Update Your Project Allocation at https://1drv.ms/w/s!AtcJs3OTsMZuiRt8k7S3z22CATjl

Fake News Related Projects

Project 1: Fake News Twitter Analysis

Consider the 2016 US election Viral Twitter dataset collected between election day (Nov 8th) and March 2017. Tweets have been labelled as containing fake news or not by two sets of people, and where fake news is categorized into one of the five categories: i) Serious fabrication; ii) Large-scale hoaxes; iii) Jokes taken at face value; iv) Slanted reporting of real facts; v) Stories where the 'truth' is contentious. The dataset can be downloaded from Fakenews on 2016 US elections viral tweets (November 2016 - March 2017) | Zenodo

1. Separate from the overall dataset two classes of tweets: one related to labelled Fake News (regardless of the category of the fake news) and the other one for Real News. Save each tweet class in a separate file. Write a script to identify the number of distinct users (user_screen_name) in Fake News class, number of distinct users in Real News and number of distinct users who participated to both Fake News and Real News.
2. Write a script for the calculus of the mean, standard deviation, kurtosis and skewness of Number of follower per user in case of Fake News and Real News dataset. Repeat this process for Favorite Count as well. Conclude whether one can discriminate the two classes using such statistical data.
3. We want to compare the activity of individual users in Fake News and Real News dataset. Select the three most active users in terms of number tweets generated and calculate the average number of tweets generated by the three users. Repeat this process for the five most active users in each dataset, and for the first 10-users in each dataset, and first 15 users for each dataset.
4. We want to compare the average time a user stays before sending a new message. Write a script that uses the date information on the dataset for each tweet to calculate the average waiting time for a random user before sending a new message in case of Fake News and Real News.
5. Study the behavior of user ids who contributed to both Fake News and Real News. Based on your observation from 2)-3)-4) and any other scrutinizing, suggest statistical index that would discriminate Fake News and Real News tweets of the same user.
6. We want to create a social network from the Twitter dataset. For this purpose, consider the mention reference in Twitter. More specifically, write a small program that allows you to identify the “mention” in each tweet message (word precedent by “@”). Now, construct a network graph where the nodes correspond to the user ID while the edge between two nodes, say A and B, indicates that tweet of user id A contains in its text message a mention of user id B. Construct social network graph for each dataset (Fake News and Real News). Use appropriate visualization to draw high level illustration of each graph.
7. Use appropriate functions in NetworkX to calculate diameter, average clustering coefficient, average degree centrality, average closeness centrality and average betweenness centrality for each dataset.
8. Calculate the degree centrality distribution and clustering coefficient distribution for each dataset and draw the corresponding plot. Discuss your result and whether this can discriminate the cases.
9. Use VADER tool(https://github.com/cjhutto/vaderSentiment) to perform sentiment analysis on tweets of each dataset. For each user id, we want to calculate the distribution of sentiment (proportion of positive, negative and neutral sentiment), then one represents the distribution of each user as a point in the ternary plot. Repeat this process for the second dataset as well, so that two distinct ternary plot will be exhibited. Conclude whether sentiment can differentiate the two datasets.

10. We want to use the information about the various Fake News categories. Using the information about the tweets, where either a given user id sends tweet messages who are categorized in different categories, or he send a tweet message that contains a mention of a user id who is assigned to another category. Write a script to perform this operation, and then output a simple graph where the nodes are constituted of the five categories and the edge indicates a link as previously described.

11. Suggest appropriate literature in fake news identification to discuss and comment on your findings at each level of the above analysis.

Project 2:

Consider the FakeNewsNet dataset, available at GitHub - KaiDMML/FakeNewsNet: This is a dataset for fake news detection research. Use the provided code to generate all Tweet attributes (Number of retweets, User followers, User following, Retweet, Number of likes, etc (whatever is available through the Twitter API)) for each of the four data categories (Glossipcop Fake News, Glossipcop Real News, Politifact Fake News, Politifact Real News). You can also consult the FakeNewsNet reference paper of Shu et al. arXiv:1809.01286 for detailed explanation of the dataset. You will realize that not all the dataset can be reconstructed as many tweet id may not be available and in API call limit.

1. For each category dataset, provide a table describing the statistical trend of the key attributes. This consists of: i) Number of tweet messages, ii) Number of distinct user ids, iii) mean, standard deviation, kurtosis and skewness of number of retweets per user id; iv) iii) mean, standard deviation, kurtosis and skewness of number of following per user id; v) mean, standard deviation, kurtosis and skewness of number of followers per user id. Discuss whether you can discriminate between fake news and real names on the basis of these attributes.
2. Draw on the same plot the distribution of follower count for Fake News and Real News of glossipcop and politifact data. Repeat the process for the distribution of followee count for fake news and Real News.
3. Explore the temporal evolution of user's engagement (according to your suggested approach to quantify the user's engagement (i.e., number of likes, some combination of followers and followees, etc.) for Fake News and Real News, and draw the corresponding plot for both glossipcop and politifact data.
4. We would like to study whether the user ids of fake news and real news dataset are genuine or not. For this purpose, study the program botometer available in https://github.com/IUNetSci/botometer-python. The program inputs a tweet user id and outputs the probability that the user id is a bot or human. You can use a threshold 0.5 beyond which a program is bot or not. The purpose is therefore to test the hypothesis whether Fake News are globally originated from bots or humans and whether Real News are generated by humans or also by bots. If the computational time is an issue to test the whole data, you can choose a random selection of the data as well.
5. We want to explore the graph structure that can be extracted from the dataset and compare the properties of fake news and real news categories. For this purpose, consider the follower relationship, where user id A is linked to user id B if either A (resp. B) is a follower of B (resp. A). We restrict only to those user ids who are associated to dataset tweets (Need to retrieve the list of followers for each user id to test whether this relation holds). Use NetworkX to calculate global attributes of this network such as overall degree centrality, diameter, clustering coefficient, size of largest component. Compare these graph attributes for Fake News and Real News for glossipcop and politifact data. Use high level illustration to draw the network of each one.
6. Draw on the same plot the degree distribution of fake news and real news for each of glossipcop and politifact data. Conclude whether some graph attributes are relevant to distinguish fake news and real news.
7. Use relevant literature from fakes news detection from social media to discuss your finding at each level of the preceding reasoning.

Project 3. Health Fakes Diffusion

This project considers the FakeHealth dataset available at https://github.com/EnyanDai/FakeHealth which includes HealthStory and HealthRelease dataset. We shall restrict to the first dataset only. You may notice that the reconstruction of the dataset from the provided tweet id will not match the original number as some tweets may be deleted or profile switched to private.

