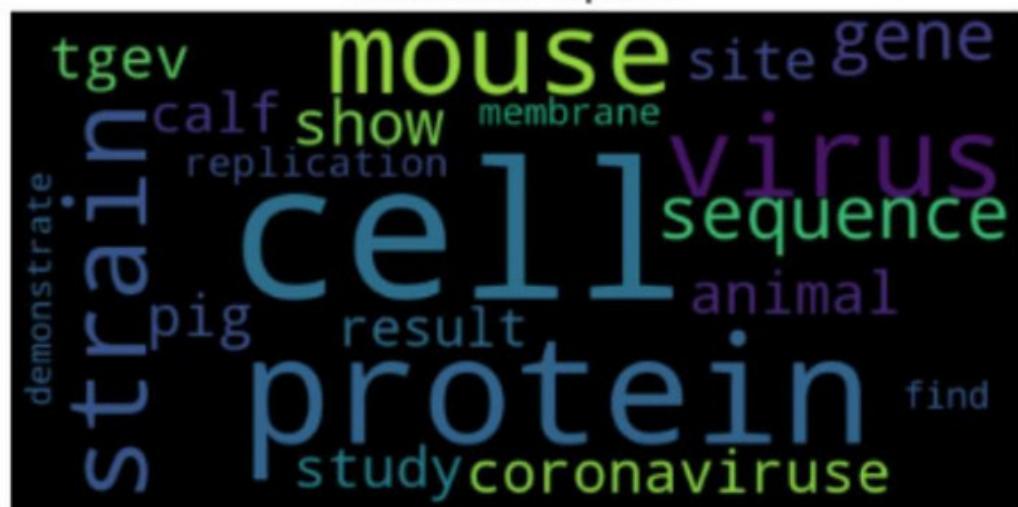
WordCloud Topic#4

**TOPIC#5: Virologia e biologia molecolare**

I termini elencati sembrano essere legati alla biologia molecolare e alla virologia. "Cell" potrebbe indicare il tipo di cellula su cui si sta effettuando la ricerca, mentre "protein" e "gene" sono chiaramente legati alla loro espressione all'interno della cellula. "Mouse" potrebbe riferirsi all'animale di laboratorio utilizzato per lo studio, mentre "virus" e "strain" indicano il tipo di agente infettivo e la sua variante specifica. "Sequence" potrebbe indicare l'analisi della sequenza genetica del virus o della cellula ospite, mentre "show" potrebbe indicare i risultati ottenuti dalla ricerca. Infine, "coronavirus" e "tgev" sono entrambi virus appartenenti alla famiglia dei coronavirus e potrebbero essere oggetto di studio per la ricerca.

In sintesi, i termini indicati sembrano riferirsi alla ricerca in biologia molecolare e alla virologia, con particolare attenzione alle proteine, alle sequenze genetiche, ai virus e alle cellule ospiti utilizzate per la ricerca.

WordCloud Topic#5



Sulla base dei risultati globali ottenuti, che indicano il modello LDA con 6 Topics come preferibile, abbiamo voluto approfondire l'analisi di tale modello con alcuni strumenti molto interessanti che permettono la generazione di grafici interattivi sotto forma di pagine HTML, permettendo così un'analisi altamente dinamica e dal contenuto informativo imparagonabile rispetto ai grafici tradizionali.

Nel nostro caso specifico utilizzeremo le librerie TSNE e pyLDAvis.

- TSNE : Tale sigla sta ad indicare il t-Distributed Stochastic Neighbor Embedding (t-SNE) che è una tecnica

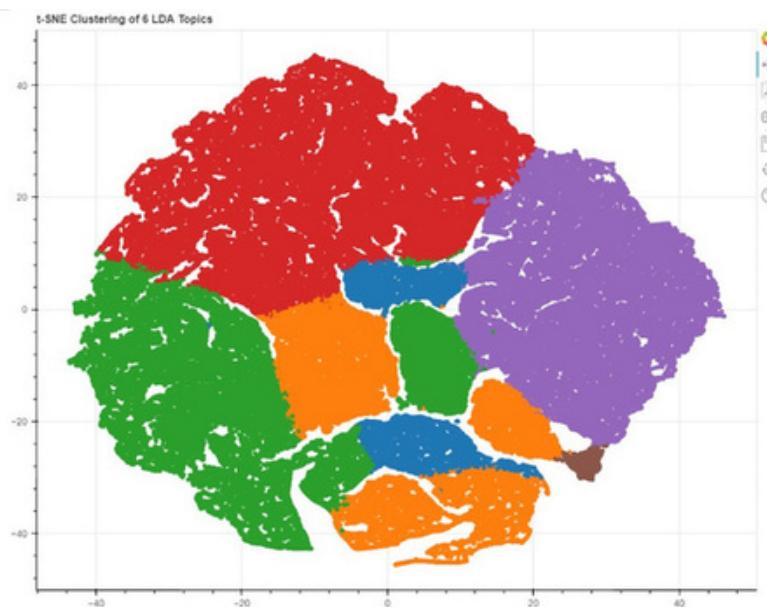
in grado di visualizzare i dati ad alta dimensione su uno spazio bi o tridimensionale. t-SNE è un algoritmo non lineare per la riduzione della dimensionalità in grado di rilevare la presenza di Cluster di diverse dimensioni tramite una funzione che consente di rappresentare punti di dati simili vicini tra loro e, allo stesso tempo, dati diversi lontani tra loro. Tale risultato è ottenibile convertendo le distanze euclidee tra i punti dati in probabilità condizionali che rappresentano le somiglianze. Questa tecnica si articola in due fasi principali:

- Una prima fase in cui viene costruita una distribuzione di probabilità che ad ogni coppia di punti nello spazio originale ad alta dimensionalità associa un valore di probabilità elevato se i due punti sono simili, basso se sono dissimili.

- Una seconda fase in cui viene definita una seconda distribuzione di probabilità, analoga alla prima, nello spazio a dimensione ridotta. L'algoritmo quindi minimizza la divergenza di Kullback-Leibler delle due distribuzioni tramite il gradiente discendente, riorganizzando i punti nello spazio a dimensione ridotta.

Di seguito sono mostrati i risultati che mostrano le caratteristiche con le quali è stato generato il grafico t-SNE nel nostro caso specifico e accanto è mostrata la visualizzazione del dataset con t-SNE vera e propria, dove è possibile vedere in maniera intuitiva che i punti rappresentanti i documenti sono raggruppati in 6 diversi cluster in base alle loro caratteristiche semantiche. I diversi colori dei cluster indicano l'assegnazione del documento ad un argomento specifico. Data la quasi ottima distinzione tra Clusters, si può affermare che tale rappresentazione sia buona, e rispetto a quei documenti che si posizionano in un cluster di colore diverso, lo fanno perché probabilmente non hanno un topic dominante molto forte

```
AAAA Informazioni relative alla generazione dei Cluster: AAAA
C:\Users\Francesco\PycharmProjects\pythonProjectNLP\venv\lib\site-packages\torch\utils\benchmark\utils.py:10: FutureWarning,
[t-SNE] Computing 91 nearest neighbors...
[t-SNE] Indexed 152702 samples in 0.203s...
[t-SNE] Computed neighbors for 152702 samples in 10.734s...
[t-SNE] Computed conditional probabilities for sample 1000 / 152702
[t-SNE] Computed conditional probabilities for sample 2000 / 152702
[t-SNE] Computed conditional probabilities for sample 3000 / 152702
[t-SNE] Computed conditional probabilities for sample 4000 / 152702
[t-SNE] Computed conditional probabilities for sample 5000 / 152702
[t-SNE] Computed conditional probabilities for sample 6000 / 152702
[t-SNE] Computed conditional probabilities for sample 7000 / 152702
[t-SNE] Computed conditional probabilities for sample 8000 / 152702
[t-SNE] Computed conditional probabilities for sample 9000 / 152702
[t-SNE] Computed conditional probabilities for sample 10000 / 152702
[t-SNE] Computed conditional probabilities for sample 11000 / 152702
[t-SNE] Computed conditional probabilities for sample 12000 / 152702
[t-SNE] Computed conditional probabilities for sample 13000 / 152702
[t-SNE] Computed conditional probabilities for sample 14000 / 152702
[t-SNE] Computed conditional probabilities for sample 15000 / 152702
[t-SNE] Computed conditional probabilities for sample 16000 / 152702
[t-SNE] Computed conditional probabilities for sample 17000 / 152702
[t-SNE] Computed conditional probabilities for sample 18000 / 152702
[t-SNE] Computed conditional probabilities for sample 19000 / 152702
[t-SNE] Computed conditional probabilities for sample 20000 / 152702
...
[t-SNE] Computed conditional probabilities for sample 146000 / 152702
[t-SNE] Computed conditional probabilities for sample 147000 / 152702
[t-SNE] Computed conditional probabilities for sample 148000 / 152702
[t-SNE] Computed conditional probabilities for sample 149000 / 152702
[t-SNE] Computed conditional probabilities for sample 150000 / 152702
[t-SNE] Computed conditional probabilities for sample 151000 / 152702
[t-SNE] Computed conditional probabilities for sample 152000 / 152702
[t-SNE] Computed conditional probabilities for sample 152702 / 152702
[t-SNE] Mean sigma: 0.002346
[t-SNE] KL divergence after 250 iterations with early exaggeration: 100.972252
[t-SNE] KL divergence after 1000 iterations: 3.497343
```



- **pyLDAvis:** grazie a questa libreria viene generato un grafico interattivo che fornisce sia una visione globale degli argomenti e di come differiscono l'uno dall'altro, sia una visione locale consentendo allo stesso tempo un'analisi approfondita dei termini più strettamente associati a ciascun singolo argomento. La visualizzazione ha due parti fondamentali:

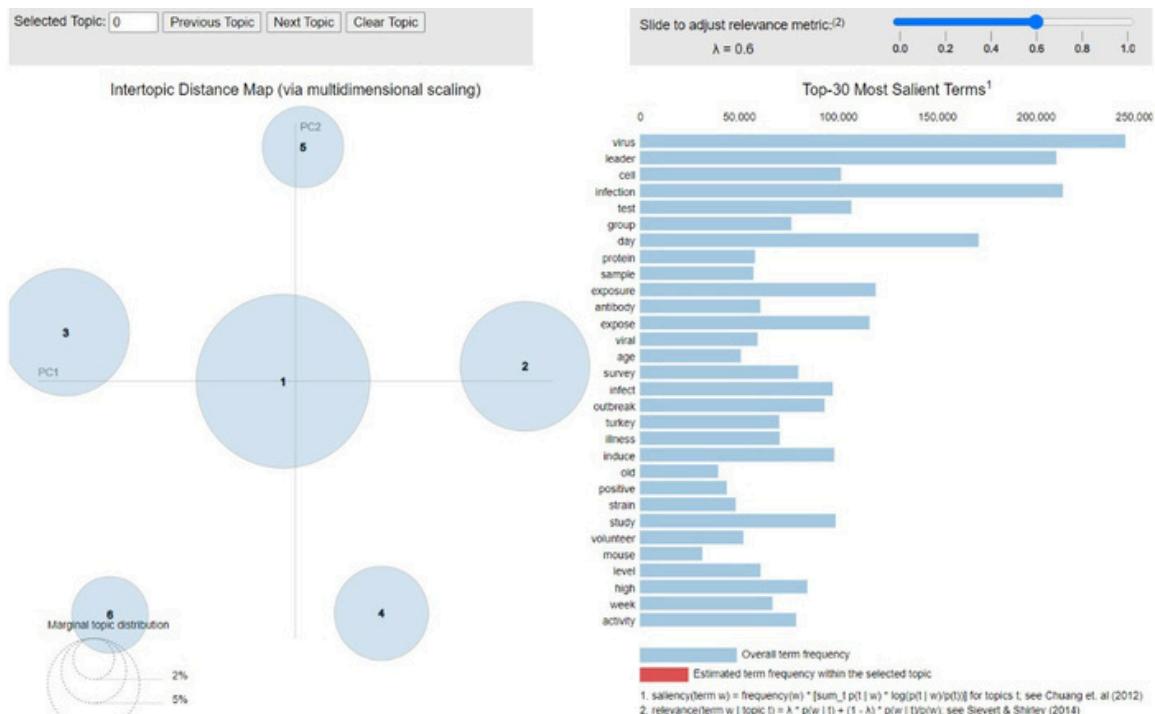
1) Il pannello di sinistra, chiamato Intertopic Distance Map, visualizza gli argomenti come cerchi nel piano bidimensionale e più il cerchio è grande, più la sua frequenza relativa rispetto al Corpus è alta.

