Online Social Network Analysis (CS-579-01)
Project 1 – Social Media Data Analysis

Dhruv Dinesh Singh - A20541901
Kunal Nilesh Samant - A20541900

Project Objectives:
You will learn how to crawl social media data, consider privacy and data usage implications, process, model and analyze the data. You will write a detailed written report and give a short oral presentation summarizing your results.

Project Outline:
1. Data Collection
2. Data Visualization
3. Network Measures Calculation

1) Data Collection:

- **DESCRIPTION:**
  We scraped data from Rotten Tomatoes "Movies on Netflix 2024 (at home)" category using Apify. First, we need to create an apify account then in apify store we search for "Rotten Tomatoes Scraper" (https://apify.com/rado.ch/rotten-tomatoes-scraper). After that we paste the website URL we need to scrape: (https://www.rottentomatoes.com/browse/movies_at_home/affiliates:netflix?page=5). Giving the input of maximum 500 results we start the scraper. It took 8 hours to scrape 864 results. Finally, we extracted the data in csv format. The data attributes we collected are [aspect ratio, audience score, box office (gross usa), cast, director, distributor, genre, original language, producer, production co, rating, release date (streaming), release date (theaters), rerelease date (theaters), runtime, sound mix, synopsis, title, tomatometer, url, view the collection, & writer]. We cleaned the raw dataset which gives us columns [audience score, box office (gross usa), genre, original language, release date (theaters), runtime, title, tomatometer] and 263 cleaned data results which is useful for further data visualization.

- **CHALLENGES:**
  During the initial stages of our data collection process, we encountered several roadblocks that influenced our decision-making and ultimately led us to choose Rotten Tomatoes as our scraping target.

At first, we set choose Twitter (X) as a potential source of data due to its vast user base, popularity and real-time nature. However, upon further investigation, we discovered that Twitter's privacy policy prohibits the scraping of user data without explicit consent, making it inaccessible for our purposes. Additionally, Twitter had made changes to its API restricting its access and availability of certain data streams, which further obstruct our ability to collect the desired information.

Considering these challenges, we explored alternative platforms and took Reddit into consideration as a possible data source. However, Reddit's uncompromising regulations regarding scraping, particularly in relation to the timeline of data retrieval, created significant obstacles. Furthermore, access to Reddit's API required costly subscriptions, making it an unfeasible option for our project. In April 2023, Reddit announced new fees for its Data APIs, its set at $0.24 per 1,000 calls.

As an acceptable replacement for web scraping, we ultimately resorted to Rotten Tomatoes. Rotten Tomatoes offers a wealth of movie-related information, including user ratings, critic reviews, and detailed movie profiles. Moreover, Rotten Tomatoes does not have the same restrictions on data access as Twitter or Reddit, making it a more accessible and feasible choice for our data collection efforts. Even though we faced some issues like most of the categories listed were by “editorial.rottentomatoes” which we can’t scrape. We were able to effectively extract the required data features and gather them into a sizable dataset for analysis by utilizing Apify's capabilities.

- **PRIVACY POLICY:**

PERSONAL INFORMATION WE COLLECT AND PURPOSES:

- Information you provide to us:
  a. Contact and account registration information
  b. Identification information, and demographics, and interests
  c. Transactional
  d. User-generated content
  e. Research and feedback
  f. Biometric identifiers
  g. Health Data
  h. Audio and video
- Information we collect automatically from you and/or your device:
  a. Device information and identifiers
  b. Connection and Usage (including information collected through cookies)
  c. Geolocation
- Information we collect from third parties:

a. Information from public and commercial sources
b. Social Media Information
c. Third Party Partners in Connection with providing the NBCUniversal Services

- Additional information that we collect for business-to-business relationships only:
  a. Business contact information
  b. Transactional information
  c. Demographics
  d. Due diligence information
  e. Events Information
  f. Social Media and Lifestyle Information