1. Provide a table summarizing the global attributes for Fake and Real part of of the dataset, which consists on i) number of tweets, average number of tweets per news (together with corresponding standard deviation, kurtosis and skewness), average number of tweets per user per news (together with corresponding standard deviation, kurtosis and skewness), average replies per news (together with corresponding standard deviation, kurtosis and skewness), average replies per tweet (together with corresponding standard deviation, kurtosis and skewness), average retweets per news (together with corresponding standard deviation, kurtosis and skewness), average retweets per tweet (together with corresponding standard deviation, kurtosis and skewness). Discuss whether any of these global attributes allow you to make a clear distinction between Fake and Real dataset.
2. Assign a single user id for each news in Fake and Real dataset and use Twitter API to retrieve the number of followers and followees.
3. Draw on the same plot the distribution of follower count for Fake and Real of HealthStory dataset. Repeat the process for the distribution of followee count for fake and Real data.
4. Show whether power law distribution can be fitted to the above plots.
5. Explore the temporal evolution of user's engagement (according to your suggested approach to quantify the user's engagement as a function of number of replies and retweets for Fake and Real and draw the corresponding plot for HealthStory data.
6. We want to investigate whether some fake data are genuine or not. study whether the user ids of fake news and real news dataset are genuine or not. For this purpose, study the program botometer available in https://github.com/IUNetSci/botometer-python. The program inputs a tweet user id and outputs the probability that the user id is a bot or human. You can use a threshold 0.5 beyond which a program is bot or not. The purpose is therefore to test the hypothesis whether Fake News are globally originated from bots or humans and whether Real News are generated by humans or also by bots. Draw a plot showing the proportion of bots in Fake data and Real data.
7. Now we want to test the hypothesis whether a fake news occurs if the initiator (user id) is communicating with bots. For this purpose, for each news (in Fake data), select 100 random user id among those associated to that news, and apply the previous botometer and output the number of users id that are found to be bots. Plot the distribution of number of bots per news in Fake data.
8. Use VADER tool (https://github.com/cjhutto/vaderSentiment), which output sentiment in terms of POSITIVE, NEGATIVE and NEUTRAL to determine the sentiment of each news in Fake and Real data. Then represent the distribution of each news statement as a point in the ternary plot for both Fake and Real data. Conclude whether sentiment can differentiate the two datasets.
9. Suggest how you can take into account the criteria C0-C10 provided in the dataset to fine-tune the reasoning in 6-7).
10. Identify relevant literature in fake new identification and health literature to back up your finding in previous sections.

Project 4: Covid-19 diffusion network

This project aims to investigate the extent to which the diffusion models can be fitted to actual data of Covid-19 infection and recovery statistics.
Consider the official statistics on covid-19 infection and recovery provided by official organizations such as https://www.worldometers.info/coronavirus. Consider a reasonable time period for a selected country of your choice, should be a small country to make the subsequent computational time feasible.

1. Use a starting date where you consider it to stand for initial state. In the statistics of the country at the chosen, calculate the initial Infection I0 as the total number of infection minus the total recovery. Use the official corona statistical source to draw a plot showing the temporal variations of the number of infections and that of the number of recovery.
2. We want to carry on the simulation using the SIR epidemic model. Use the implementation provided in NDLIB library to perform the calculus. Set the number of nodes of the network equal to the total population and a very small probability for Erdos random graph of 0.001. Choose a infection probability beta and recovery probability beta of your choice (you may inspire from the data trend). Run the EDLIB and plot the temporal variation of the Number of infection and recovery over time.
3. Now we want to use the data of official statistics to tune the probability of infection and recovery to find a way to match the variations plotted in 2) with that of 1). Suggest an empirical approach where, for instance, you vary incrementally the values of alpha and gamma until you visualize a figure infections and recovery count closely match that of official statistics.
4. Now we want to use the official dataset statistics to estimate the probability of infection and recovery. Suggest a simple approach to calculate these attributes using the available historical dataset. Then input these values to the SIR model and run the simulation to display the variation of the infections and recovery. Discuss the relevance of the SIR model for this purpose.
5. Now we want to treat the death count provided in the statistics. Consider using the SI model for this purpose. Similarly to 1), draw the timely evolution of the number of death.
6. Next use the implementation provided in EDLIB for the SI model and suggest a simple model to generate the simulated model that displays the total number of death.
7. Suggest an empirical and incremental variation of the infection probability in SI model until the death variation is close to the real dataset. Discuss the relevance of such approach and probability value.
8. Consider the SEIRS (Susceptible-Exposed-Infected-Recovered-Susceptible) model described in https://github.com/ryansmcgee/seirsplus. Set an initial value of parameters of the model and display the temporal evolution of the infection and recovery counts.
9. Similarly to 3), suggest an empirical and possible an incremental approach to attempt to match the infection and recovery counts with that of real dataset.
10. Similarly to 4), use the official statistics to infer the infection, recovery probabilities and other parameters of the model. Then run the model again and display the new graph showing the variations of the infections and recovery counts.
11. Use relevant literature to back your reasoning and finding in previous steps.

Project 5: Analysis of Smoking Cessation

The project aims to study the online community of smoking cessation users in Twitter social network.

1. Identify two hashtags related to smoked cessation. Examples include #Stopsmoking, #smokefree, #Stoptabac, etc.. Show your reasoning to identify few other equivalent hashtags as well. Collect a sufficient number of tweets related to each hashtag (around one thousand tweets). For this purpose, you can use for instance Tweepy (see tutorial at https://riptutorial.com/tweepy), also see examples of text processing in Python in NLTK online book at https://www.nltk.org/book/. You would need to create your Twitter API account credential. The key in the collection process is that there is an important number of tweets that contain other hashtags as well. It is also important to leave the collection open to other non-English tweets. Save the attributes of the tweets for each hashtag in excel database. This includes the Twitter ID, tweet message, list of followers of the tweet user, whether it is a retweet or not, location, if available.
2. Draw a histogram showing the popularity of the main hashtags highlighting the number of tweets per individual hashtag and in another graph the number of distinct Tweet users per individual hashtag.
3. Draw another histogram showing the proportion of tweets where location information is mentioned (if location attribute is activated in the tweet) and another one for the language of the tweet messages. Represent the finding through pie chart.
4. We now want to build a social graph where each node corresponds to a hashtag and an edge between hashtag A and hashtag B indicates that there is at least one tweet which contains both hashtag A and hashtag B. Implement a small python program that allows you to generate the above social network graph from the collected tweets.
5. Summarize in a table the main global properties of the above graph: Number of nodes, Number of edges, average degree centrality and its variance, average in-betweeness centrality and its variance, average path length and its variance, diameter, clustering coefficient, size of largest component using appropriate NetworkX functions. Comment on the obtained results highlighting any inherent limitation or characteristic of the data collection process.
6. We want to see the extent to which some global attributes can be imitated using random graph. For this purpose, use appropriate functions of NetworkX concerning the small word model whose number of node is equal to the total number of nodes of the graph in 5) but whose probability p can be chosen in such a way the clustering coefficient of this random graph approximates that of the graph in 5). Make incremental change of the probability value p until this approximation holds.
7. Identify the five highest ranked nodes in terms of degree centrality, Katz's centrality, PageRank centrality, Closeness centrality, Betweeness centrality of the graph obtained in 5).
8. Use appropriate NetworkX functions (or other alternatives) to display the distribution of the degree centrality and that of the local clustering coefficient and local betweenness centrality and closeness centrality.
9. Now we want to comprehend the key players in the graph whether they are genuine or not. For this purpose, consider the 10 most ranked in terms of degree centrality, and scrutinize the tweets which are linked to those hashtags (10 most ranked). The scrutinizing consists in applying the botometer available in https://github.com/IUNetSci/botometer-python. The program inputs a tweet user id and outputs the probability that the user id is a bot or human. You can use a threshold 0.5 beyond which a program is bot or not. The purpose is therefore to test whether a given user id (Tweet user of the tweet that contains the hashtag) is a bot or not. Draw a relevant plot which shows the proportion of bots in the top ranked hashtag.
10. We would like to test the amount of support assigned to each hashtag. For this purpose use the information about number of retweets and number of replies of each tweet involved in the hashtag and suggest an expression that quantifies the amount of support.
10. Comment the results obtained and summarize the key finding regarding behavior of users with respect to smoking cessation. Seek some literature to reinforce your interpretation.

