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 **iv**

List of tables **v**

1 Milestone - 1 **1**

1.1 Project Task 1

1.2 Data Extraction 1

List of figures

1.1 The csv file which contains director and movie information 2
1.2 The directory structure 2
1.3 The structure of the credit.json file 3
1.4 Comparison of crew details before and after normalization 4

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.

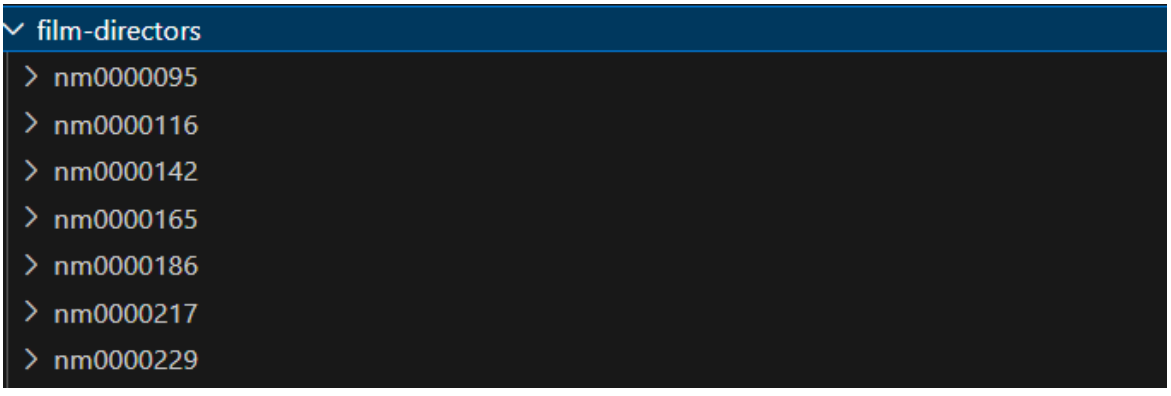

Fig. 1.2 The directory structure

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.

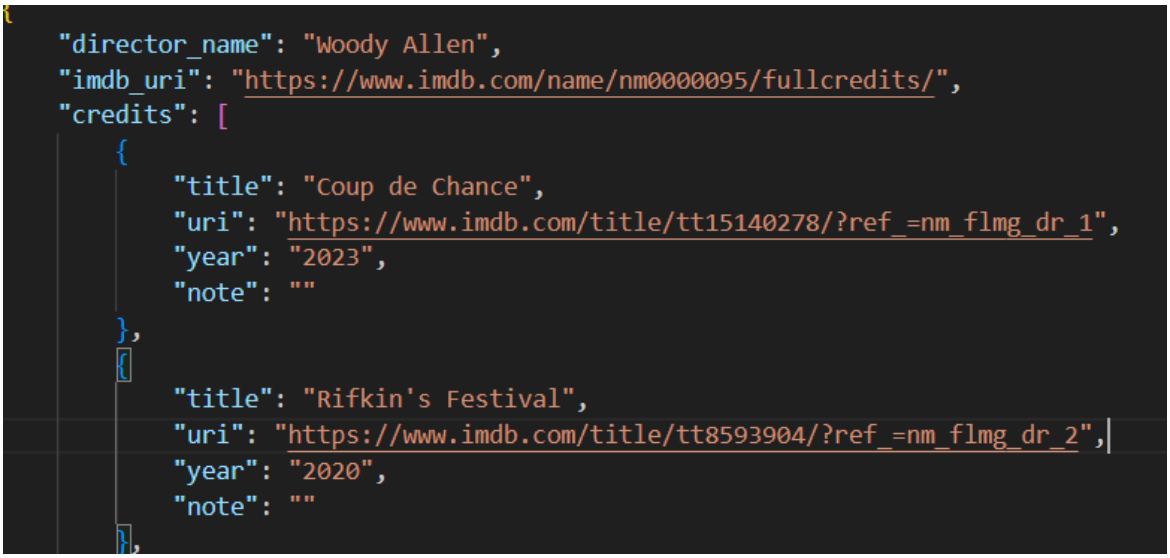

Fig. 1.3 The structure of the credit.json file

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

Fig. 1.4 Comparison of crew details before and after normalization