Data Collection Report

Rebeka Mukherjee, Archit Rathore, Yash Gangrade

1. How is the data obtained?

   The complete data is fetched from this Kaggle dataset: `https://www.kaggle.com/rounakbanik/the-movies-dataset/`. The dataset is downloadable after creating a Kaggle account and comes with a public domain creative commons license.

2. How large is your data?

   The dataset contains metadata for 45,000 movies. Data points include cast, crew, plot keywords, budget, revenue, posters, release dates, languages, production companies, countries, TMDB vote counts and vote averages. The compressed dataset is roughly 230 MBs.

3. In what format are you storing the dataset?

   We want to convert the data into a graph that can further lead to spectra based analysis methods. For this purpose we are going to first convert the user ratings into vectors and use the pairwise similarities as edge weights to induce an undirected graph over the vertices (which will be movies). We are therefore going to store

   (a) the rating vectors (as a matrix)

   (b) the pairwise similarities (again as a matrix)

   (c) the movie graph as an adjacency matrix

4. Did you need to process the original data to get it into an easier, more compressed format (e.g., convert from one format to another one)?

   No, the data is in csv format and thus easily readable using a large number of existing frameworks.

5. How would you simulate similar data?

   A (naive) generative model for a movie dataset can be described as following:

   (a) Sample a genre from a probability distribution over all possible genres.

   (b) Pick individuals (cast and staff like director, actors etc) from the conditional distribution P(individuals—genre)

   (c) Generate storyline by conditioning on the genre and director.

   This however is extremely naive, and there should be cross interactions between genres and people and storyline and so on, but that type of complicated model is probably impossibly to formulate.