Project 6: Authorship Network Analysis

This project aims to investigate the social network of co-authorship from DBLP (http://dblp.uni-trier.de/).

1- Perform an initial exploration of the database, e.g., going to Statistics attribute in the tab Home. You will find various statistics. Check the authors with the highest number of co-authors. Make sure to disambiguate the outcomes to not account for authors that are affiliated to more than one institutions. Choose two authors (ideally from distinct disciplines) among the top five who scored best in terms of number of co-authors in the database.

2- For each of the two selected authors in 1), retrieve the statistics of each of his/her co-authors in terms of number of his/her co-authors again, and calculate the average value. You may use the availability of the DPLP search API. Compare the average co-authorship statistics for each of the two authors in 1).

3- We now consider the full co-authorship network of each selected author in 1), so the nodes of the network correspond to author names. Besides, for sake of simplicity, we shall consider that an edge between two nodes (authors) is established if the two authors have co-authored at least two papers together. Retrieve the full network structure for each case in 1). Use NetworkX to display dense parts of each network.

4- Use networkX to calculate and summarize in a table the global attributes to each network; namely, clustering coefficient, average degree centrality and its variance, average closeness centrality and its variance, average betweenness centrality and its variance, diameter, size of giant component.

5- Trace the degree centrality distribution for each network and check whether Power-law distribution can be fit using simple statistical test available in python cure fitting library.

6- Use a community detection algorithm of your choice to identify the main communities for each network. Suggest a heuristic search that would allow you to assign potential interpretation to each detected community

7- Repeat tasks 3-6 when using the affiliations of the authors as nodes instead of authors' names. Write pseudo-code that allows you to generate the corresponding network and use appropriate function to visualize some interesting parts of the network.

8. Consider the result in terms of diameter and global clustering coefficient obtained in 4), identify appropriate Erdos random graph whose diameter and global clustering coefficient are close to the real network. (Need to vary the probability p from 0 and 1 with a network whose number of nodes is equal to that of the real network).

9- Use relevant literature to comment on the obtained results.

Project 7: Citation Network

Consider the citation network Citation-network V1, available from https://www.aminer.cn/citation. The dataset contains 629,814 papers and more than 632,752 citation relationships (2010-05-15), where for each paper, are provided Title, Authors, Year, Publication venue, Index ID of the paper, IDs of all references listed in the paper, and finally, Abstract of the paper.

1. Consider the network where the nodes correspond to paper IDs and an edge from node A to node B is established if paper A cites paper B. Use networkX functions to provide in a table the main global characteristics of this graph in terms of number of edges, number of nodes, diameter, clustering coefficient, number of components, average degree centrality.
2. Use NetworkX to calculate the in-degree centrality, out-degree centrality and PageRank centrality of each node and save the result in a file. Plot the in-degree, out-degree and pageRank centrality distributions on the same graph and comment on the variations between these three measures.
3. Check whether power law distribution can be fit to in-degree /out degree /pageRank degree distributions. Justify your answer.
4. Use label propagation implementation in NetworkX to determine the various communities in the network. List the various communities outputted by the algorithm, summarizing in a table the main characteristics of each community in terms of nodes, edges, clustering coefficient and diameter.
5. Write a program that allows you the extent to which the identified community agree with the publication venue attribute in the dataset.
6. Now assume that the nodes correspond to author name and a link between two nodes indicates that the two authors co-written the same paper. Use networkX to determine the global properties of the network as in 1).
7. Identify the author who has highest number of collaborators. Assume this author plays the role of Erdos in Erdos network. Calculate the histogram of Erdos number when considering all authors. What is the maximum, minimum and median Erdos number across all authors in the network.
8. We shall consider the effective distance of two authors, say, i and j defined by
   $d_{ij} = 1 - \log(f_{ij}/F_i)$,
   where f_{ij} corresponds to the number of times authors i and j co-authored together and F_i is the number of the total number of collaborations held by author i.
   Suggest a script that calculates the d_{ij} for all authors. Identify the authors corresponding to the ten most largest effective distances.
9. Discuss the results of d_{ij} with respect to Erdos number and comment whether possible connection exists between the two concepts. Identify relevant literature in the field to justify your arguments.

Project 8. Graph and Semantic Analysis of Movie database 1

Explore the IMDB 5000 Movie database available at https://data.world/data-society/imdb-5000-movie-dataset . The excel file can also be accessed at shared Google drive https://1drv.ms/x/s!AtcJs3OTsMZuiRctyVt3IVt4lSup

The database contains several attributes for each movie, including main actor, second actor, third actor, director, movie genre, various user ranking attributes, budget, keywords. We shall consider a network where the nodes correspond to the actor names and the link is established whenever the two actors are played together in at least one movie.