Gli indici all'interno del cerchio indicano la popolarità di ciascun topic e, in particolare con il numero 1 viene indicato l'argomento più popolare e con il numero 6 l'argomento meno popolare.

La poca distanza tra due "bolle" rappresenta la maggior somiglianza tra argomenti e viceversa. Tuttavia, questa è solo un'approssimazione della matrice di similarità dell'argomento originale poiché la distribuzione spaziale di tutti e 6 gli argomenti viene riadattata alle due dimensioni in quanto il grafico utilizzato è un grafico bidimensionale. Un singolo argomento può essere selezionato per un esame più attento facendo clic sul suo cerchio o inserendo il suo numero nella casella "argomento selezionato" in alto a sinistra.

2) Il pannello di destra raffigura un grafico a barre orizzontali le cui barre rappresentano i 30 termini più utili per interpretare l'argomento attualmente selezionato a sinistra; i termini vengono mostrati seguendo un ordine decrescente di rilevanza. Quando nessun argomento è selezionato nel grafico a sinistra, il grafico a barre mostra i primi 30 termini più "salienti" nel corpus. La salienza di un termine è una misura sia di quanto sia frequente il termine nel corpus e sia di quanto sia utile nel distinguere tra diversi argomenti. La barra blu rappresenta la frequenza complessiva del termine e la barra rossa indica la frequenza stimata del termine all'interno dell'argomento selezionato. Quindi, se una barra appare sia rossa che blu, significa che il termine è presente in più di un topics.

Il cursore λ consente di classificare le parole in base alla pertinenza del termine; si tratta di una metrica regolabile che bilancia la frequenza di un termine in un particolare argomento con la frequenza del termine nell'intero corpus di documenti. Per impostazione predefinita, i termini di un argomento sono classificati in ordine decrescente in base alla loro probabilità specifica per argomento ($\lambda = 1$). Quando aggiustiamo la pertinenza con un lambda più basso significa che i termini che sono frequenti in tutti gli argomenti vengono penalizzati, quindi riducendo λ le parole che sono molto frequenti nell'intero corpus assumeranno una rilevanza minore. Il valore "ottimale" suggerito è di 0,6.

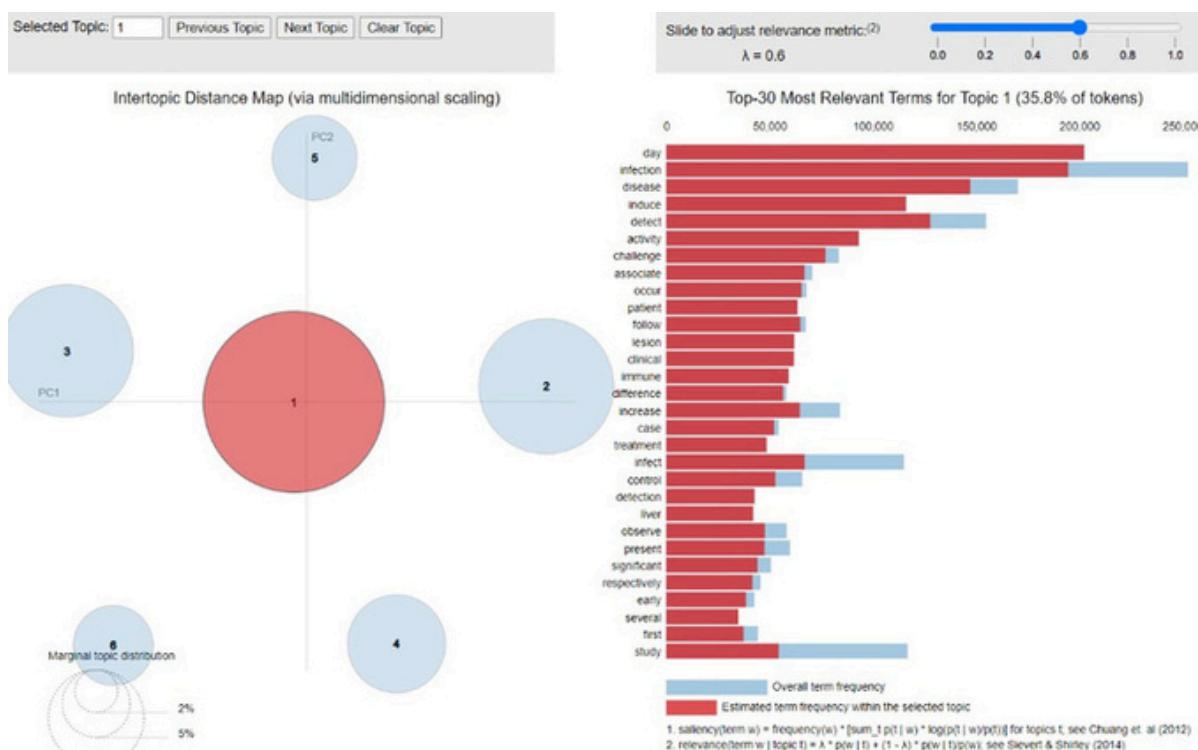


ANALISI GLOBALE:

Secondo un'analisi generale rispetto al Corpus delle fonti documentali analizzate, i 30 Termini più rilevanti sembrano riguardare l'ambito della salute e della malattia, in particolare la malattia COVID-19 e il virus SARS-CoV-2. Essi includono riferimenti al virus stesso, alla sua struttura e alle sue proprietà, alla cellula ospite infettata, all'infezione e alla diffusione del virus nel corpo, alla diagnosi dell'infezione tramite test, alla risposta immunitaria umana alla malattia, all'esposizione al virus e alla presenza di anticorpi, alla diffusione dell'infezione in gruppi di persone, all'età e ad altre caratteristiche dei pazienti infetti, alla valutazione degli effetti dell'attività fisica sulla salute in relazione alla malattia, alla diffusione del virus in diverse comunità e popolazioni, alla diffusione del virus anche in animali come i tacchini, e agli studi scientifici volti a comprendere la malattia e a sviluppare nuovi trattamenti e vaccini.

Riassumendolo in 4 parole, gli argomenti trattati possono essere interpretati come in riferimento alla **"Ricerca sull'infezione da Coronavirus"**

Tale possibile interpretazione sembra essere perfettamente in linea con la base di argomenti oggetto di studio.



TOPIC #1:

Il senso comune che emerge dai termini elencati sembra essere inherente allo studio della malattia da coronavirus, i suoi sintomi, il suo trattamento e la sua prevenzione.

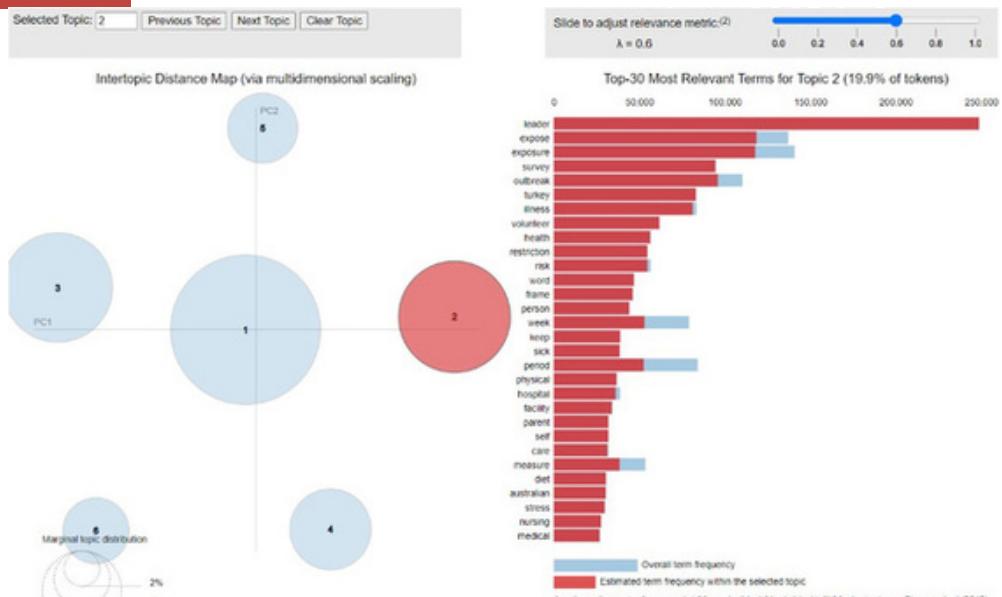
In particolare, il virus può causare infezioni che si manifestano con sintomi clinici come lesioni e aumenti di attività immunitaria, e la sua diffusione può essere controllata mediante la diagnosi precoce, la rilevazione e il monitoraggio dei casi di infezione.

Inoltre, il trattamento delle infezioni da coronavirus rappresenta una sfida importante, poiché il virus è in grado di indurre una serie di patologie diverse, che possono colpire vari organi del corpo come il fegato e l'orecchio, e richiede un'attenta osservazione dei pazienti.

