Project 4: Movie Recommendation

By: Aaron Dzaboff

1.
Problem Formulation

What Are We Solving For?

Develop a recommendation system for Netflux based on similarity of users

Goal: Return a list of suggested movies based on similar users

2.

Data Mining

- **What sources were used**
- **How I collected the information**

What was Mined?

Sources: MovieLens Dataset

What was Collected?

- Collected information on over 62,423 movies
- Collected information on ratings of those movies. In total there are 25,000,095 ratings from 162,541 users between January 09, 1995 and November 21, 2019
- Data Collected Includes:
  - userId
  - movieId
  - rating
  - timestamp
  - gender
  - zip code
  - movie genre
  - movie title

3.
Data Preprocessing/Cleaning

- **Data came well put together so there was no cleaning/preprocessing done**

4.
Analysis

- **Methodology**
- **Query Examples**

Methodology

- Split the dataset into 80% training and 20% query sets.
- Defined a feature vector describing users
- Utilized LHS hashing to compute distances/similarities between users

Query Example #1: User 8

- Movies user 8 has rated highly:

Title:	Genre:	Rating:
Seven	Mystery/Thriller	5.0
Happy Gilmore	Comedy	5.0
Braveheart	Action/Drama	5.0
Crimson Tide	Drama/Thriller	5.0

- Suggested Movies based on similar users:

Movie:	Genre:
Speed	Action
Jackie Brown	Drama/Thriller
Clerks	Comedy
Disclosure	Drama/Thriller

Query Example #2: User 23

- Movies user 23 has rated highly:

Title:	Genre:	Rating:
Simple Twist of Fate	Drama	5.0
M	Crime/Thriller	5.0
A Man Without a Face	Drama	5.0
Fargo	Comedy/Drama	4.0
The Rock	Action/Thriller	4.0

- Suggested Movies based on similar users:

Movie:	Genre:
To Kill a Mockingbird	Drama
Good Will Hunting	Drama/Romance
Manon of the Spring	Drama
Gandhi	Drama
As Good As it Gets	Comedy/Drama

5.
Limitations

How Can This Project be Improved?

- Could have added more bins to improve search quality
- Can’t account for changes in user preferences
- Eccentric movies are innately hard to recommend because people either love or hate them
- Input data may not always be accurate as people are not perfect at providing ratings

Thanks!

Any questions?