1. Suggest appropriate preprocessing to make the dataset clean.
2. Use Networkx appropriate functions in order to study the properties of this network by summarizing in a table its key characteristics, which contain: i) Number of nodes, ii) Number of edges, iii) Overall clustering coefficient, iv) Diameter, v) Number of components and size of its largest component, vi) Average path length, vii) Maximum degree, average degree, and minimum degree centrality; vii) maximum / average and minimum eigen vector centrality, ix) maximum / average / minimum betweenness centrality.
3. Save in a file the degree centrality of each node and draw the degree distribution. Then check whether a power-law distribution can be fitted or not. Justify your answer
4. Save in a file the result of eigenvector centrality of each node. Suggest a subdivision (histogram bins) of these values that take into account the variability of the centrality values. Draw the corresponding degree distribution and check whether a power law can be fit.
5. Repeat question 4) when using betweenness centrality.
6. Provide the ten highly ranked actors according to each of the centrality measure: degree centrality, eigenvector centrality, betweenness centrality. Comment on the overlapping and discrepancy between the three centrality measures.
7. Use appropriate NetworkX functions to determine communities using clique, k-clique, and Girvan-Newman algorithms. Summarize in a table the main characteristics of each community.
8. We would like to investigate the communities identified by those algorithms for the original movie network. For this purpose, we would like to test some hypotheses. The first hypothesis is that the community corresponds to movies belonging to same category, e.g., Action movies, Romance, Science Fiction, etc. The second hypothesis is that community corresponds to movies belonging to same series (episode). Use the information of your movie database that you have selected (e.g., Internet Movie Database) to test the two hypotheses. Propose a more formal evaluations.
9. Discuss whether some communities match other attributes of the database (keywords, genre, country, budget..)
10. Provide a methodology and a script that computes actor ranking based on the ranking of movies the actor has participated (a simple way to do so is to average the ranking of movies the actor is involved with but other options are also possible).
11. Now we would like to compare how the ranking data matches with various centrality measures in order to identify whether a given centrality measure better agrees with the ranking outcomes. For this purpose, construct a histogram distribution showing the proportion of actors whose ranking values fall within a given subdivision of the histogram (you should transform the ranking data into a normalized scale, e.g., within [0,1] interval in order to ease comparison with other measures). Similarly, construct for each centrality measure in 1) a histogram showing the proportion of the actors whose centrality measure falls within the corresponding histogram bin. Again for ease of comparison, the centrality measures should be transformed to normalized scale. Finally use a simple distance between histograms in order to calculate the distance between the ranking histogram and the underlined centrality measure. Identify the best matching centrality measure accordingly.
10. Use potential literature from entertainment and movie making in order to comprehend the results of your finding.

Project 9. Graph and Semantic Analysis of Movie database

Consider again the IMDB 5000 Movie database available at https://data.world/data-society/imdb-5000-movie-dataset . The excel file can also be accessed at shared Google drive https://1drv.ms/x/s!AtcJs3OTsMZuiRctyVt3IVt4lSup

The database contains several attributes for each movie, including main actor, second actor, third actor, director, movie genre, various user ranking attributes, budget, keywords. We shall consider a network where the nodes correspond to the actor names and the link is established whenever the two actors are played together in at least one movie.

1. Consider the information of rating attribute of movies in terms of total Facebook Likes. Normalize the Likes score with respect to the number of voters. Plot the rating distribution (histogram of rating over all movies), and show the mean, median and standard deviation of rating values.
2. Write a script that calculates the (average, standard deviation) rating (Likes) per actor by averaging over all movies the actor participated in. Identify the top scoring actors accordingly.
3. Use the information of the actor Likes in the database to find out how the rating per actor obtained from 2) agrees or deviates from the actor Likes attribute. Draw on the same plot the two evaluations as a vector of actors.
4. Use Networkx appropriate functions in order to study the properties of this network by summarizing in a table its key characteristics, which contain: i) Number of nodes, ii) Number of edges, iii) Overall clustering coefficient, iv) Diameter, v) Number of components and size of its largest component, vi) Average path length, vii) Maximum degree, average degree, and minimum degree centrality; vii) maximum / average and minimum eigenvector centrality, ix) maximum / average / minimum pageRank centrality.
5. Save in a file the PageRank centrality of each node and draw the degree distribution. Then check whether a power-law distribution can be fitted or not. Justify your answer
6. Identify the node in the network with the highest degree. Assume this node plays the role of Erdos number. Write a script that allows you to calculate the Erdos number of each actor, then plot the Erdos number distribution.
7. Suggest a script that calculates the correlation between Erdos number and actor's Likes as provided in the database, and the correlation with actor's rating as inferred in 2). You should consider statistical Person correlation coefficient.
8. Use appropriate NetworkX functions to determine communities using label propagation algorithm. Summarize in a table the main characteristics of each community in terms of nodes, edges, diameter, and clustering coefficient.
9. Explore the correspondence of the identified communities with genre, keywords, budget, … attributes provided in the database.
10. Use potential literature from entertainment and movie making in order to comprehend the results of your finding.

Project 10. Mining Violence in Suomi24

The interest focuses on mining the discussion related to violence in Suomi24 Finnish forum, one of the largest Finnish internet corpus where users discuss all topics.

1. Use the online version of Suomi24 in www.suomi24.fi . Alternatively, if you have enough computational resources, you can also download the Suomi24 corpus history from The Suomi24 Corpus 2001-2017, VRT version 1.1 published in Download service | Kielipankki.
2. Construct a list of Finnish keywords related to violence (should be broad enough to include all aspects, i.e., abuse and insult related wording, hate speech related words, common bullying related abbreviation). Elaborate your own methodology to identify a large scale cyber-bullying related terms.
3. Run a simple keyword matching in Suomi24 dataset or in the online portal (crawl all search outcomes) in order to extract only those posts and the associated threats where there is a matching. Save the newly constructed database, which contains both the identified posts and associated threads.
4. Draw a bar plot showing the proportion or number of hits for each individual violence word. Draw also another plot showing the proportion of hits found on the title of the threads only.
5. Construct a social network in the following way. The nodes of the network are constituted of the set of all threats of the search outcome. An edge from a threat A to a thread B is established whenever the same violence keyword is mentioned at one post of thread A and one post of threat B.
6. Study the properties of this constructed network by reporting the number of nodes, number of edges, maximum degree, average degree, global clustering coefficient, diameter, average path length, size of giant component, size and number of communities as well as the associated quality measure.
7. Draw the degree distribution and check whether a power law distribution can be fit
8. Repeat questions 5-6, when introducing a threshold regarding the numbers of mentions of same keywords among two threads before deciding to draw an edge between the two nodes. Namely, an edge between thread A and thread B is established if there are at least k violence keywords contained in both thread A and thread B. (You can start by k=2, k=3, k=5,..). Draw a plot showing the evolution of each attribute of the network (size of giant component, average degree centrality, average path length, diameter and clustering coefficient) according to the value of k.
9. We want to test the extent to which the reciprocity relationship is fulfilled. More specifically, we want to find out whether a violence attack automatically generates a reciprocal attack. For this purpose, you need to take into account the timestamp of the posts. Therefore, we assume that whenever a thread contains an even number of violence keywords, then the reciprocity is fulfilled for the underlined thread (node). Draw a bar plot showing the proportion of threads whose reciprocity is satisfied and those not.
10. Comment on the key findings by identifying key sociology studies that support your argumentations.

Project 11: Mapping Covid-19 Vaccine Discussions in a Blog Forum

This project aims to investigate the mental health discussion taking part around Covid 19 vaccination available in Have you had covid vaccine side effects? - Health and Wellness -Doctors, illness, diseases, nutrition, sleep, stress, diet, hospitals, medicine, cancer, heart disease - City-Data Forum (city-data.com). The thread contains large number of posts. Interestingly each post contains statistical information about the author in terms of number of posts made by the author, reputation and number of reads as well as location of the author.