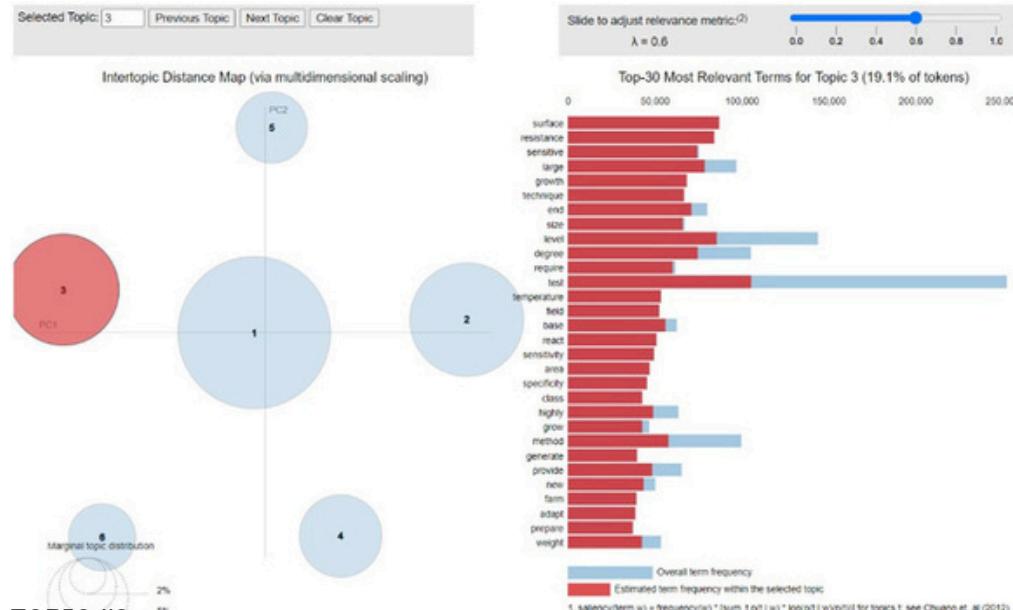
In generale, i risultati di diversi studi hanno dimostrato che la rilevazione tempestiva e il controllo della diffusione del virus sono fattori chiave per prevenire la sua diffusione e limitare l'impatto sulla salute pubblica.

Assimilabile perfettamente al risultato delle WordCloud **TOPIC#4: Tempistiche dell'infezione ed effetti sulla salute.**

4.6 QUESITO F



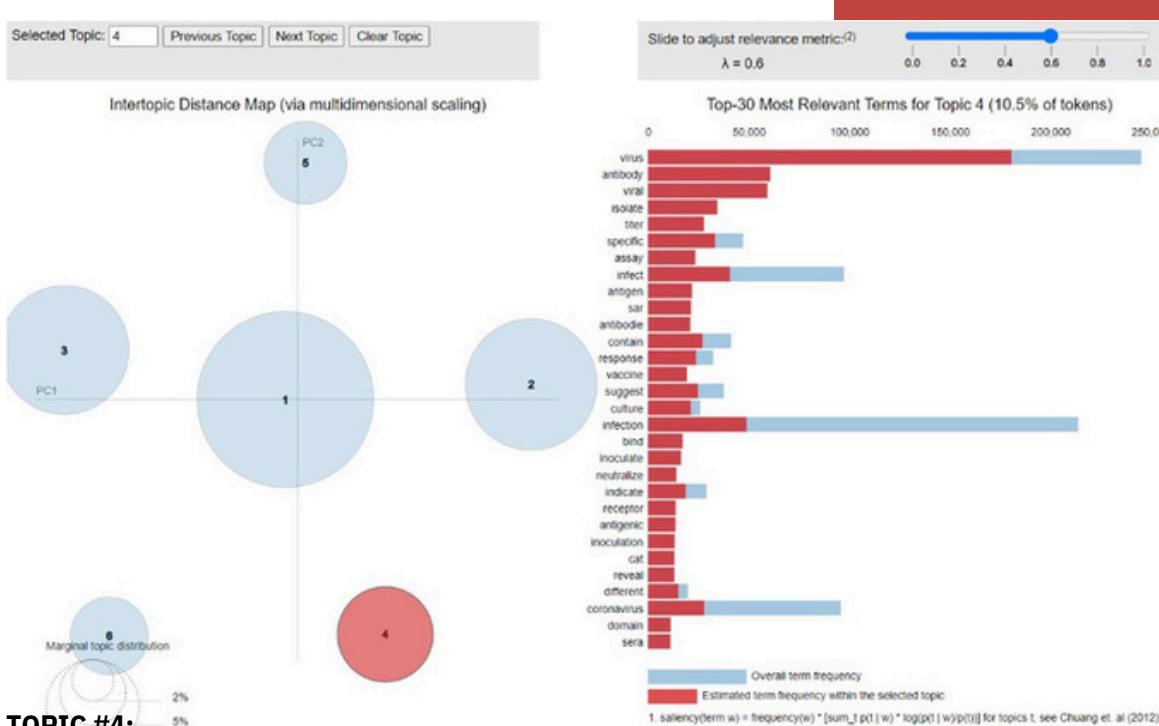
TOPIC #2: Il senso comune che emerge dai termini elencati sembra essere inerente alla gestione della pandemia da coronavirus e ai relativi effetti sulla salute pubblica e individuale. In particolare, l'esposizione al virus può avvenire attraverso una serie di fattori, come ad esempio l'esposizione diretta a persone malate o a luoghi a rischio come strutture sanitarie e ospedaliere. La gestione della pandemia richiede misure di prevenzione, come la cura personale e il monitoraggio della dieta, nonché un'attenta valutazione dei rischi per la salute associati all'esposizione al virus. Inoltre, il mantenimento di una buona salute mentale e fisica, l'adozione di misure di autogestione della cura e la formazione di una rete di sostegno familiare e sociale sono importanti per mantenere il benessere individuale e collettivo. Infine, gli operatori sanitari e infermieristici giocano un ruolo fondamentale nella gestione della pandemia, offrendo cure mediche e supporto emotivo ai pazienti. Assimilabile perfettamente al risultato delle WordCloud **TOPIC#3:Salute pubblica e Gestione epidemie**.



TOPIC #3:

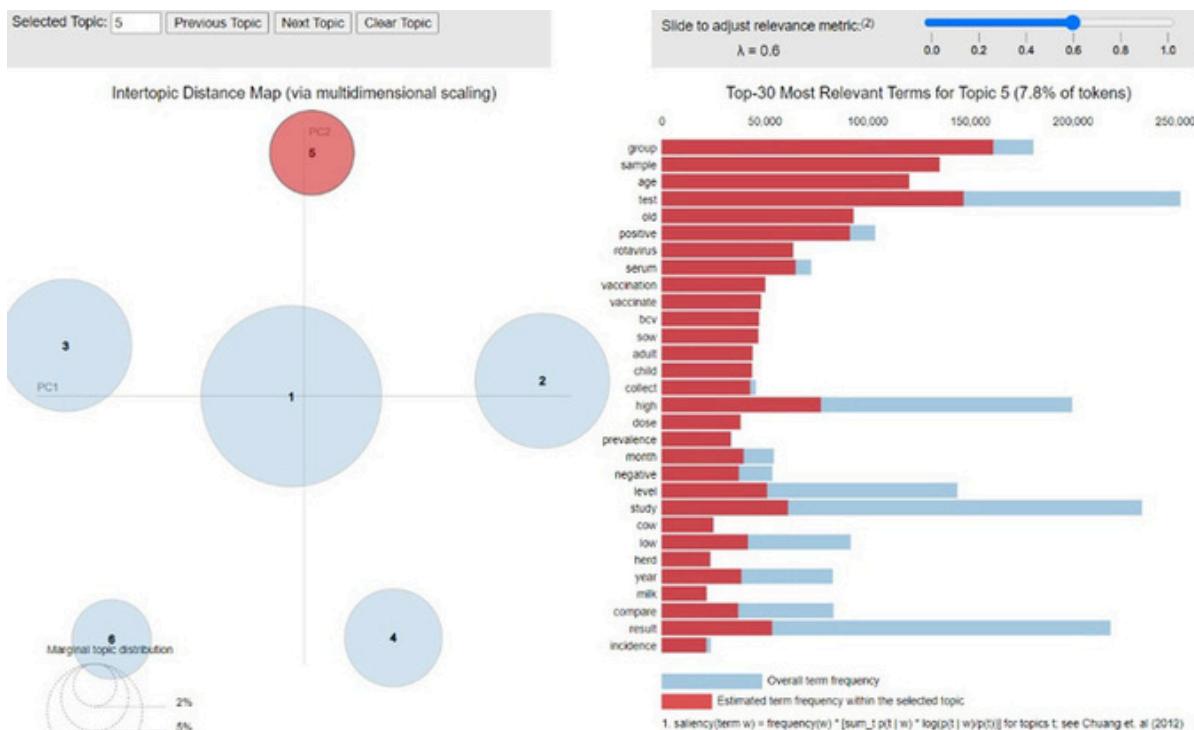
I termini elencati sembrano essere inerenti alla comprensione della crescita e della diffusione del virus SARS-CoV-2, insieme alla necessità di nuove tecniche, metodi e strumenti di test per rilevare il virus su superfici e in diversi contesti, come le fattorie e le strutture sanitarie. La sensibilità e la specificità dei test e la temperatura e l'umidità ambientale sembrano essere fattori importanti da considerare per garantire risultati accurati e prevenire la diffusione del virus.

Assimilabile perfettamente al risultato delle WordCloud **TOPIC#2:Valutazione resistenza virus e test**.



Studi riguardanti il virus SARS-CoV-2 hanno evidenziato la presenza di anticorpi specifici e la possibilità di isolare, coltivare e inoculare il virus, dimostrando inoltre l'efficacia delle vaccinazioni per prevenire l'infezione e neutralizzare la risposta antigenica.

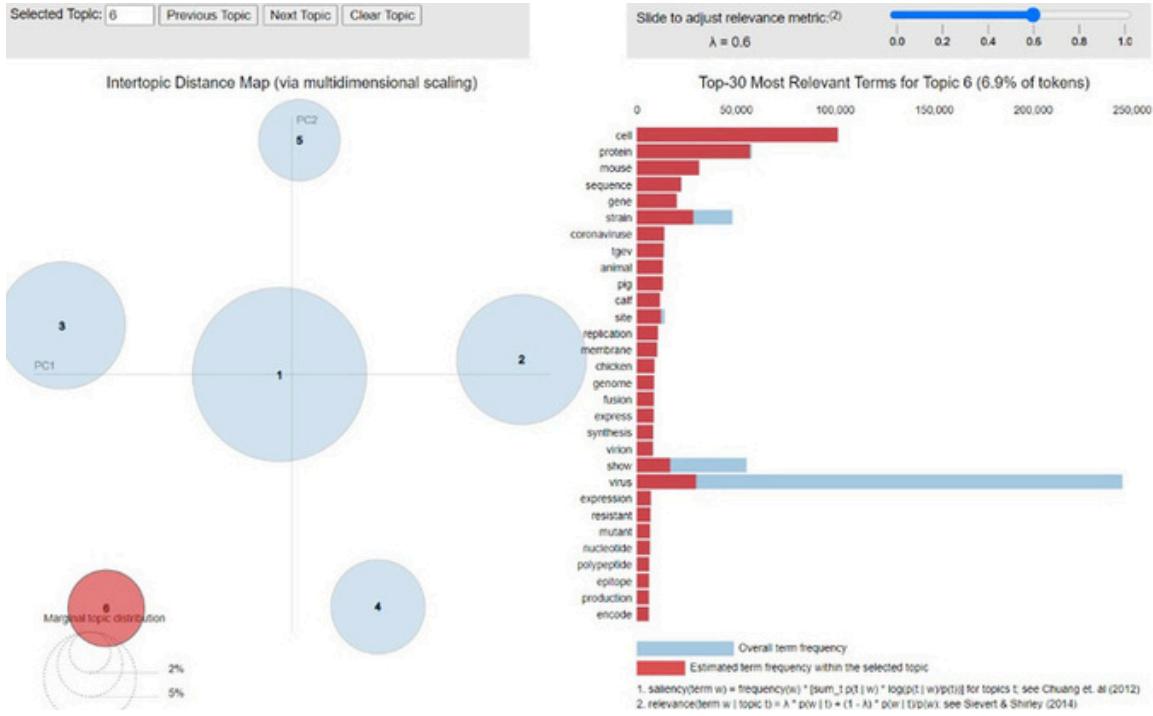
Assimilabile in parte al risultato delle WordCloud **TOPIC °1: Comprensione, controllo e risposta delle infezioni virali**, dandogli un'accezione ancora più specifica rispetto al tema Vaccini.