- **If you cannot find these policies, please describe where you looked for them:**
- 1. ACCEPTANCE OF TERMS
  a. These Terms of Use set forth the terms and conditions that apply to your use of the Services. You agree that you have read, understand and agree to be legally bound by these Terms. By using the Services, you will be deemed to have agreed to these Terms of Use.
  b. If you do not agree to these Terms, you may not use the Services.
  c. Fandango may, modify these Terms, and such changes will be effective 30 days following either notification to you or our posting of the changes to the Services. Your continued access or use of the Services after we post changes to these Terms will be deemed acceptance of these Terms as modified. We encourage you to check back here for any such changes from time to time.
  d. Any form of transfer or sublicense, or unauthorized access, distribution, reproduction, copying, retransmission, publication, sale, or exploitation (commercial or otherwise), of any portion of the Services, including but not limited to all content, services, digital products, tools or products, is hereby expressly prohibited.
- 2. PERMITTED USE
  a. Our Site and Services are for your personal and non-commercial use. They contain material that is derived in whole or in part from material supplied and owned by Fandango and other sources. Such material is protected by copyright, trademark and other applicable laws. Unless otherwise agreed to in writing by Fandango, you agree that you will not use the Services, or duplicate, download, publish, modify or otherwise distribute or use any material in the Services for any purpose, except for your personal, non-

commercial use. You also agree that you will not link to any page on the Site other than the home page (for example, "deep linking"), without Fandango's prior written consent. Use of the Services or any materials or content on the Services for any commercial or other unauthorized purpose is prohibited. You acknowledge that storing, distributing or transmitting unlawful material could expose you to criminal and/or civil liability. You may not download (other than page caching) or modify the Services or any portion of them unless we have provided you with express written consent. You shall not make a derivative use of the Services (or any part thereof) for any purpose, nor shall you download or copy information of users, or otherwise engage in data mining or similar data gathering.

- 3. REGISTRATION, ACCOUNTS AND PASSWORDS
  a. If you establish a personal account with us, you agree to provide true and accurate data about yourself on our account registration form, and to update and keep such data current. You will receive a password and account upon completing the registration form. You are solely responsible for maintaining the confidentiality of your password and account, and you are solely responsible for all use of your password or account, whether authorized by you or not. You shall not allow other persons access to or use of such username or password. You shall not post your username or password on any website nor transmit it through unsecured sites. You agree to (a) immediately notify Fandango of any unauthorized use of your password or account or any other breach of security and (b) ensure that you exit from your account each time you use the Services. Access and use of password-protected and/or secure areas of the Services is restricted to users who have been given a valid password by Fandango. We may terminate your membership and access to the Services if we learn that you have provided us with false or misleading registration data. If we feel your username and password are insecure or otherwise problematic, we may require you to change it or terminate your account.