1. Use your own way to crawl the whole data and available statistical attributes (through API, beautifulsoup, copy and past at last resource if no automatic procedure can be implemented). Show your reasoning how this has been performed.
2. Use the location information of the authors to provide the distribution of the location in terms of number of posts generated. Show whether the Power law distribution can be fit.
3. Build a simple program that allows you to output the length of the post in terms of number of words / characters it contains.
4. Create your own subdivision of the length of the posts (e.g., length less than k1, length between k1 and k2, length between k2 and k3, …) and draw a histogram showing the number of hits in each bin. Comment on the distribution of the hits accordingly.
5. Repeat step 4) for the top 5 regions in terms of number of posts generated. Comment whether length of the post can be used as an attribute to discriminate the regions in this dataset.
6. Many posts in the dataset are written as reply to some other posts (This occurs when at the beginning of the post, there is a Quote where the name of the user is also mentioned). Consider a network graph constructed using this mentioning relation where nodes are the user names and an edge between two user names is established if one user name is mentioned in the quote of the post of the other user name (no need to be reciprocal). use appropriate NetworkX functions to plot this graph.
7. Provide a table showing the global attributes of this social graph in terms of number of nodes and edges, diameter, number of connected components, average clustering coefficient, average degree centrality and average degree closeness centrality.
8. Plot the degree centrality distribution and the local clustering coefficient distribution. Comment whether a power law distribution can be fit to the plot.
9. Use Girvan-Newman algorithm to find communities in the above network through appropriate use of NetworkX functions. Compare the size of the generated communities in a table.
10. Use the author's Reputation information to identify communities that have higher reputation. You can simply consider the reputation of a community as the sum of the reputations its members.
11. Discuss and comment, and use appropriate health literature in order to reinforce your interpretations.

Project 12. Social Network Blog Analysis

Choose an active blog community of your choice with an available API to ease data collection.
Proceed in the following way to construct the social network graph.

- Start with a list of most cited blogs at a specific time of your choice and select a time window (it should include the time of most cited blog) that you can use to collect posts and blogs occurring within that time interval.
- Make some reasonable assumptions in terms of the maximum number of posts that will be retrieved.
- Typically, each post contains a link of the parent blog, date of the post, post content and a list of all links that occur in the post's content.

1. Elaborate on the choice of blogs and size of data collection.
2. Plot the number of posts per day over the span of the collected dataset
3. We would like to represent the collected data as a cluster graph where clusters correspond to blogs, nodes in the cluster are posts from the blog, and hyper-links between posts in the dataset are represented as directed edges. Only consider out-links to posts in the dataset. Therefore, remove links that point to posts outside the collected dataset or other resources on the web (images, movies, other web-pages), and also those edges that point to themselves if any. This is to keep track of timestamp for temporal analysis.
4. Study the global properties of the established network: number of noes, number of edges, clustering coefficient, diameter, size of giant component, average in-degree centrality and out-degree centrality and their associated variance, average path length and its variance, average closeness centrality and its variance, average in-betweeness centrality and its variance.
5. Trace the in-degree centrality distribution and check whether a power-law distribution can be fit using appropriate statistical testing
6. Trace the out-degree centrality distribution and check whether a power-law distribution can be fit using appropriate statistical testing
7. Now investigate the temporal variation of popularity. For this purpose, collect all in-links to a post and plot the number of links occurring after each day following the post. This creates a curve that indicates the rise and fall of popularity. By aggregating over a large set of posts, you should a obtain a more general pattern.
8. Check whether a power-law distribution can be fit
9. identify appropriate literature to comment on the obtained results and the limitations

Project 13. Covid-19 and Hashtag diffusion

This project investigates a large scale Covid-19 Twitter dataset available at https://github.com/lopezbec/COVID19_Tweets_Dataset. The dataset is organized by hour (UTC) and each hour contains five tables: (1) "Summary_Details", (2) "Summary_Hastag", (3) "Summary_Mentions", (4) "Summary_Sentiment", and (5) "Summary_NER (Named-Entity-Recognition)". The dataset is made of billions of tweets and still is constantly updated. The summary hashtag consists of the top five popular hashtags in tweets collected at a given hour (UTC). It also provides for each tweet Likes count, Retweet count and sentiment label.

1. Use appropriate tool to save the dataset in appropriate format. The actual text of the tweet message is not needed (tweet id will be enough). Using the information in the Sumary_Hashtag attribute of the data, draw a plot showing the distribution of the hashtags in terms of number of tweets citing the hashtag. Does the graph follow a power law distribution? Use statistical significance of curve fitting to show whether such fitting is significant or not.
2. Repeat the preceding for the named entity as provided in Summary_NER attribute, and indicate whether a power law distribution can be fit or not.
3. We want to focus on the timely evolution of the hashtags. Consider the five most frequent hashtag (cited by largest number of tweets) that you may infer from 1).
4. We want to reconsider the top hashtags by taking into account the replies and likes count. For this purpose, assume that the score of the hashtag is calculated so that in first case (replies), we add the replies count of each tweet that contains that hashtag. While in the first case, we will use likes count instead of relies count. By doing so, identify the top five hashtags according to replies count, and the top five hashtags according to likes count.
5. For each of these hashtag, suggest a plot which shows the timely evolution of this hashtag over a period of few months. You may create a weekly subdivision, where you count the total number of mentioning of this hashtag for each week, and then draw a plot of count versus weeks. Also for each week calculate the statistics of the hashtag count in terms of average, standard deviation, kurtosis and skewness. You may also plot to show on the same graph the evolution of the mean and standard deviation. Identify, whether you may notice cases where some weeks have zero count and then start picking up again. Discuss the evolution of the various hashtags according to replies and likes count.
6. We would like to evaluate the evolution of each hashtag in terms of sentiment score. For this purpose, use the logits data provided in sentiment attribute of the dataset. More specifically, for each week, add the logist_negative (as well logist_positive, and logist_neutral) of all tweets mentioning the underlined hashtag. The hashtag will therefore be assigned a sentiment label that has the highest value among negative, positive and neutral logist values. Use a plot where you represent the evolution of the positive by its positive score, while negative sentiment is represented by a negative value (where the value corresponds to the total logist_negative). Discuss how hashtag count correlates with sentiment score.
7. Now we want to model the speed of hashtag diffusion over the network. Consider that the propagation speed is defined by the following.

$$Ps = (R_1 + R_2 + .. + R_n) / n$$

where R_i is total count of retweet of all tweets mentioning the hashtag S in week i. n is the total number of weeks

Use the above formula to calculate the speed of the hashtag at three different periods that you may distinguish: starting time, peak time and flat time where the associated tweets are getting less replies score.

8. Comment on the results using identified literature of Covid-19 of your choice.

Project 14 Reddit Community Analysis

Consider the Reddit data, accessible from reddit user posting behavior (mid-2013) (figshare.com)
The dataset contains user activity from over 876,961 Reddit users across 15,122 subreddits, created in mid 2013. Each entry in this dataset provides a user ID and a list of subreddits within which that user is active. A user is considered active in a particular subreddit if at least 10 of their 1,000 most recent posts or comments were made in that subreddit. Follow the network construction described in Detection and Analysis of Subreddit Communities (samgriesemer.com) (Section 3).