Il senso comune che emerge dai termini elencati sembra essere inherente allo studio dell'incidenza della vaccinazione contro il coronavirus in diversi gruppi di età, includendo adulti, anziani e bambini, con un'analisi dei livelli di anticorpi e della prevalenza del virus in ciascun gruppo.

Assimilabile in parte al risultato delle WordCloud **TOPIC #0: Studio epidemiologico del rotavirus**, in quanto centra maggiormente l'attenzione sullo studio epidemiologico rispetto alle caratteristiche del virus e dunque del vaccino, dove il rotavirus potrebbe avere un funzionamento simile al coronavirus.

4.6 QUESITO F



TOPIC #6:

Il coronavirus è un virus che infetta animali come maiali, vitelli e polli, e può causare malattie respiratorie. Il virus si replica all'interno delle cellule dell'animale ospite utilizzando le proteine della membrana cellulare. La sequenza di nucleotidi del genoma del virus codifica per le proteine necessarie per la replicazione virale e la produzione di virioni infettivi. Gli scienziati studiano anche le diverse mutazioni e ceppi del virus, come TGEV, per capire la sua resistenza

agli anticorpi e per sviluppare vaccini efficaci. Inoltre, gli scienziati identificano gli epitopi (siti di legame per gli anticorpi) presenti sul virus per sviluppare test diagnostici e terapie.

L'argomento in questione potrebbe riguardare la biologia molecolare e la virologia del coronavirus, comprese le modalità di replicazione del virus all'interno delle cellule ospiti, la codifica proteica e la produzione di virioni infettivi.

Assimilabile perfettamente al risultato delle WordCloud **TOPIC#5: Virologia e biologia molecolare**.

INTERPRETAZIONE GENERALE DEI TOPIC IN BASE AI RISULTATI

Una volta ottenuti i risultati sia delle WordCloud che dei Grafici Interattivi pyLDAvis, è stato possibile confrontarne i risultati.

Nonostante il numero associato a ciascun Topic sia risultato diverso, tale problema è dato da un aspetto meramente tecnico, infatti tramite un confronto è stato possibile individuare a livello semantico gli stessi 6 Topic in entrambi gli strumenti.

I Grafici interattivi hanno quindi confermato il risultato semantico offerto dalle WordCloud e hanno fornito informazioni più dettagliate rispetto ad alcuni Topic, permettendoci così di individuare un'interpretazione per ciascuno di essi che verrà utilizzata per distinguere i topic in seguito e che viene di seguito proposta:

TOPIC#1: STUDIO EPIDEMIOLOGICO MALATTIE INFETTIVE.

La sintesi generale è che i termini elencati riguardano l'importanza dello studio epidemiologico per valutare la diffusione di malattie infettive come il rotavirus e il coronavirus, attraverso l'analisi di campioni di diverse età e l'utilizzo di test sierologici per rilevare la presenza di anticorpi specifici e del virus stesso.

TOPIC#1: RISPOSTA ALLE INFESIONI VIRALI E VACCINI.

La sintesi generale è che i termini elencati riguardano l'importanza della comprensione, del controllo e della risposta alle infezioni virali, inclusa la produzione di anticorpi specifici e la ricerca sui virus specifici come il coronavirus. In particolare, gli studi sul virus SARS-CoV-2 hanno dimostrato l'efficacia delle vaccinazioni per prevenire l'infezione e neutralizzare la risposta antigenica.

TOPIC#2: SENSIBILITA' E SPECIFICITA' DEL VIRUS E DEI TEST:

La sintesi generale è che i termini elencati riguardano l'importanza della valutazione della resistenza del virus e dei test diagnostici con focus verso l'importanza della sensibilità e della specificità. La valutazione dell'effetto dell'ambiente sulla diffusione del virus e la necessità di nuove tecniche di rilevamento su diverse superfici e in diversi contesti sono fattori cruciali per prevenire la diffusione del virus.

TOPIC#3: GESTIONE EPIDEMIE E SALUTE PUBBLICA

La sintesi generale è che i termini elencati riguardano l'importanza della gestione delle epidemie e alla salute pubblica, con attenzione alla figura del leader, alla diffusione dell'epidemia, all'esposizione e al rischio, nonché al ruolo degli operatori sanitari. È importante mantenere la salute mentale e fisica, adottare misure di prevenzione e formare una rete di sostegno familiare e sociale per garantire il benessere individuale e collettivo.

TOPIC#4: TEMPISTICHE:DIAGNOSI,SINTOMI,DIFFUSIONE E TRATTAMENTO

La sintesi generale è che i termini elencati riguardano l'importanza della gestione delle malattie infettive, con particolare attenzione alla diagnosi, ai sintomi, alla diffusione, alla risposta immunitaria e al trattamento. Inoltre, si sottolinea l'importanza della prevenzione e della rilevazione tempestiva per limitare l'impatto sulla salute pubblica.

TOPIC#5: VIROLOGIA E BIOLOGIA MOLECOLARE

La sintesi generale è che i termini elencati riguardano l'importanza della biologia molecolare e la virologia, in particolare la ricerca su proteine, sequenze genetiche, virus e cellule ospiti utilizzati per la ricerca.

Entrambi gli argomenti si concentrano sull'analisi del coronavirus e dei virus simili, inclusa la replicazione virale all'interno delle cellule ospiti, la produzione di virioni infettivi, la resistenza agli anticorpi e lo sviluppo di vaccini efficaci e terapie.

ANALISI TEMPORALE DEI TOPICS

Al fine di poter rispondere ai sotto-quesiti F:

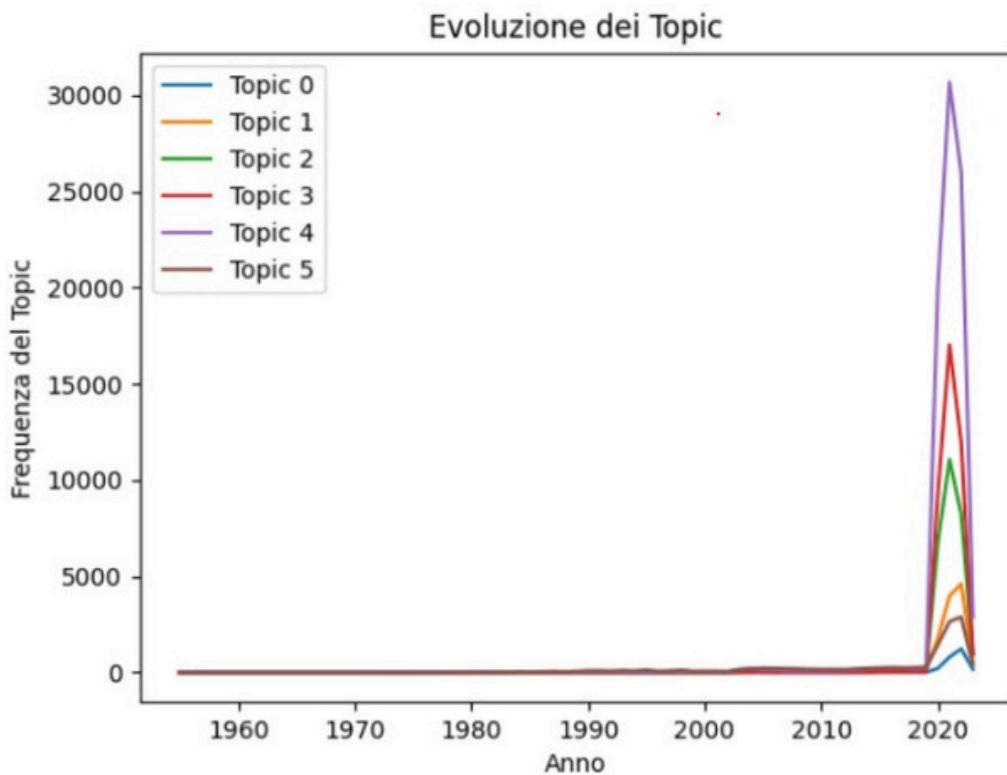
-Qual'è stato l'andamento dei Topic individuati durante gli anni analizzati?

-Come l'avvento della Pandemia ha modificato l'andamento dei Topic trattati, tra quelli individuati?

Si è optato per l'elaborazione di un'Analisi temporale dei Topics al fine di mettere in evidenza eventuali trend nelle argomentazioni trattate negli articoli raccolti nel periodo compreso tra il 1949 e il 2023.

Al fine di visualizzare graficamente i trend dei 6 Topics trattati negli anni, è stato generato un Grafico che misura le pubblicazioni per ciascun topic per ogni anno e attraverso l'utilizzo di una retta spezzata mostra dunque l'interesse mostrato per ciascun Topic durante gli anni presi in considerazione.

Il periodo preso in considerazione è quello che va dal 1949 al 2023.



Grazie all'analisi storica e allo strumento di visualizzazione è possibile rilevare i seguenti aspetti:

- Così come nella distribuzione delle Pubblicazioni viste nell'Analisi Esplorativa, notiamo che anche l'andamento dei Topic subisce un primo lieve rialzo dopo gli anni 2000 e un picco esponenziale a partire dalla fine del 2019 in concomitanza con lo scoppio della Pandemia Sars-Cov19 che trova il suo apice nel 2021 e scende nel 2022 e a seguire.