2) Data Visualization –

- **NETWORKX**

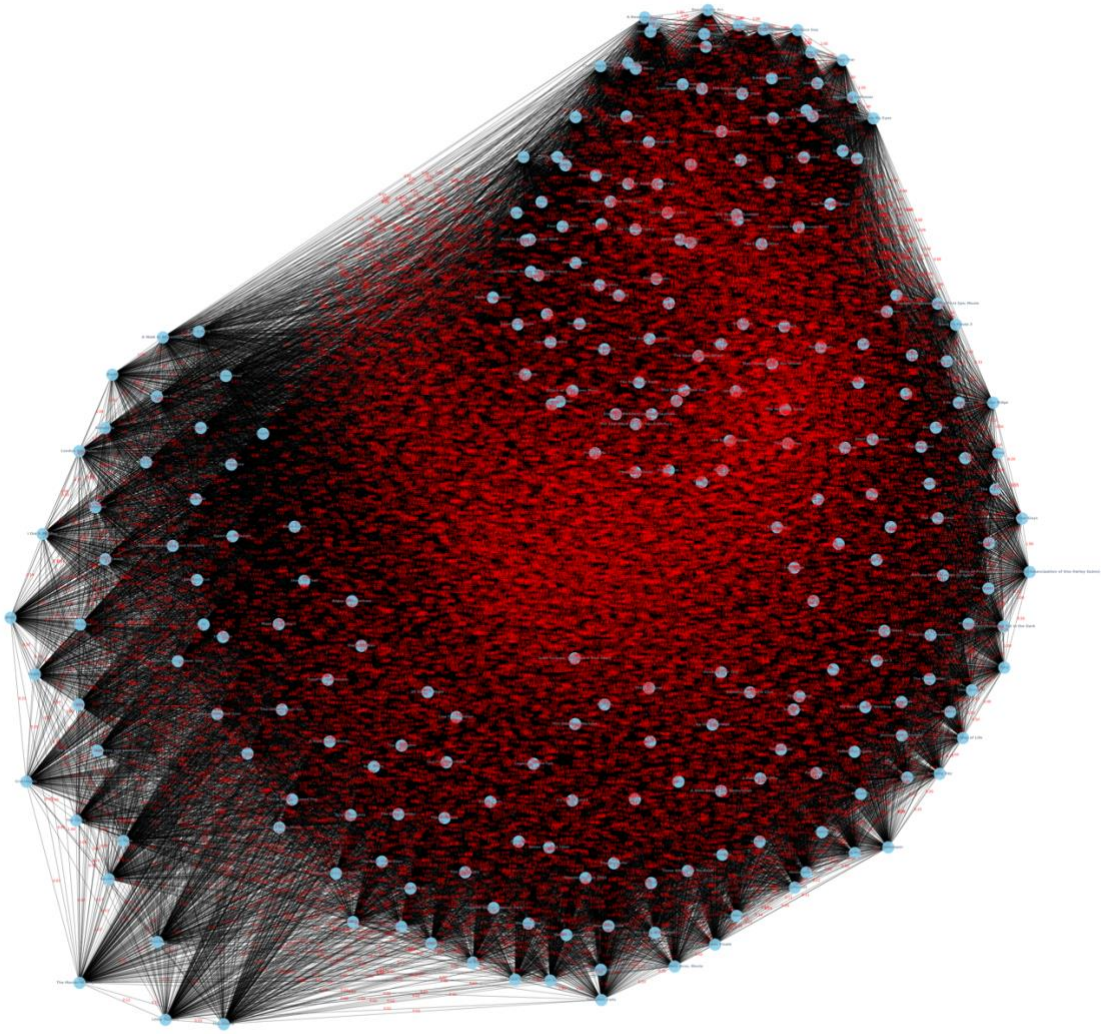
  NetworkX is a Python library designed for the creation, manipulation, and analysis of complex networks (or graphs). It provides an extensive set of tools and functions for working with graphs and their properties, making it a valuable tool for researchers, scientists, and engineers involved in network analysis and graph theory.

  Key features of NetworkX include:

1. Graph Representation: NetworkX supports the creation and manipulation of both directed and undirected graphs. It allows you to represent graphs using various data structures, such as adjacency matrices, adjacency lists, and more.
2. Node and Edge Attributes: Nodes and edges in a graph can have associated attributes. This flexibility allows you to attach additional information to elements in the graph, making it suitable for a wide range of applications.
3. Algorithms: NetworkX provides a rich collection of algorithms for tasks such as finding shortest paths, computing centrality measures, detecting communities, and more. These algorithms are implemented efficiently, making them suitable for analyzing large networks.
4. Graph Generation: NetworkX includes functions to generate various types of graphs, including classic graph structures (e.g., complete graphs, cycles) and random graphs (e.g., Erdős-Rényi graphs, Barabási-Albert graphs).
5. Visualization: The library allows for the visualization of graphs using matplotlib, a popular Python plotting library. This makes it easy to create visual representations of complex networks, aiding in the interpretation of results.