1. Use networkX functions to provide in a table the main global characteristics of this graph in terms of number of edges, number of nodes, diameter, clustering coefficient, number of components, average degree centrality.
2. Use NetworkX to calculate the degree centrality, betweenness and closeness centralities of each node and save the result in a file. Plot the degree, betweenness and closeness centrality distributions on the same graph and comment on the variations between these three measures.
3. Check whether power law distribution can be fit to each type of the three similarity measures. Justify your answer.
4. We want to investigate the behavior of the top influencing nodes. Consider the top 5 users with the highest degree / betweenness / closeness centrality values (the 5 users may be different for each of the three measures). For each of these users, write a script to find the distribution of length of his posts.
5. Repeat 4) for the 5 users yielding the least centrality values. Comment whether the post length is a discriminating factor.
6. Now we shall focus on the timestamp of the posts, taking into account the date of the earliest and latest post in the dataset. Given that individual post can be commented by other users at different timestamp, we can set the age or lifetime duration of each individual post. Suggest a script that allows you to draw the distribution the ages of posts.
7. Summarize in a table the average age and standard deviation of posts associated to the five top scoring users in 4) and 5).
8. Use label propagation implementation in NetworkX to determine the various communities of the network. List the various communities outputted by the algorithm, summarizing in a table the main characteristics of each community in terms of nodes, edges, clustering coefficient and diameter.
9. Calculate the PangeRank centrality of each node of the network using appropriate function of NetworkX. Plot the PageRank centrality distribution and check whether a power law fit can be established.
10. We want to evaluate the impact of the outputted communities. Write a script that calculates the total of PageRank centrality of all nodes of the same community. Rank the communities according to their PageRank score.
11. Identify relevant literature to comment on the main finding of your results.

Project 15: Analysis of Smoking Cessation

The project aims to study the online community of smoking cessation users in Twitter social network.

1. Identify two hashtags related to smoked cessation. Examples include #Stopsmoking, #smokefree, #Stoptabac, etc.. Use some statistics from literature to motivate your choice.
2. Collect a sufficient number of tweets related to each hashtag (few thousands tweets). For this purpose, you can use for instance Tweepy (see tutorial at https://riptutorial.com/tweepy), also see examples of text processing in Python in NLTK online book at https://www.nltk.org/book/. You would need to create your Twitter API account credential. See an example of program of collecting tweets for a given hashtag in https://www.promptcloud.com/blog/scrape-twitter-data-using-python-r/. You should ensure the data is large enough to ensure there are connections among a large number of tweet users and the existence of users who have several tweets. Try to repeat the process using more than one Twitter account, if the data is not sufficient, to ensure satisfactory of the collected data. Save the attributes of the tweets for each hashtag in a csv database. This includes the Twitter ID, tweet message, list of followers of the tweet user, whether it is a retweet or not, location, if available.
3. Identify the top ten influencers in each hashtag in terms of number of tweets generated. Determine the average and standard deviation of number of tweets per user.
4. By taking into account the timestamp of the tweet, draw the histogram of the number of tweets of each hashtag per time interval.
5. Identify from the Twitter accounts instances that likely corresponds to pharmaceutical products and adverts. Suggest how you can use NetworkX functions to identify such Twitter accounts.
6. We want to investigate the existence of other (sub) hashtags in each of the two collected Twitter dataset in 1). Use simple string search in tweet messages to identify the existence of such sub-hashtags.. Report the results in a table summarizing the proportion of the sub hashtags for each case.
7. Write down small program that allows you to identify the “mention” in each tweet message (word precedent by “@”). Now, construct a new network graph where the nodes correspond to the user ID while the edge between two nodes indicates that these two nodes share the same mention. Use appropriate NetworkX functions to determine global properties of the network in terms of number of nodes, edges, diameter and clustering coefficient. Put the result in a table.
8. Now we want to explore the hashtag content of the tweet message. Write a script that allows you to determine the various hashtags in each tweet message. Then consider the network constructed when nodes correspond to hashtags and edge between two nodes indicates that these nodes (hashtags) are contained in the same tweet message. Provide a short summary of the property of this graph in terms of number of nodes, number of edges, number of components, diameter, average/standard deviation degree and global clustering coefficient using appropriate NetworkX functions. Write down the results in a table.

You may see examples for network construction in https://github.com/computermacgyver/twitter-python

9. Use three methods of your choice (already implemented in Networkx) of community discovery in order to determine the online communities of the network in 8). Provide in a table the summary of these communities in terms of number of nodes / edges, diameter and clustering coefficient.
10. Comment the results obtained and summarize the key finding regarding behavior of users with respect to smoking cessation. Seek some literature to reinforce your interpretation.

Project 16: Twitter Dataset Analysis

Consider the Lerman Twitter 2010 Dataset available at
https://www.isi.edu/~lerman/downloads/twitter/twitter2010.html
Twitter_2010 data set contains tweets containing URLs that have been posted on Twitter during October 2010. In addition to tweets, the dataset contains the followee links of tweeting users, allowing the reconstruction of the follower graph of active (tweeting) users. URLs 66,059 tweets 2,859,764 users 736,930 links 36,743,448 Tweets.

1. Download the dataset in excel format and use the followee link table in order to study the social network graph properties of the underlined graph. Use NetworkX in order to report in a table the overall properties of the network: number of nodes, number of edges, average degree centrality and its variance, average In-betweeness centrality and its variance, average closeness centrality and its variance, average page-rank centrality and its variance, average clustering coefficient and average shortest path length and its variance, and size of its giant component. Plot the corresponding graph using appropriate NetworkX functions.
2. Draw heat map of degree centrality distribution. [see examples of heat maps in https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0220061
3. Consider the number of tweets per users ID. Trace the distribution of number of Tweets per user. Check whether this distribution can be fit to a power law distribution.
4. Repeat the preceding by considering the number of followers per user ID.
5. Repeat the preceding by considering the number of followees/friends.
6. Now we want to focus on the most active Twitter IDs. Trace the histogram (in terms of number of tweets generated) of the 30 most active Twitter users.
7. Now we want to find out whether some active users were genuine users or not. For this purpose, study the program botometer available in https://github.com/IUNetSci/botometer-python. The program inputs a tweet user id and outputs the probability that the user id is a bot or human. You can use a threshold 0.5 beyond which a program is bot or not. Therefore, consider the following hierarchical construction. Find out the proportion of bots among the top 5 active users. Repeat this process for the top 10 active users, top15, top20, to 25 and top 30. Report the result in a table.
8. Use appropriate NetworkX functions in order to identify various communities present in the graph using edge-betweeness evaluation (using Girvan-Newman algorithm for instance) and evaluate the quality of these communities. Draw the various communities using distinct colors.
9. Now we want to test whether community members share the same sentiment. For this purpose, use VADER tool (https://github.com/cjhutto/vaderSentiment), which output sentiment in terms of POSITIVE, NEGATIVE and NEUTRAL to determine the sentiment of each tweet message. Write a program that input tweet messages of each node (user ID) of a given community, and outputs the overall sentiment. Represent the sentiment of each community as the proportion of positive, negative and neutral it does contain. Trace the histogram (in terms of proportion of positive, negative and neutral) of each community.
10. Now we want to make use of the timestamp information available in the Tweet attributes. Provide a histogram with 20 bins which shows the number of tweet occurrences at each bin.
11. Suggest your own approach to test the speed of links provided in tweets messages using the above timestamp information
12. Use appropriate literature to support your finding and discuss your results

Project 17. Analysis of Climate Change Community.