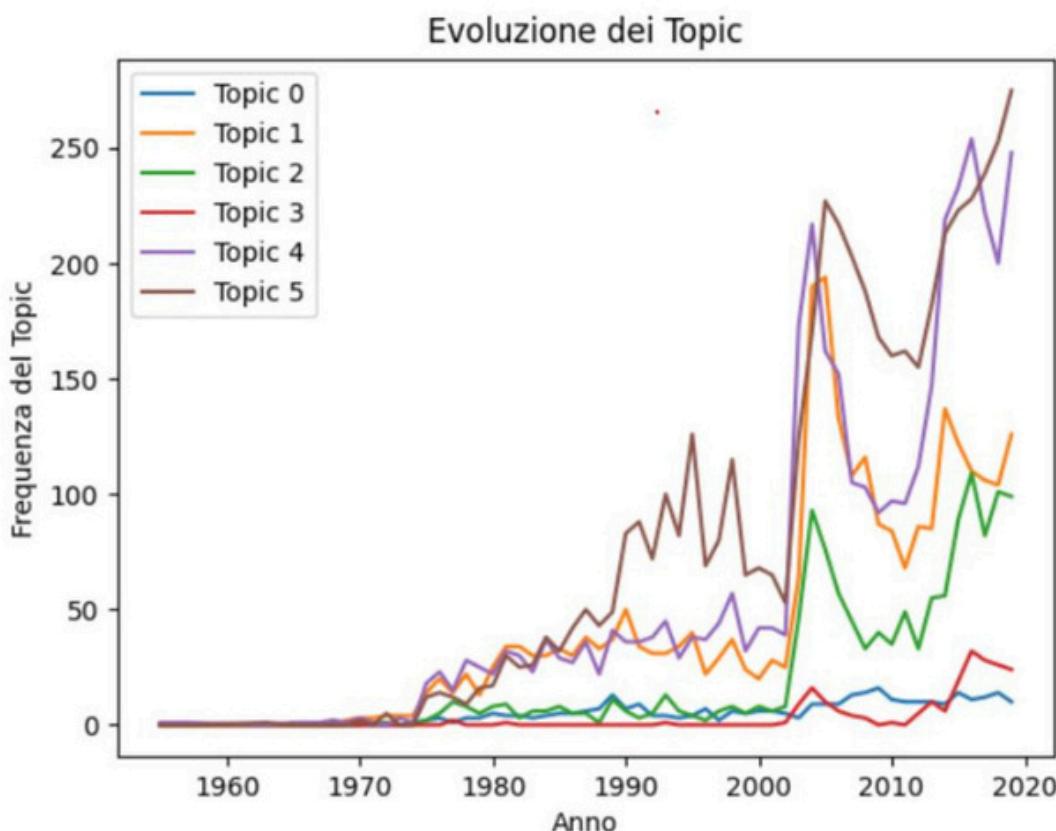
- Analizzando il picco delle pubblicazioni a partire dal 2019 possiamo vedere come, nonostante tutti i Topic abbiano un naturale rialzo, I topic #5, #0, #1 riguardanti rispettivamente la Virologia e Biologia Molecolare, lo Studio Epidemiologico e la Risposta alle infezioni e vaccini aumentino meno rispetto ai restanti topic.

Tra i topic restanti il #2, la Sensibilità e Specificità dei Test e del Virus raggiunge livelli di frequenza doppi rispetto ai primi tre, mentre ed il #3, la Gestione delle Epidemie e della Salute Pubblica risulta averne il triplo.

Infine il topic #4, relativo alle Tempistiche rispetto alle diagnosi,sintomi,diffusione e trattamento risulta essere il topic più trattato, quasi il doppio rispetto al Topic #3.

Data la distribuzione particolarmente sbilanciata tra gli anni precedenti alla pandemia e quelli successivi, si è optato per la creazione di ulteriori Grafici che misurassero l'andamento dei Topic ma con focus su periodi distinti al fine di visualizzare ulteriori comportamenti dei topic negli anni, visibili su scala differente.

FOCUS ANDAMENTO STORICO 1949-2019:



Grazie al focus al periodo precedente la pandemia è possibile valutare meglio gli andamenti dei Topic individuati durante i 70 anni che vanno dal 1949 al 2019.

E' possibile notare come la ricerca non fosse molto attiva in ambito "Coronavirus" per nessuno dei topic individuati fino agli anni 70, a metà degli stessi anni 70 abbiamo un primo leggero picco, che vede i Topic #1,#4,#5 riguardanti rispettivamente la Risposta alle infezioni e vaccini, le Tempistiche rispetto alle diagnosi,sintomi,diffusione e trattamento e la Virologia e Biologia Molecolare iniziare a distaccarsi rispetto agli altri Topic.

Mentre la Biologia e Virologia mantengono alti valori rispetto agli altri tra il 1990 e il 2000 un secondo picco si registra nei primi anni degli anni 2000 che coinvolge i Topic #1,#4,#5 ma anche il topic #2, la Sensibilità e Specificità dei Test e del Virus , anche se resta a livelli dimezzati rispetto ai topic precedenti, mentre i topic #0,lo Studio Epidemiologico e #3, la Gestione delle Epidemie e della Salute Pubblica mantengono livelli bassi.

Dopo un leggero calo,l'andamento dei 6 Topic tende a salire progressivamente negli anni fino al 2019 e vede in testa i Topic #5 e #4, a seguire con livelli dimezzati rispetto ai precedenti i topic #1 e #2 e per ultimi con livelli molto più bassi i topic #3 e #0.

Confrontando tali andamenti con un'analisi degli eventi storici effettivamente si può notare che negli anni '70, la ricerca sui coronavirus era ancora agli inizi e non vi erano ancora state epidemie di coronavirus umani di grande entità. Tuttavia, in quegli anni si verificarono alcuni eventi significativi che potrebbero aver influenzato la ricerca sui virus a RNA in generale, compresi i coronavirus, ovvero:

- Uno di questi eventi fu la scoperta dei retrovirus, un gruppo di virus a RNA che utilizzano un'enzima chiamata trascrittasi inversa per trasformare il loro RNA in DNA, che poi viene integrato nel genoma dell'ospite. La scoperta dei retrovirus e della loro capacità di integrarsi nel DNA dell'ospite ebbe un grande impatto sulla ricerca in ambito virale, portando alla scoperta di nuove classi di virus e alla comprensione di importanti meccanismi biologici, come la replicazione del DNA.

4.5 QUESITO F

-Un altro evento significativo che potrebbe aver influenzato la ricerca sui coronavirus negli anni '70 fu l'epidemia di influenza suina del 1976 negli Stati Uniti, causata dal virus dell'influenza suina A (H1N1). Questa epidemia portò all'introduzione di nuovi metodi di sorveglianza epidemiologica e di ricerca sui virus influenzali, che potrebbero aver beneficiato anche la ricerca sui coronavirus e altri virus a RNA.

-Inoltre, durante gli anni '70 furono sviluppati nuovi metodi per la coltivazione di virus in laboratorio e la tecnologia dell'ingegneria genetica cominciava a svilupparsi, aprendo nuove possibilità di ricerca in ambito virologico.

In generale, negli anni '70 ci furono diversi progressi scientifici e tecnologici che potrebbero aver favorito la ricerca sui virus a RNA, compresi i coronavirus, e questo può evidenziare dal primo picco a metà degli anni settanta e l'andamento del Topic riguardo la Biologia e Virologia.

Con riguardo invece al picco degli anni 2000, si può notare come prima dell'emergenza del coronavirus

Sono stati numerosi studi accademici sulla famiglia dei coronavirus, che comprende anche virus della SARS (SARS-CoV) e il virus della MERS (MERS-CoV), che hanno causato epidemie in passato.

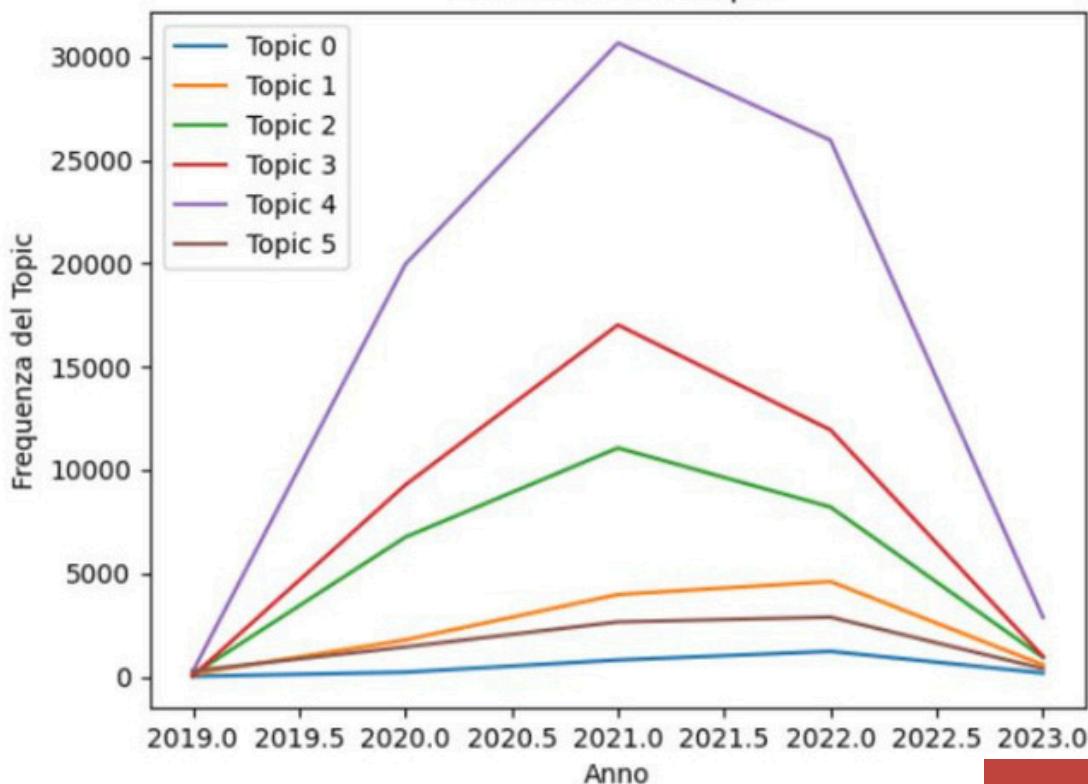
Negli anni 2000, in particolare, la ricerca sui coronavirus ha ricevuto un certo grado di attenzione accademica. Nel 2003, la SARS-CoV è stata la causa di un'epidemia che ha colpito principalmente l'Asia e il Canada, causando oltre 8000 casi e circa 800 decessi. L'epidemia di SARS ha portato ad un maggiore interesse accademico sulla famiglia dei coronavirus, in particolare sulla biologia e la patogenesi di questi virus.

In seguito, ci sono stati numerosi studi che hanno cercato di comprendere la biologia e la patogenesi dei coronavirus, non solo della SARS-CoV ma anche di altri virus della stessa famiglia. Questi studi hanno incluso la ricerca sulla struttura e la funzione delle proteine virali, sulle modalità di trasmissione e sulle strategie per prevenire e trattare le infezioni da coronavirus.

In sintesi, sebbene la ricerca sui coronavirus non sia stata così ampia e intensa come lo è stata durante la pandemia di COVID-19, la famiglia dei coronavirus ha attirato l'attenzione degli scienziati già dagli anni 2000, soprattutto dopo l'epidemia di SARS.

FOCUS ANDAMENTO STORICO 2019-2023:

Evoluzione dei Topic



Grazie al focus sul periodo durante la pandemia e a seguire è possibile valutare meglio gli andamenti dei Topic individuati durante gli anni che vanno dal 2019 al 2023.

E' possibile notare come la ricerca subisca un esplosione proprio in concomitanza dello scoppio della pandemia, subendo una crescita esponenziale rispetto al periodo precedente.

I picchi per tutti i Topic vengono raggiunti tra il 2021 e il 2022, segnando un calo nel periodo 2022-2023, a prescindere dal livello del 2023 che è influenzato dalla non completezza delle pubblicazioni riferite a tale anno, confermando come l'avvento della pandemia abbia aumentato l'interesse di tale tematica generale nella ricerca accademica.

