Network Science Project - Building Real World Applications

Hemalata Nayak

Supervisor: Dr. Alexander Nwala

Department of Applied Science
College of William and Mary

This report is submitted for the final project in "Network Science" course
PhD First Semester

April 18, 2014

Abstract

Table of contents

List of figures

List of tables

Chapter 1

Milestone - 1

1.1 Project Task

Our project aims to analyze the collaborative networks of renowned film directors, focusing on key collaborators and their roles. We will investigate patterns of collaboration, such as directors consistently working with the same individuals, and explore how these collaborations contribute to the creative process and the overall quality of the films. Additionally, we will examine how some directors prioritize diversity and inclusion by hiring collaborators from historically marginalized groups, such as African Americans and women, to broaden the perspectives and voices in the U.S. film industry. Through network analysis, we aim to gain insights into the dynamics of these collaborations and their impact on the industry's creative labor pool and representation.

It involves four steps as following:

- Data Extraction/cleaning
- Network Generation
- Network Visualization
- Analysis

1.2 Data Extraction

The project starts by extracting relevant information (all details of directors and their movies) from a CSV file (‘100 film directors.csv‘) containing data on directors and their movies. The snapshot of the CSV file is shown in Figure 1.1. The following list of directors includes the last name, first name, sex, ethnicity/race (A=Asian, Asian American (incl India), B=Black, I=Indigenous (Native American, Maori), L=Latin American, W=White), labels (H=top 25 highest grossing directors (excluding animation directors) and Q=identifies as LGBTQ), and IMDb URIs of 101 directors.

1	LastName	FirstName	Sex	Ethnicity_Race	Labels	IMDb_URI
2	Abrams	J.J.	M	W	H	https://www.imdb.com/name/nm0009190/
3	Allen	Woody	M	W		https://www.imdb.com/name/nm0000095/
4	Anderson	Paul Thomas	M	W		https://www.imdb.com/name/nm0000759/
5	Anderson	Wes	M	W		https://www.imdb.com/name/nm0027572/
6	Araki	Gregg	M	A	Q	https://www.imdb.com/name/nm0000777/

Fig. 1.1 The csv file which contains director and movie information

I have extracted all the IMDb URIs of each director from the csv file and stored it in a directory. The data are organized as shown in the Figure 1.2. The directory structure follows the format where "film-directors" is the main directory, and each sub-directory is named according to the unique director ID for each director. In each sub-directory, we will find a credit.json file containing all the details of the movies as shown in Figure 1.3 the director has worked on.

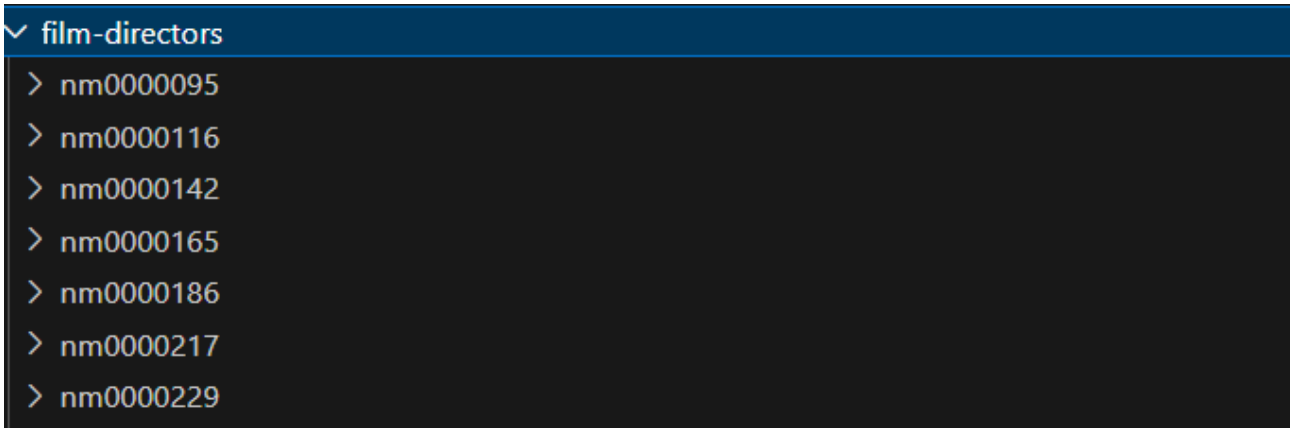

Fig. 1.2 The directory structure

I then scraped the movie URIs in the credit.json file for all directors, extracting the crew information for each movie and storing it in the respective directory.