The project aims to investigate the diffusion process of Climate change topic.

Use Twitter API to collect at least 2000 tweets related to hashtags *#globalwarming, #climatechange, #agw* (an acronym for "anthropogenic global warming"), *#climate*and*#climaterealists*, #climatestrikeonline, but feel free to suggest any other climate change hashtags of your choice if deemed more popular. The key in the collection process is that there is important number of tweets that contain other hashtags as well. It is also important to leave the collection open to other non-English tweets as well to ensure large coverage.

1. Draw a histogram showing the popularity of the main hashtags highlighting the number of tweets per individual hashtag and in another graph the number of distinct Tweet users per individual hashtag.
2. Draw pie chart illustrations showing regional location of the tweets associated to each of the above main hashtags using the location attribute of the tweet (whenever available).
3. Use other pie chart illustrations to show the language of the tweets for each of the above main hashtags.
4. Use VADER tool (https://github.com/cjhutto/vaderSentiment), which output sentiment in terms of POSITIVE, NEGATIVE and NEUTRAL to determine the sentiment of each tweet of the dataset. Then represent the distribution of each tweet as a point in the ternary plot.
5. We now want to build a social graph where each node corresponds to a hashtag and an edge between hashtag A and hashtag B indicates that there is at least one tweet which contains both hashtag A and hashtag B. Implement a small python program that allows you to identify hashtags in tweet messages and generate the above social network graph.
6. Summarize in a table the main global properties of the above graph: Number of nodes, Number of edges, average degree centrality, diameter, clustering coefficient, size of largest component using appropriate NetworkX functions. Comment on the obtained results highlighting any inherent limitation or characteristic of the data collection process.
7. Plot the degree distribution and local clustering coefficient distribution.
8. Use label propagation algorithm in NetworkX to find communities in the above network. Compare the size of the generated communities and their associated diameter and clustering coefficient in a table. Using your understanding of the name of the hashtag, speculate whether each community can be assigned some interpretation.
9. Now we want to comprehend the key players in the graph whether they are genuine or not. For this purpose, consider the 10 most ranked in terms of degree centrality, and scrutinize the tweets which are linked to those hashtags (10 most ranked). The scrutinizing consists in applying the botometer available in https://github.com/IUNetSci/botometer-python. The program inputs a tweet user id and outputs the probability that the user id is a bot or human. You can use a threshold 0.5 beyond which a program is bot or not. The purpose is therefore to test whether a given user id (Tweet user of the tweet that contains the hashtag) is a bot or not. Draw a relevant plot which shows the proportion of bots in the top ranked hashtag.
10. We would like to test the amount of support assigned to each hashtag. For this purpose use the information about number of retweets and number of replies of each tweet involved in the hashtag and suggest an expression that quantifies the amount of support.
11. Comment the results of the previous steps using some literature from climate change in order to reinforce your argumentation.

Project 18. War Twitter Analysis.

The project aims to investigate the diffusion process of Ukraine war topic using hashtags.

Use Twitter API to collect at least 5000 tweets related to hashtags #ukrainewar, #war #army #military #kiev #ua #specialforces #donbass #donbasswar #airsoft #nomockal #warukraine #tactics #azovsea #militarystile #azov #russia #donetsk #soldiers #ukrainenews #odessa #ukrainianarmy #lviv #victory #nato #kyiv #militaryukraine #news #freesentso, but feel free to suggest any other Ukraine war hashtags of your choice if deemed more popular. The key in the collection process is that there is important number of tweets that contain other hashtags as well. It is also important to leave the collection open to other non-English tweets as well to ensure large coverage.

1. Draw a histogram showing the popularity of the main hashtags highlighting the number of tweets per individual hashtag and in another graph the number of distinct Tweet users per individual hashtag.
2. Draw pie chart illustrations showing regional location of the tweets associated to each of the above main hashtags using the location attribute of the tweet (whenever available).
3. Use other pie chart illustrations to show the language of the tweets for each of the above main hashtags.
4. Use VADER tool (https://github.com/cjhutto/vaderSentiment), which output sentiment in terms of POSITIVE, NEGATIVE and NEUTRAL to determine the sentiment of each tweet of the dataset. Then represent the sentiment of each tweet as a point in the ternary plot.
5. We now want to build a social graph where each node corresponds to a hashtag and an edge between hashtag A and hashtag B indicates that there is at least one tweet which contains both hashtag A and hashtag B. Implement a python program that allows you to identify hashtags from Tweet content message and generate the above social network graph.
6. Summarize in a table the main global properties of the above graph: Number of nodes, Number of edges, average degree centrality, diameter, clustering coefficient, size of largest component using appropriate NetworkX functions. Comment on the obtained results highlighting any inherent limitation or characteristic of the data collection process.
7. Plot the degree distribution and local clustering coefficient distribution.
8. Use label propagation algorithm in NetworkX to find communities in the above network. Compare the size of the generated communities and their associated diameter and clustering coefficient in a table. Using your understanding of the name of the hashtag, speculate whether each community can be assigned some interpretation.
9. Now we want to comprehend the key players in the graph whether they are genuine or not. For this purpose, consider the 10 most ranked in terms of degree centrality, and scrutinize the tweets which are linked to those hashtags (10 most ranked). The scrutinizing consists in applying the botometer available in https://github.com/IUNetSci/botometer-python. The program inputs a tweet user id and outputs the probability that the user id is a bot or human. You can use a threshold 0.5 beyond which a program is bot or not. The purpose is therefore to test whether a given user id (Tweet user of the tweet that contains the hashtag) is a bot or not. Draw a relevant plot which shows the proportion of bots in the top ranked hashtag.
10. We would like to test the amount of support assigned to each hashtag. For this purpose use the information about number of retweets and number of replies of each tweet involved in the hashtag and suggest an expression that quantifies the amount of support.
11. Comment the results of the previous steps using some literature from security studies to reinforce your argumentation.

Project 19. War Facebook Analysis.

The project aims to investigate the diffusion process of Ukraine war topic using hashtags through Facebook.

Use Facebook Graph API (Graph API (facebook.com), need to apply for account) to collect at least 1000 posts related to hashtags #ukrainewar, #war #army #military #kiev #ua #specialforces #donbass #donbasswar #airsoft #nomockal #warukraine #tactics #azovsea #militarystile #azov #russia #donetsk #soldiers #ukrainenews #odessa #ukrainianarmy #lviv #victory #nato #kyiv #militaryukraine #news #freesentso, but feel free to suggest any other Ukraine war hashtags of your choice if deemed more popular. The key in the collection process is that there is important number of posts that contain other hashtags as well in their message content. It is also important to leave the collection open to other non-English language as well to ensure large coverage.