Effettuando un confronto con il periodo precedente, possiamo notare come il Topic #4, le Tempistiche rispetto alle diagnosi, sintomi, diffusione e trattamento, che già possedeva alti livelli rispetto agli altri nel 2019, mantenga questo primato per tutti gli anni analizzati segnando livelli molto maggiori rispetto agli altri Topic.

Con riguardo al topic #3, la Gestione delle Epidemie e della Salute Pubblica, ne osserviamo un aumento esponenziale anche rispetto al suo posizionamento precedente che lo porta da penultimo a secondo topic più trattato.

A seguire troviamo il Topic #2, la Sensibilità e Specificità dei Test e del Virus, che guadagna una posizione e subisce una crescita esponenziale rispetto agli ultimi tre topic rimanenti.

Interessante l'andamento dei Topic #1 e #5, ovvero la Risposta alle infezioni e vaccini e la Virologia e Biologia Molecolare che non subiscono una crescita esponenziale come i topic #4, #3, #2 e soprattutto passano da posizioni "alte" del 2019 a posizioni "basse" rispetto agli altri Topic, segnando un calo dell'interesse della ricerca scientifica in tali campi.

Specialmente il topic #5 che passa da Topic di maggior interesse a penultimo.

Il topic #0, lo Studio Epidemiologico, che anche nei periodi precedenti non raggiungeva alti livelli, mantiene costante il suo andamento rispetto agli altri.

Ci sono diverse ragioni per cui alcuni argomenti possono subire una crescita più rapida rispetto ad altri. In primo luogo, l'interesse per un argomento può essere influenzato dagli eventi attuali o dalle tendenze sociali. Ad esempio, durante la pandemia di COVID-19, la ricerca sulla gestione delle epidemie è stata naturalmente al centro dell'attenzione ed è stata guidata dall'urgenza e dalla necessità di affrontare una crisi sanitaria globale. Inoltre, la pandemia ha portato all'attenzione pubblica una serie di questioni, come la sensibilità e specificità dei test e del virus, che forse non erano state considerate in modo così ampio in precedenza.

Al contrario, altri argomenti come la risposta alle infezioni e i vaccini potrebbero essere diventati meno urgenti a causa dell'aumento delle vaccinazioni e della diminuzione dei casi di COVID-19.

In secondo luogo, alcuni argomenti potrebbero essere più "caldi" di altri in termini di ricerca scientifica e sviluppo tecnologico. Infine, la disponibilità di finanziamenti e risorse può influire sull'interesse per un argomento. Se un campo ha finanziamenti e risorse limitati, potrebbe subire una crescita più lenta rispetto a un campo con maggiori risorse a disposizione.

TOPIC COLLABORATION NETWORK

Al fine di poter rispondere al sotto-quesito E:

Si può individuare l'andamento dei Topic tramite il Grafo? E dunque possibile evidenziare il passaggio di un topic da interessante a meno interessante e viceversa tramite lo strumento del grafo?

Si è voluto optare per la creazione di un algoritmo che fondesse le analisi svolte sui Grafi e i risultati ottenuti dalla Topic Modeling.

A tal riguardo, essendo il grafo riferito alle entità Autori e la Topic Modeling riferita invece alle entità Documenti, è stato necessario implementare un algoritmo che risolvesse tale problematica, associando a ciascun co-autore di un'opera il topic della stessa e andando dunque ad aggiungere al Dataset originario le colonne:

- **Dominant_Topic:** contenente il topic dominante tra i 6 individuati dalla Topic Modeling per ciascun autore, espresso con numeri dallo 0 al 5.

- **Topic_Perc_Contrib:** contiene la percentuale di contribuzione del topic dominante per ogni specifico autore.

- **Keywords:** contenente le parole chiave più rappresentative per ciascun topic per ogni specifico autore.

Inoltre, essendo possibile che un singolo Autore abbia lavorato a distinti Documenti e quindi a distinti Topic nel periodo di tempo preso in considerazione, si è reso necessario implementare del codice che tenesse traccia delle occorrenze di ciascun topic trattato da ogni singolo Autore e che associasse a tale Autore il topic trattato dallo stesso più volte e dunque con maggiori occorrenze.

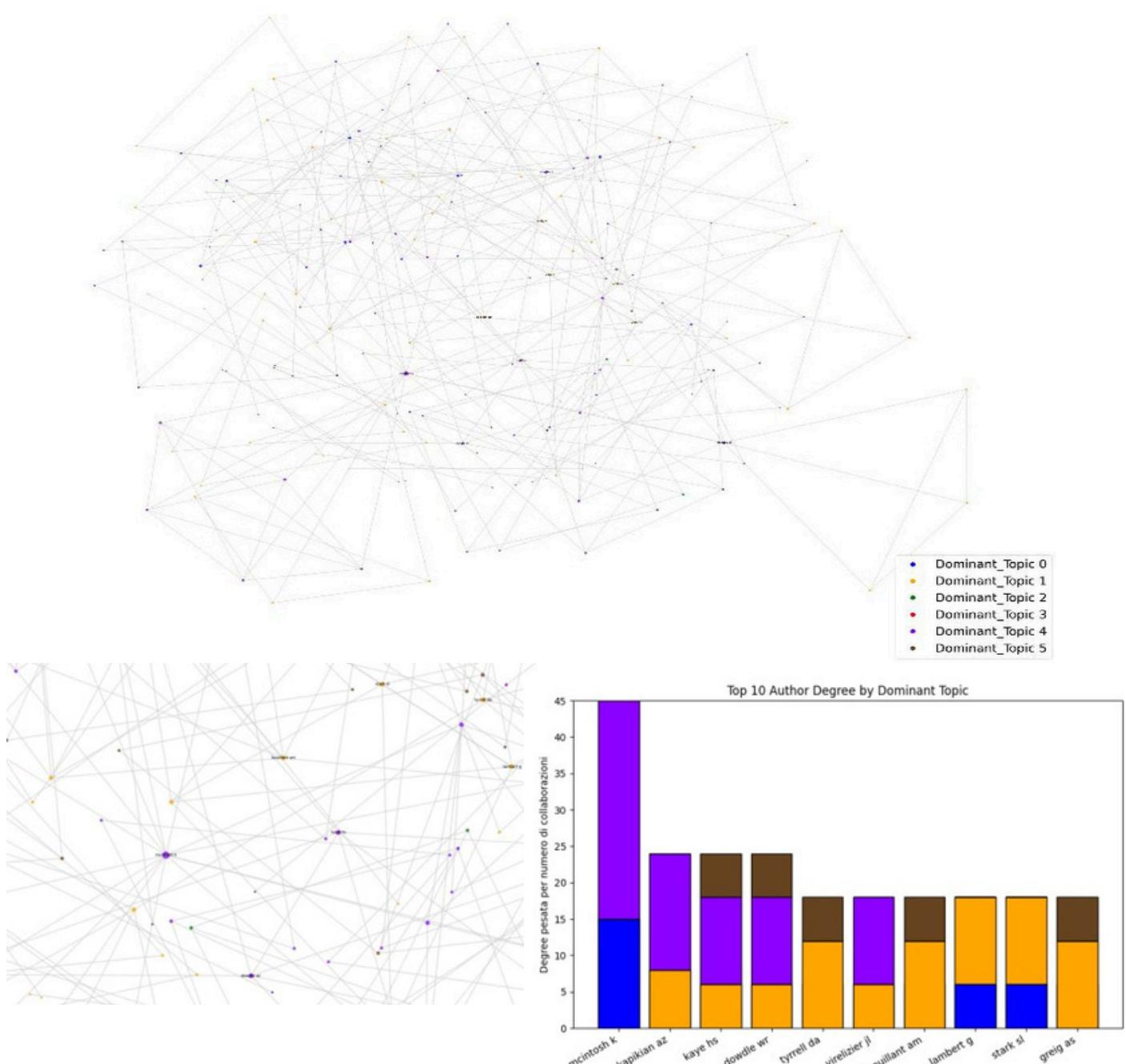
Per poter visualizzare in maniera chiara tutto questo, si è implementato nel codice precedentemente utilizzato per la creazione del Grafo della Rete di Collaborazione tra Autori, un ulteriore algoritmo generasse un grafo dove i nodi rappresentati i singoli autori sono colorati tramite un codice colore associato a ciascun topic dominante per ogni singolo autore.

Inoltre, per fornire un grado di informazione ulteriore, tale grafo è stato accompagnato dalla creazione di un ulteriore Grafico a Barre Sovraposte, che permette di visualizzare il contributo di ciascun Topic trattato da ogni singolo autore, in modo da non visualizzare esclusivamente quello dominante.

A tal riguardo vengono di seguito proposti il Grafo delle Collaborazioni tra Autori evidenziati secondo Topic di riferimento e il Grafico dei Top 10 Autori secondo la loro Degree ed evidenziati con i vari topic trattati dagli stessi.

Al fine di avere una visualizzazione più nitida dei nodi del grafo, sono stati generati più grafi e grafici a seconda dei vari periodi di tempo presi in analisi, filtrando di volta in volta il grafo secondo determinati livelli di degree e di numero di pubblicazioni all'aumentare del numero di pubblicazioni nel corso del tempo.

Grafo del Coautorato evidenziato per Topic nel 1949_1975 (Degree=1; Pubblicazioni >= 1):



