Assignment n.3

Social Network Analysis

This assignment is an individual assignment. Each students is assigned a different dataset which required to analyse by computing different measures and properties using Gephi.

Each student must produce a Report, descriptions, measures and graphs containing the following informations 1) to 8).

Reference instructions for the requested task can be found at
http://www.martingrandjean.ch/gephi-introduction/

For each dataset it is required to

1)Compute the following statistics:

- Number of node/edges
- Min, max and average nodes degree
- Diameter
- PageRank max and average
- Clustering Coefficent
- Number of Connected Components (provide numbers and a snapshot of the network where each component is assigned a different color)

2)Compute the main 4 communities by adjusting the appropriate parameters in modularity (provide a snapshot of the network where each component is assigned a different color)

3)Compute the main 2 communities by adjusting the appropriate parameters in modularity (provide a snapshot of the network where each component is assigned a different color)

4)Is there a component including at least 50% of the nodes?

5) Is there a community in case 2) including at least 30% of the nodes?

6) Calcolate degree and eigen value, then visualize network ***using degree, eigenvector centrality , betweeness centrality, closeness*** as node dimension and an appropriate layout.

7) Compute the main 4 community on the network limited to the nodes with highest 25% in-degree

8) Write your network analysis conclusions 2basing on the previous steps

Data Set Aggiuntivi 2020

Quest'anno dataset sono relativi a notizie "cronaca" su argomenti di attualità.

Utilizzando la funzione di ricerca avanzata online del giornale la Repubblica

http://ricerca.repubblica.it/ricerca/repubblica?query=&view=repubblica&ref=HRHS

- si effettui la ricerca su ogni singolo giorno del periodo indicato con la/le parole chiave indicate, servendosi se lo si ritiene utile dello scrit python fornito dal docente
- si salvi il testo della pagina web come "testo"
- Filtrando i dati di intestazione e footer del file di testo si individuino i titoli dei singoli articoli, si eliminino preliminarmente le "stop word" e si costruisca la rete così definita:
  - un nodo per ciascuno *NomeMaiuscolo* (come definito di seguito)
  - un nodo per ciascuna parola *ParolaNormale*, ovvero che non sia una "stop word" che compare nel titolo
  - un arco *non orientato* per ciascuna coppia di nodi di tipo *NomeMaiuscolo* e/o *ParolaNormale* che compaiono nel titolo di uno stesso articolo, incluse le versioni graficamente diverse di tutti i NomeMaiuscolo

  Per costruire la rete si faccia riferimento ad uno dei formati testo importabili in Gephi, potrà essere necessario scrivere uno script o un programma che legge i dati dal file scaricato e li formatta opportunamente.

NomeMaiuscolo: qualsiasi stringa o sequenza di stringhe che inizi con il primo carattere maiuscolo e gli altri minuscolo che a)non sia in inizio riga, b) non segua un punto "." o un apice singolo ' o doppio " o un punto esclamativo "!" o interrogativo "?". Esempio nella seguente frase: "Macron e Merkel hanno incontrato TRUMP e parlato Kim Joun Un riguardo ai missili USA e alle sanzioni Onu."

verrano identificati i nodi "Merkel" "Kim Jong Un" e "Onu" e creati gli archi (Mekel, KimJongUn) (Merkel, Onu) (Kim Jong Un, Onu), se in un articolo precedente fosse stato creato il nodo "Usa" allora anche "USA" sarebbe stato riconosciuto e generati gli archi (Usa, Merkel)(Usa, Onu)(Usa,Kim JongUn)

Ad esempio, se un titolo ottenuto con la parola chiave "covid" contiene la frase "Donald Trump combatte il virus"

1) si eliminino le stop word, in questo caso "il"
2) si generino i *NomiMaiuscoli*, in questo caso "Donald Trump"
3) si generino i nodi "DonaldTrump", "combatte" e "virus", aggiungendoli alla rete se non esistono
4) si aggiungano alla rete, se non esistono, gli archi
   (DonaldTrump , combatte, 1)

   (DonaldTrump , virus, 1)

   (combatte, virus, 1)

   Oppure si incrementi di 1 il loro peso di tali archi, se esitono già
5) si ripeta per ogni titolo
6) si importi in gephi il file CSV o in altro formato compatibile così creato

A partire dalla parola chiave o dalla “frase chiave”, si devono creare i dataset in due parti (a) e (b):

Prima parte dataset (a): Periodo dal 07/01/2020 al 07/02/2020

Seconda parte dataset (b): Periodo dal 15/03/2020 al 14/04/2020

Prima si **analizzino separatamente** le due reti così ottenute con Gephi, utilizzando i criteri e le modalità prima descritte, si analizzino **poi congiuntamente** e si predisponga un report da consegnare assieme ai dati, in cui per ogni rete separata e per quella congiunta si effettuano considerazioni, motivate dai dati dell’analisi. Oltre ai dati ed i commenti/spiegazioni il report dovrà contenere i valori numerici richiesti, alcuni grafici di distribuzione e visualizzazioni significative della/e rete/i.

Data set 1 Parole chiave: **governo 187-277**

Data set 2 Parole chiave: **salvini 111-25**

Data set 3 Parole chiave: **covid 6-329**

Data set 4 Parole chiave: **cina 92-87**

Data set 5 Parole chiave: **lombardia 41-147**

Data set 6 Parole chiave: **conte 91-94**

Data set 7 Parole chiave: **chiusura 67-126**

Data set 8 Parole chiave: **“restare a casa” 1-43**

Data set 9 Parole chiave: **mascherine 14-192**

Data set 10 Parole chiave: **ospedali 22-153**

Data set 11 Parole chiave: **borrelli 9-21**

Data set 12 Parole chiave: **speranza 65-111**