- **Short Description of the graph analysis software that you used, your reasoning for choosing the software and the format of the data input file.**
  Graph Information:
  The graph nodes are the Titles of the movies and the edges represent the relationship between them. The relationship between each node is based on the Tomatometer score each movie has received. The Higher the rating of the movie the less weight is on the edge. If a movie has a bad rating, then the weight of the edge between a highly rated movie would be greater. The graph is a collection of 264 nodes and numerous edges that connect each node with each other. The weight of the edges is in red color and the size of the graph is humungous. The code that generates the graph has a figure size of 50 by 50 which has the same graph but it is crowded but a commented code right underneath the plt.figure(figsize=(50, 50)) has a size of 250 by 250 and to get an output of the graph you need a heavy working system which can handle the load of creating a 25000 × 25000 PNG image. The 25000 × 25000 PNG image of the network is included in the zip file. The top of the graph shows nodes with fewer ratings and the bottom shows the nodes with the highest rating.

- **screenshot of our visualized graph along with any information needed for the reader to understand the visualization.**
  - High quality png file is attached in the zip file.



3) Network Measures –

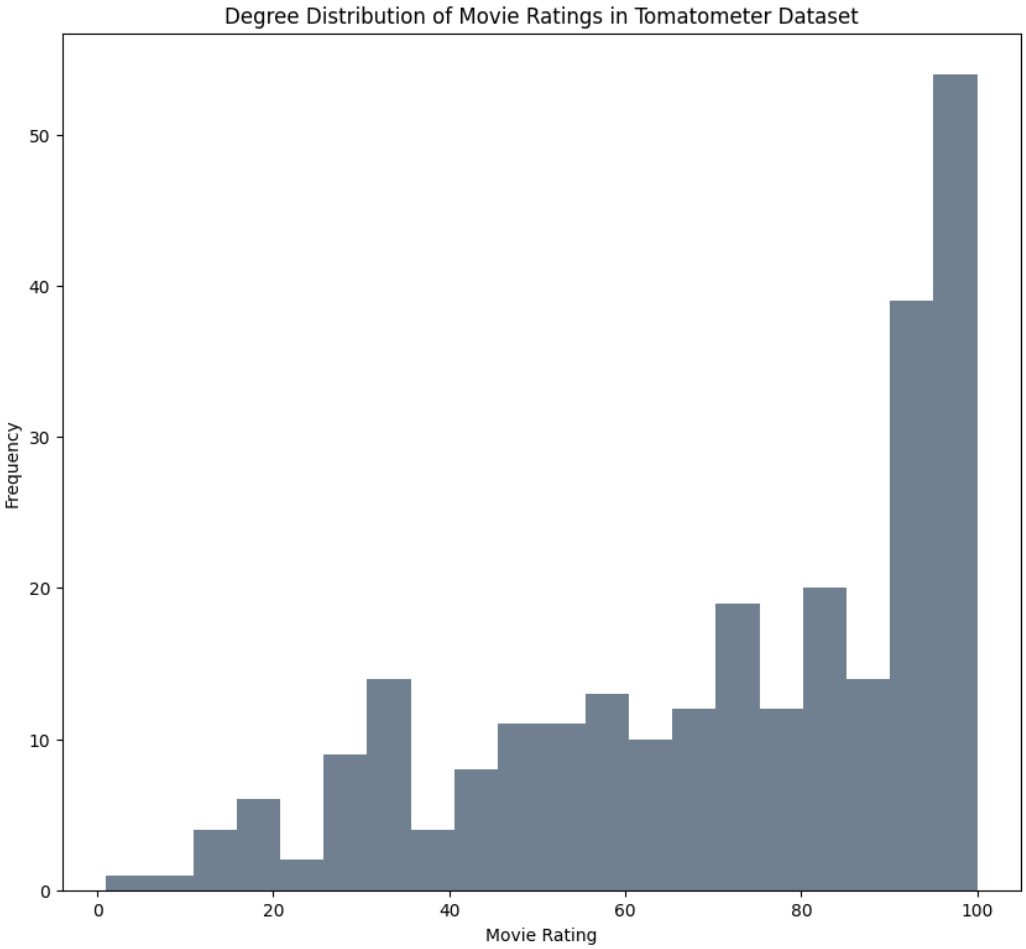
- As a key tool in our project, we used PageRank to assess the significance and relevancy of web pages inside a network. PageRank thinks a webpage is important if other important webpages link to it. It begins by giving each webpage a starting score called as PageRank score to show how important it is in the network. Then, it keeps adjusting these scores in each round, considering the links from other webpages. This goes on until the scores settle down and stop changing much.
- PageRank looks at how many links come to a webpage and where they come from. If a webpage gets lots of links from respected or good sources, its