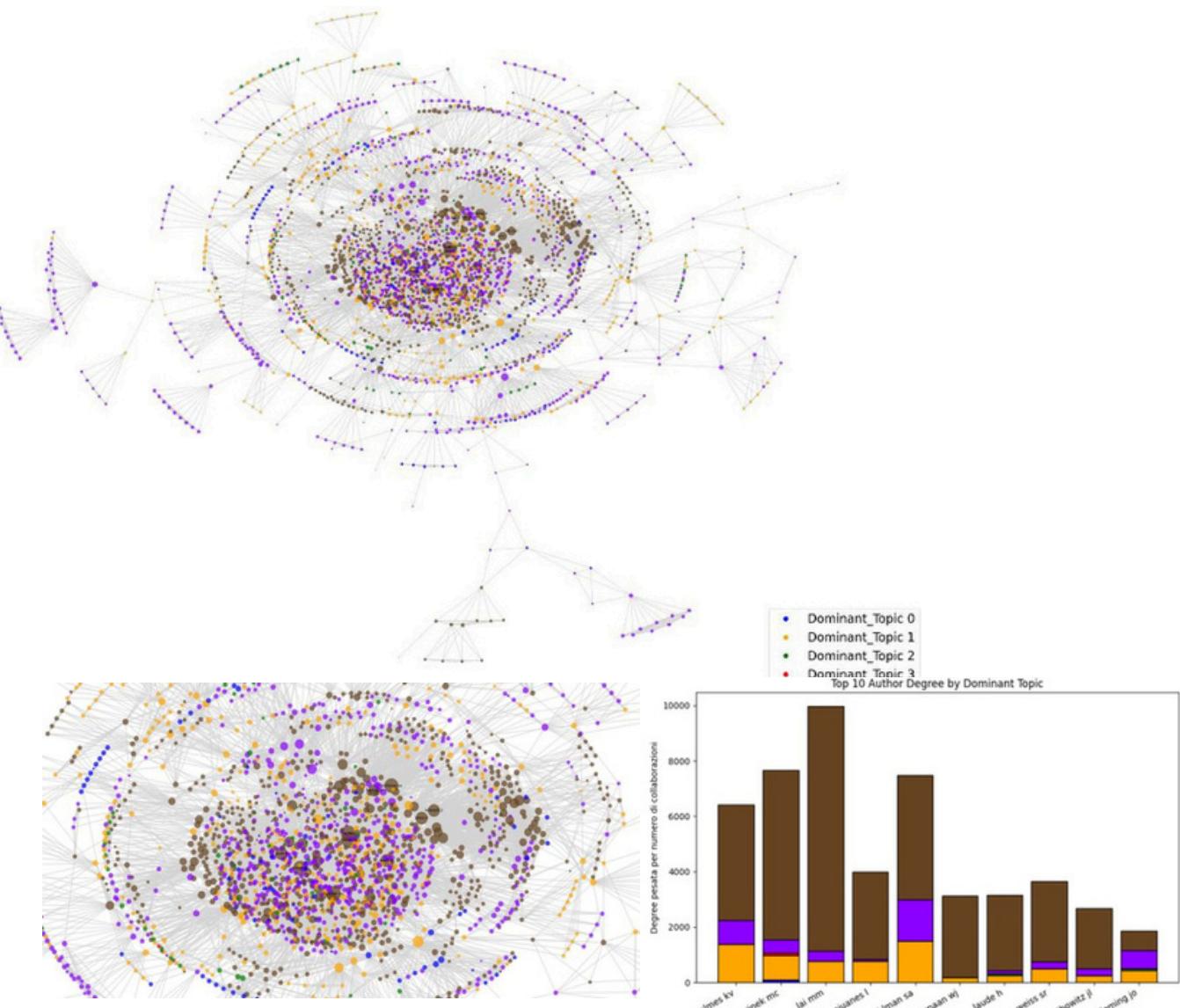
Grazie all'informazione contenuta nel Grafo e nel Grafico della Top10, riusciamo a visualizzare le collaborazioni tra autori che hanno scritto almeno un'opera e che hanno almeno una collaborazione durante il periodo dal 1949 al 1975.

Inoltre è possibile associare a ciascun autore i topic da lui trattati ed associare il colore del topic dominante al nodo, in tal modo è possibile trarre le seguenti conclusioni:

Gli argomenti più trattati risultano essere il #4 e il #1, ovvero quelli riferiti rispettivamente alle **Tempistiche rispetto alle diagnosi, sintomi, diffusione e trattamento** ed alla **Risposta alle infezioni e vaccini**.

A seguire il topic #0, lo **Studio Epidemiologico** e il #5, la **Virologia e Biologia Molecolare** e poi gli altri.

Nonostante la ridotta numerosità delle pubblicazioni, tali risultati ci confermano quanto osservato nella fase di analisi dell'andamento storico dei Topic anche in termini di collaborazioni, infatti i topic maggiormente trattati risultano anche quelli con maggiori collaborazioni come mostrato dal grafo e dai nodi con degree maggiore.

Grafo del Coautorato evidenziato per Topic nel 1976_1999 (Degree=5; Pubblicazioni >= 1):


Grazie all'informazione contenuta nel Grafo e nel Grafico della Top10, riusciamo a visualizzare le collaborazioni tra autori che hanno scritto almeno un'opera e che hanno almeno cinque collaborazioni durante il periodo dal 1976 al 1999. Inoltre è possibile associare a ciascun autore i topic da lui trattati ed associare il colore del topic dominante al nodo, in tal modo è possibile trarre le seguenti conclusioni:

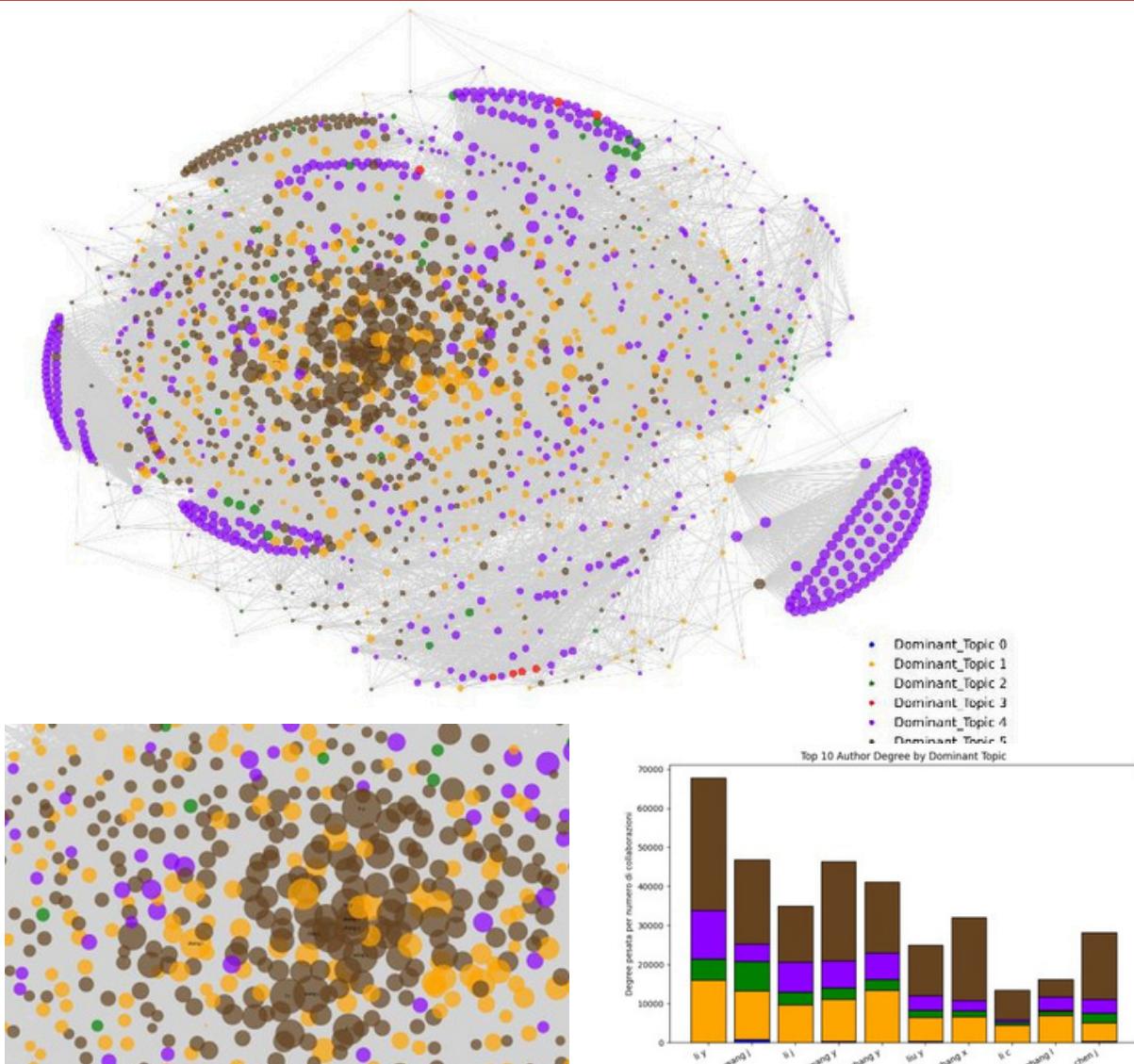
L'argomento più centrale risulta essere chiaramente il #5, la **Virologia e Biologia Molecolare**, seguito da una nube diffusa di #1 e #4, ovvero quelli riferiti rispettivamente alla **Risposta alle infezioni e vaccini** e alle **Tempistiche rispetto alle diagnosi, sintomi, diffusione e trattamento**.

A seguire il topic #0 e #2, lo **Studio Epidemiologico e la Sensibilità e Specificità dei Test e del Virus** e infine il #3, la **Gestione delle Epidemie e della Salute Pubblica**.

Tali risultati ci confermano quanto osservato nella fase di analisi dell'andamento storico dei Topic anche in termini di collaborazioni, infatti i topic maggiormente trattati risultano anche quelli con maggiori collaborazioni come mostrato dal grafo e dai nodi con degree maggiore.

E' possibile infatti rilevare l'interessamento al topic #5 rispetto agli altri argomenti in questo ventennio.

Grafo del Coautorato evidenziato per Topic nel 2000_2018 (Degree=50; Pubblicazioni >= 1):



Grazie all'informazione contenuta nel Grafo e nel Grafico della Top10, riusciamo a visualizzare le collaborazioni tra autori che hanno scritto almeno un'opera e che hanno almeno cinquanta collaborazioni durante il periodo dal 2000 al 2018.

Inoltre è possibile associare a ciascun autore i topic da lui trattati ed associare il colore del topic dominante al nodo, in tal modo è possibile trarre le seguenti conclusioni:

L'argomento più centrale risulta essere ancora chiaramente il #5, la **Virologia e Biologia Molecolare**, seguito dal topic di #1, **Risposta alle infezioni e vaccini**.

Interessante la visualizzazione delle collaborazioni per quanto riguarda il topic #4, **Tempistiche rispetto alle diagnosi,sintomi,diffusione e trattamento**, che risulta formare delle comunità distinte in tale ambito.