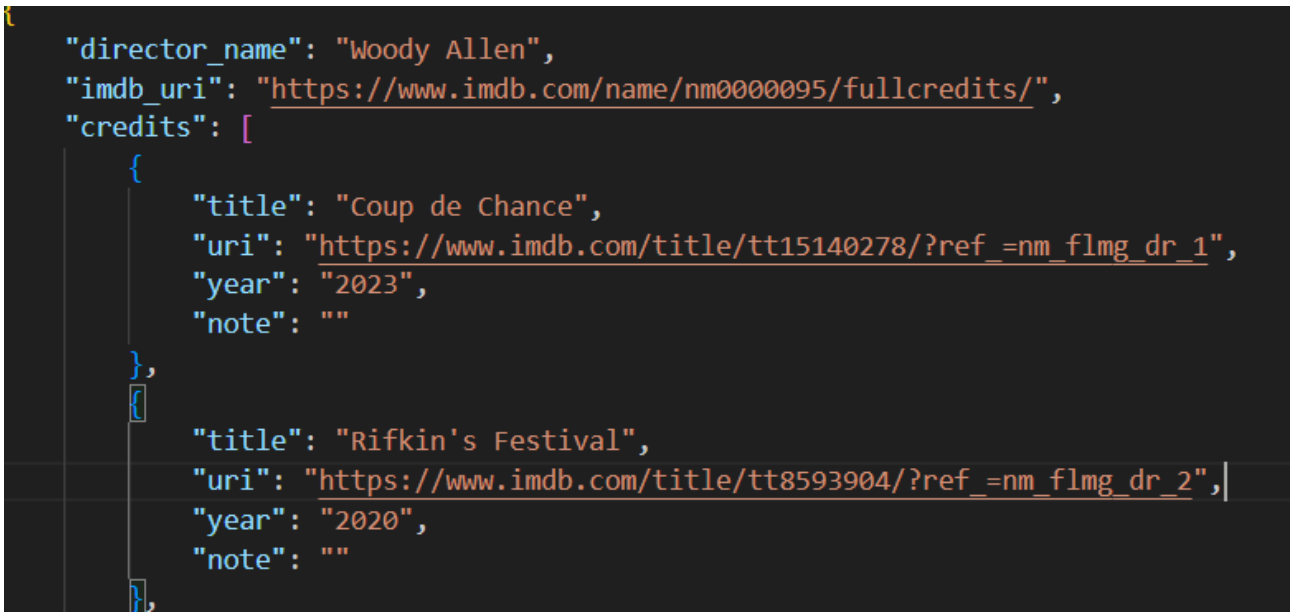

Fig. 1.3 The structure of the credit.json file

After extracting the crew information, the next step involved normalizing all variants of cast and writing credit to their root roles. This normalization process ensured consistency in how roles were represented across different movies. The normalized information was then stored in a directory named "Normalized_data" within each director's directory. The differences in file content before and after normalization are illustrated in Figure 1.4.

So far data extraction and cleaning is completed. The next steps involve analyzing the extracted data to create the collaborative network and visualize the network to find the relationships among directors and their collaborators.

(a) Crew details for a movie before normalization

(b) Crew details for a movie after normalization

Chapter 2

Milestone 2

In the previous section I had extracted the data and normalized it. But there were issues with downloading some files, hence they were empty. So I modified that code to handle such issues and re downloaded data and then I normalized roles and sub roles only for the movies which are longer than 70 mins and have saved it in the directory named "Normalized data".

2.1 distribution of entities in the director-crew dataset:

Then I addressed some of the research questions for sanity check. Let's see below.
Q1. Number of featured movies?
Ans: The total no. of featured movies is found to be 1383.
Q2. Number of directors, number movies per director, and average number of movies per director?
Ans: There are 101 directors. Average number of movies per director is: 13.693. The no. of movies per director is given in the table 2.1.
Q3.Number of crews?
Ans: The total number of crews is 8004.
Q4.Number of roles and frequency of each role?
Ans: There are total 17 roles.Out of which 8 are main roles and others are subroles. The frequency of each role is given the table 2.2.

The next step is generating the network and visualize it. And then I will address the research question.

Table 2.1 Number of movies per director ID

Director ID	Number of Movies	Director ID	Number of Movies	Director ID	Number of Movies
nm0000095	53	nm0000116	12	nm0000142	41
nm0000165	36	nm0000186	21	nm0000217	40
nm0000229	38	nm0000231	25	nm0000233	14
nm0000318	20	nm0000338	28	nm0000343	23
nm0000361	32	nm0000386	18	nm0000399	20
nm0000464	17	nm0000487	14	nm0000490	41
nm0000500	24	nm0000517	10	nm0000520	13
nm0000600	17	nm0000631	30	nm0000709	25
nm0000759	9	nm0000777	11	nm0000876	13
nm0000881	15	nm0000941	10	nm0001005	11
nm0001054	21	nm0001060	16	nm0001068	9
nm0001081	11	nm0001331	11	nm0001392	15
nm0001631	4	nm0001741	11	nm0001752	37
nm0001814	24	nm0002132	9	nm0004716	8
nm0005069	12	nm0009190	6	nm0027572	12
nm0036349	6	nm0037708	11	nm0122344	12
nm0138927	9	nm0160840	12	nm0169806	8
nm0190859	9	nm0200005	5	nm0269463	9
nm0281945	8	nm0298807	21	nm0327944	9
nm0336620	14	nm0336695	8	nm0362566	11
nm0366004	7	nm0392237	7	nm0420941	4
nm0426059	6	nm0476201	6	nm0501435	6
nm0510912	11	nm0570912	11	nm0583600	6
nm0590122	7	nm0619762	16	nm0634240	11
nm0668247	12	nm0697656	6	nm0716980	8
nm0751102	10	nm0751577	9	nm0796117	15
nm0853380	8	nm0868219	12	nm0893659	10
nm0898288	13	nm0905152	7	nm0905154	8
nm0911061	23	nm0946734	12	nm1119645	11
nm1148550	8	nm1218281	6	nm1347153	43
nm1443502	3	nm1490123	12	nm1503575	3
nm1560977	5	nm1716636	4	nm1802161	5
nm1883257	11	nm1950086	4	nm2011696	3
nm2125482	4	nm3363032	4		

Table 2.2 Number of movies per role/subrole

Role/Subrole	**Number of Movies**
Produced by - producer	3057
Produced by - producer produced by	592
Produced by - producer produced by pga	211
Writing Credits	3820
Sound Department - rerecording mixer	2113
Sound Department - sound designer	536
Sound Department - supervising sound editor	851
Directed by	2742
Cinematography by	1940
Casting By	1909
Music by	1523
Production Design by	1359
Costume Design by	1265
Makeup Department - hair department head	383
Makeup Department - makeup department head	461
Special Effects by - special effects supervisor	623
Special Effects by - visual effects supervisor	1