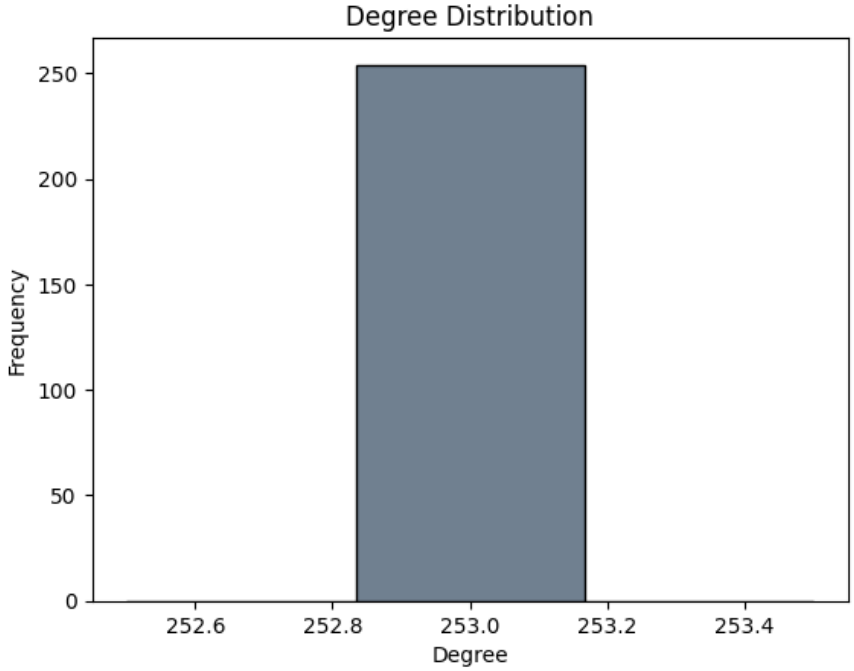
PageRank score will be higher compared to one with fewer links or links from less reputable sources.

- PageRank can also help to find and reduce spam or fake reviews by spotting manipulation patterns. If users with artificially boosted PageRank scores gained through cheating, the system could highlight or give less importance to those reviews in its calculations.

a. Node - Household Saints: PageRank Score = 0.00320551999352305
b. Node - Fallen Leaves: PageRank Score = 0.005218707614211272
c. Node - I.S.S.: PageRank Score = 0.002963692470657998
d. Node - Tótem: PageRank Score = 0.005218707614211272
e. Node - Four Daughters: PageRank Score = 0.0054826306374731815
f. Node - Freud's Last Session: PageRank Score = 0.0028109387859202705
g. Node - I Did It My Way: PageRank Score = 0.002184218719351957
h. Node - Monster: PageRank Score = 0.005373972918116932
i. Node - The Crime Is Mine: PageRank Score = 0.004942481478146552
j. Node - Aquaman and the Lost Kingdom: PageRank Score = 0.002953248400662138

- Use your chosen graph analysis software to obtain degree distribution and plot it as a histogram. In addition to this, choose two other network measures to report on. Choose any two from those that we've learned about. Report on these measures in an appropriate format.
- Your report will include a description of how you used the graph analysis software to get each of the three measures along with the measures and corresponding visualizations as appropriate.