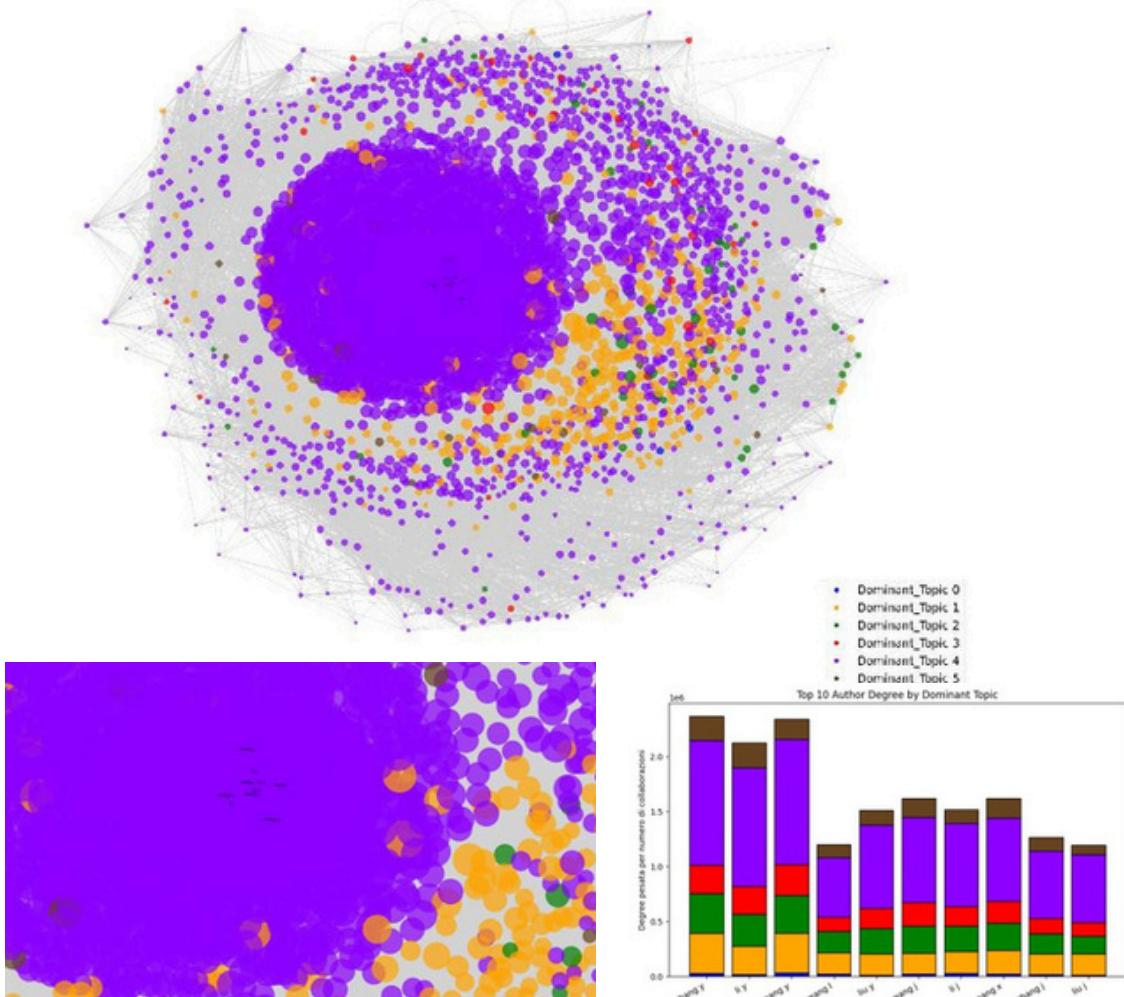
A seguire i topic #2 e #3, la **Sensibilità e Specificità dei Test e del Virus** e la **Gestione delle Epidemie e della Salute Pubblica**. Infine il topic #0, lo **Studio Epidemiologico**.

Tali risultati ci confermano quanto osservato nella fase di analisi dell'andamento storico dei Topic anche in termini di collaborazioni, infatti i topic maggiormente trattati risultano anche quelli con maggiori collaborazioni come mostrato dal grafo e dai nodi con degree maggiore.

E' possibile infatti rilevare il continuo interessamento ai topic #5,#4 e #1 e il crescente interesse verso il topic #2,la **Sensibilità e Specificità dei Test e del Virus**.

Rimane di meno interesse il topic #3, **Gestione delle Epidemie e della Salute Pubblica** e perde l'interesse nel tempo il topic #0, lo **Studio Epidemiologico**.

Grafo del Coautorato evidenziato per Topic nel 2019_2023 (Degree=250; Pubblicazioni >= 2):



Grazie all'informazione contenuta nel Grafo e nel Grafico della Top10, riusciamo a visualizzare le collaborazioni tra autori che hanno scritto almeno due opere e che hanno almeno duecentocinquanta collaborazioni durante il periodo dal 2019 al 2023.

Inoltre è possibile associare a ciascun autore i topic da lui trattati ed associare il colore del topic dominante al nodo, in tal modo è possibile trarre le seguenti conclusioni:

L'argomento più centrale durante il periodo pandemico risulta essere il #4, **Tempistiche rispetto alle diagnosi,sintomi,diffusione e trattamento**.

A seguire a livelli inferiori troviamo i topic #3,#2 e #1,rispettivamente **Gestione delle Epidemie e della Salute Pubblica , la Sensibilità e Specificità dei Test e del Virus e la Risposta alle infezioni e vaccini**.

Infine troviamo il topic #5, **la Virologia e Biologia Molecolare** e per ultimo il topic #0,**lo Studio Epidemiologico**.

Tali risultati ci confermano quanto osservato nella fase di analisi dell'andamento storico dei Topic, infatti i topic maggiormente trattati nel periodo analizzato risultano essere gli stessi mostrati dal grafo e dai nodi con degree maggiore.

E' possibile infatti rilevare il l'ancora più interesse per il topic #4, un incremento importante dell'interesse verso il topic #3, **Gestione delle Epidemie e della Salute Pubblica** ed il continuo alto interesse per i topic #2,**la Sensibilità e Specificità dei Test e del Virus** e #1, **Risposta alle infezioni e vaccini**.

Il topic #0,**lo Studio Epidemiologico** mantiene il suo andamento, mentre anche tramite il Grafo è possibile osservare come il topic #5, **la Virologia e Biologia Molecolare** sia incredibilmente in disuso rispetto ai livelli degli anni precedenti.

RISULTATI TOPIC COLLABORATION NETWORK

In conclusione possiamo affermare che è possibile rispondere positivamente alla domanda "Si può individuare l'andamento dei Topic tramite il Grafo? E dunque possibile evidenziare il passaggio di un topic da interessante a meno interessante e viceversa tramite lo strumento del grafo?"

Come mostrato dalla nostra analisi e dai grafici generati infatti, siamo riusciti tramite lo strumento del grafo ad individuare l'andamento dei Topic negli anni ed evidenziare il maggior o minor interesse per le tematiche individuate nel corso degli anni attraverso la visualizzazione delle reti di collaborazione.

Tali risultati sono perfettamente in linea con le analisi fornite in precedenza e descrivono la seguente situazione: - Nonostante le poche osservazioni, i due argomenti con maggiori collaborazioni durante il periodo dal 1949

al 1975 sono stati la "Tempistica rispetto alle diagnosi, sintomi, diffusione e trattamento" e la "Risposta alle infezioni e vaccini" ed a seguire "Lo studio Epidemiologico".

-Dal 1976 al 1999, sono state evidenziate maggiori collaborazioni per la "Virologia e Biologia Molecolare", con il topic #5 che risulta essere l'argomento centrale.

Tuttavia, anche in questo periodo vi sono numerose collaborazioni nella "Risposta alle infezioni e vaccini" e la "Tempistica rispetto alle diagnosi, sintomi, diffusione e trattamento" e a seguire "Studio epidemiologico", mentre minori collaborazioni riguardano la "Sensibilità e Specificità dei Test e del Virus", "Gestione delle Epidemie e della Salute Pubblica".

-Nel periodo dal 2000 al 2018, con il diffondersi di nuovi ceppi di coronavirus, il topic "Virologia e Biologia Molecolare" continua ad essere l'argomento centrale con il maggior numero di collaborazioni, seguito dalla "Risposta alle infezioni e vaccini".

È interessante notare come la "Tempistica rispetto alle diagnosi, sintomi, diffusione e trattamento" abbia formato delle comunità distinte all'interno del grafo del coautore, indicando un aumento di attenzione per questo argomento.

Da notare in tale periodo il maggior interesse verso la "Sensibilità e Specificità dei Test e del Virus", mentre la "Gestione delle Epidemie e della Salute Pubblica" insieme allo "Studio Epidemiologico" sono trattati meno.

Infine, dal 2019 al 2023, con la pandemia di COVID-19, la "Tempistica rispetto alle diagnosi, sintomi, diffusione e trattamento" è diventata l'argomento principale della ricerca scientifica sull'ambito "Coronavirus", seguita dalla "Gestione delle Epidemie e della Salute Pubblica, dalla "Sensibilità e Specificità dei Test e del Virus" e dalla "Risposta alle infezioni e vaccini".

Perde molto interesse "Virologia e Biologia Molecolare" che si attesta a bassi livelli insieme allo "Studio Epidemiologico".

In sintesi, le maggiori collaborazioni nella ricerca scientifica sull'ambito "Coronavirus" dal 1949 al 2023 sono state in tema "Risposta alle infezioni e vaccini", "Tempistica rispetto alle diagnosi, sintomi, diffusione e trattamento" e la "Virologia e Biologia Molecolare". Tuttavia, l'avvento della pandemia ha spinto la ricerca ad aumentare l'interesse, e di conseguenza le collaborazioni ,per tematiche quali la "Gestione delle Epidemie e della Salute Pubblica" e la "Sensibilità e Specificità dei Test e del Virus" rispetto ad argomenti quali lo "Studio Epidemiologico" e la "Virologia e Biologia Molecolare".

5 CONCLUSIONI

In conclusione possiamo affermare che, grazie all'applicazione delle metodologie proprie della Scienza dei Network, è stato possibile analizzare le Reti di Collaborazione tra Autori nella Ricerca Scientifica in ambito “Coronavirus” e si è riusciti nell'obiettivo di descrivere alcuni aspetti di tale fenomeno ed in particolare come esso si sia modificato nel tempo anche a seguito della pandemia Sars-Covid19.

In dettaglio, attraverso le analisi effettuate è stato dunque possibile fornire risposta ai quesiti proposti inizialmente:

- E' stato evidenziato come è strutturato il Grafo delle Collaborazioni.
- E' stato evidenziato come il Grafo delle Collaborazioni sia divisibile in più componenti che indicano dei cluster di collaborazioni più intense tra autori.
- E' stato evidenziato come l'avvento della Pandemia Sars-Covid19 abbia influito sui cluster di collaborazione e le strutture sopracitati.
- È stato evidenziato come, nel periodo pre-pandemia, la letteratura su tema coronavirus rispecchi la classifica delle riviste di settore più autorevoli e come questo invece non avvenga durante e post periodo pandemico.
- E' stato evidenziato come effettivamente i First Author abbiano un ruolo di "ponte" tra comunità di collaborazione.
- E' stato evidenziato come sia possibile individuare gli argomenti maggiormente trattati nella Ricerca Scientifica in ambito “Coronavirus” e la loro evoluzione nel tempo anche tramite l'analisi dei grafi combinata alla Topic Modeling.

I limiti dell'analisi proposta sono molteplici e possono fungere da spunto utile a future ricerche.

In particolare, visto l'elevatissimo numero di osservazioni che ha reso difficile l'analisi effettuata con gli strumenti a disposizione, si è ricorso in alcuni casi a un filtraggio per rendere i risultati più intellegibili.

Una futura ricerca, avendo strumenti adatti, potrebbe analizzare in maniera complessiva il dataset fornendo risultati più completi o alternativamente sarebbe interessante effettuare dei focus per specifiche caratteristiche di autori o opere.

A seguire, va specificato che le nostre analisi si sono limitate all'utilizzo delle metodologie della scienza dei network utili a rispondere ai quesiti proposti e non vanno quindi da intendersi come esaustive.

Un futuro lavoro potrebbe dunque utilizzare differenti metriche e metodologie al fine di rispondere a ulteriori quesiti in merito all'analisi delle Reti di Collaborazione tra Autori nella Ricerca Scientifica in ambito “Coronavirus”.

- Documentazione ufficiale [NetworkX](#).
- Documentazione ufficiale [Gensim](#).
- Documentazione ufficiale [Matplotlib](#).
- Documentazione ufficiale [Pandas](#).
- Sito di PubMed.
- Sito MDPI
- Compendio insegnamento di Economia dei Network - A.A 2021-2022 - Prof. Stefano Matta - Università degli Studi di Cagliari
- Sito SCImago Journal Rank o SJR indicator
- D. Easley – J. Kleinberg, Networks Crowds and Markets, Reasoning about a Highly Connected World, Cambridge University Press.