1. Draw a histogram showing the popularity of the main hashtags highlighting the number of posts per individual hashtag.
2. Use a pie chart illustrations to show the language of the posts for each of the above main hashtags.
3. Use VADER tool (https://github.com/cjhutto/vaderSentiment), which output sentiment in terms of POSITIVE, NEGATIVE and NEUTRAL to determine the sentiment of each post of the dataset. Then represent the sentiment of each tweet as a point in the ternary plot.
4. We now want to build a social graph where each node corresponds to a hashtag and an edge between hashtag A and hashtag B indicates that there is at least one post which contains both hashtag A and hashtag B. Implement a python program that allows you to identify hashtags from post content message and generate the above social network graph.
5. Summarize in a table the main global properties of the above graph: Number of nodes, Number of edges, average degree centrality, diameter, clustering coefficient, size of largest component using appropriate NetworkX functions. Comment on the obtained results highlighting any inherent limitation or characteristic of the data collection process.
6. Plot the degree distribution and local clustering coefficient distribution.
7. Check whether power law distribution can be fit to the degree distribution.
8. Use label propagation algorithm in NetworkX to find communities in the above network. Compare the size of the generated communities and their associated diameter and clustering coefficient in a table. Using your understanding of the name of the hashtag, speculate whether each community can be assigned some interpretation.
9. We would like to use the information about Likes rating of the posts. For this purpose, calculate the rating of each hashtag by averaging the ratings of all posts where this hashtag occurred. Draw the histogram showing the distribution of the hashtag ratings.
10. We want to correlate the information about the hashtag rating with centrality values of each node (hashtag). For this purpose, calculate Person correlation between degree centrality and hashtag rating.
11. Repeat 10) when using Eigenvalue centrality measure instead of degree centrality.
12. Repeat 10) when using PageRank centrality measures.
13. Comment the results of the previous steps using some literature to reinforce your argumentation.

Project 20: Mapping Parenthood Responsibility in Vauva.fi debate forum

This project aims to some parenthood aspects from Finnish specialized **Vauva.fi** forum.

For this purpose start with enquiry the forum with keywords related to first pregnancy (i.e., first child, first time pregnant, etc..). We would like to comprehend aspects related to disappointment caused by first time pregnancy.

For this purpose, we shall be interested to those threads which bear negative feeling about pregnancy. The collection should also concentrate on threads which contain large number of posts (replies), at least more than 10 replies. Try to either manually or automatically (using beatifulsoup, vauva API or any other scrapping software of your choice) to collect at least the 100 threads which contains the highest number of replies and also record the timestamp interval of each thread. We are also specifically interested to categories:

i) parents and family,
ii) society,
iii) health issues
iv) Social services
v) Finance and wealth

For each of the above categories set a number of keywords that you think are more or less associated to it. Although, the whole blogs is in Finnish, you can use automatic Google translate to comprehend the meaning for non-Finnish students.

1. We would like to test the strength of each category in the collected database. For this purpose, use a simple string matching of the keyword lists (of each category) in the whole collected database to find out the support of each category in the whole database. Trace a bar plot showing the proportion of each category.
2. Use the like support indicator in the database to find out the correlation between the string matching and the like support indicator. For instance you can sum up the like and hate indicators of each post where a string matching is found, and see whether the total sum of the like /dislike of posts associated to each category follows the same ordering as the string matching result.
3. We would like to construct a social network graph where the nodes are the collected threads. An edge between two nodes (threads) is established if these threads share at least two categories among the five categories above. In other words, the two threads contain keywords related to at least two specified categories. Use appropriate NetworkX function to plot the constructed network graph.
4. Summarize in a table the main global attributes of the constructed graph in terms of Number of nodes, Number of edges, Overall clustering coefficient, Average path length, Size of giant component, diameter, average degree centrality and its associated variance, Average In-Betweeness centrality and its variance, Average path length and its variance, clustering coefficient, and diameter.
5. Use spectral clustering algorithm implemented in NetworkX in order to identify the communities in the graph and use the coverage function in NetworkX to quantify the quality of this clustering.
6. Use the context of the dataset to identify plausible interpretations of each community. For instance, does each community share specific attributes of the dataset, or is impacted by the size of number of posts, etc..?
7. Repeat the graph construction when narrowing the constraints on the edge establishment such that an edge between two threads is established if they share 3 categories, another graph for 4 categories and another one for all five categories. Draw for each case, the underlined case, the underlined graph and provide for each case a table that summarizes the aforementioned global attributes of the graph.
8. Discuss and comment the results using appropriate literature in order to reinforce your interpretations.

Project 21. Water Research Citation Network Analysis

We would like to explore the citation analysis in "Water Resources Research" journal community – see https://agupubs.onlinelibrary.wiley.com/journal/19447973 –

1. Construct a small database containing the list of papers published in the journal in the last five years. You are free to suggest your own approach to collect the data (either automated through API, if any, semi-automated). The database should contain titles of the papers, author names and their affiliations, and keywords used in the journal metadata and list of references for each paper. For instance, Scopus API, SciFinder API provide you with hints to gather articles biometric data (although not required). Save the collected database in a csv format.
2. Perform a simple histogram construction that allows you to rank the authors in terms of number of publications in the journal and number of collaborators (co-authors).
3. Perform another histogram that allows you to rank the institutions that publish most papers in the journal. Next, perform another histogram that allows you to rank the keywords that are attached to the largest number of articles.
4. We would like to construct a graph using the paper titles and their reference list. We consider there is a link from paper 1 to paper 2 if in the list of references of paper 1 contains paper 2. Design a program that allows you to implement the above strategy. Provide in a table the global properties of this graph in terms of number of nodes / edges, diameter, clustering coefficient, number of components, average / standard deviation degree using appropriate functions in NetworkX.
5. Use the concept of hubs and authorities as implemented in hits algorithm of NetworkX to calculate the page rank of each paper. Provide a histogram ranking the hubs in the article database
6. We would like to explore the network of authors in similar way as Erdös construction. From 2), assume the authors who has the largest number of collaborators as a reference Erdös.
7. Design and implement a program that would allow you to calculate the newly established Erdös number of each author.
8. Visualize the graph of authors with Erdös number 1 and 2.
9. Discuss your findings in the light appropriate bibliometric literature. You can also use alternative citations (e.g., Google citations for the authors with the top Erdös number to see if there is any match).

10. We want to investigate the network of institutions. For this purpose, write down a code that allows you to focus on the affiliations only part, and construct a graph where the nodes represent the author's affiliation and an edge between two nodes indicate that the underlined two institutions co-authored the same paper. Provide global properties of this graph as in 4).

11. Plot the degree distribution of this graph and check whether a power-law distribution can be fit. Comment on your finding using relevant literature.

Project 22: Open to new suggestion

If you have a concise idea in graph analysis that you want to pursue for personal reasons, feel free to get in touch to discuss the details