Degree Distribution of Movie Ratings in Tomatometer Dataset
Frequency
Movie Rating
0
10
20
30
40
50
0
20
40
60
80
100


Degree Distribution
Frequency
Degree
0
50
100
150
200
250
252.6
252.8
253.0
253.2
253.4

4) **Discussion of Results –**

- **Your report will include a discussion of the results of the data visualization and network measures.**

  The final network graph shows the relationship between highly rated movies and low scored movies. The low scored movies are placed higher in the graph and the high rated movies are placed lower in the graph , Each Node is connected via an edge which has a weight assigned to it. If 2 movies have similar score it will have less weight but if they have a difference of for example 100 to 1 then the weight of the edge would be 1.

  1. Average Clustering Coefficient: The average clustering coefficient of the network is 1.0, indicating a fully connected graph where all nodes form closed triangles.

  2. Degree Distribution: The degree distribution of the graph is mentioned as 253, which refers to the total number of nodes or the sum of degrees across all nodes.

  3. PageRank Scores: PageRank is a measure of the importance of nodes in a network. The provided scores represent the PageRank for specific nodes (movies) in the network. Higher PageRank scores generally indicate more influential nodes.

  - Example nodes and their PageRank scores:
    - "Fallen Leaves": 0.00522
    - "Tótem": 0.00522
    - "Four Daughters": 0.00548
    - "The Zone of Interest": 0.00520
    - "Turning Red": 0.00548
    - ...

  These PageRank scores can be interpreted as the relative importance or influence of each movie in the network, based on the connectivity and relationships within the graph. Higher scores suggest movies that are more central or well-connected within the network.

- **What insights do these results provide?**

  The results provide us with the insights of a how each movies is different from each other. This graph can be used further to analyze the likely hood of someone watching a movie based on the score it receives and also if a user is confused about what movie to watch they can use the weight of the edge to decide what movie would be the better choice to watch.

- **What further questions do these results raise?**
  Questions like can the analysis of the network be expanded if the collection of out movies expand. Also, can the analysis method be improved to make the network graph more efficient.

- **What would your next step to investigate further be?**
  To investigate more further the movies can be divided based on the genre, the year of release, the director and then a graph can be developed to show the relationship between them.

5) Reference –

- Your report will cite all tutorials, packages, software and libraries you used in your data collection and analysis:
- For data collection and analysis, we used various tools and software to make things easier. Here's a list of software, packages, and libraries we used:
  a. Rotten Tomatoes: We scraped data from Rotten Tomatoes, a well-known website that collects reviews, to get details about movies on Netflix in 2024.
  b. Apify: We utilized Apify, a web scraping and automation platform, for data collection. Apify provided us with the necessary infrastructure and tools to scrape data from websites efficiently.
  c. Apify Store: Also, we used the Apify Store, which is like a shop for ready-made tools and solutions (Scraper), to find extra features and speed up our scraping tasks.
  d. Google Colab: We used Google Colab to run Python code and do data analysis tasks in the cloud. Google Colab gave us access to strong computing resources and tools for working together, allowing us to run our code and analyze big datasets smoothly in a cloud setting.
  e. Python Packages:
     - networkx: We used the networkx package for creating, manipulating, and analyzing complex networks (or graphs), particularly in our analysis of the web structure and link relationships.
     - pandas: The pandas package was under use for data manipulation and analysis. It allowed us to efficiently handle and process large datasets, including cleaning, filtering, and aggregating data.
     - matplotlib.pyplot: We utilized matplotlib.pyplot for data visualization. This package enabled us to create various plots and visualizations to illustrate our findings and insights effectively.

- glob: We used the glob package to handle and work with files. It helped us find file paths that matched certain patterns, making it easier to organize and handle our data files.

6) Video –

- The Video is attached in the zip file.