Movie Recommendation Service with Apache Spark & Flask

A collaborative filtering approach using the MovieLens dataset

By: Ty Hak
Date: June 5, 2024

Building a Movie Recommendation Service with Apache Spark & Flask

Content: Movie Recommendation Service

URL:
https://www.codementor.io/@jadianes/building-a-recommender-with-apache-spark-python-example-app-part1-du1083qbw

Introduction

This project demonstrates the development of a movie recommendation service using Apache Spark and Flask. It employs collaborative filtering techniques to provide personalized movie suggestions based on user ratings, showcasing how machine learning algorithms can be effectively scaled using distributed computing.

Use Case: Building a movie recommendation system to suggest movies to users based on their preferences and past ratings.

Goal: Enhance user experience by providing personalized movie recommendations.

Dataset that we are using:

Small: 100,000 ratings, 3,600 tags, 9,000 movies, 600 users

Full: 27,000,000 ratings, 1,100,000 tags, 58,000 movies, 280,000 users

Technical Details: *Which aspect of Spark is applied*

- **SparkContext Configuration**: Configured for local mode to enable testing and development.
- **Data Loading and Parsing**: Utilized RDD (Resilient Distributed Dataset) to load, parse, and process data.
- **Collaborative Filtering**: Implemented Alternating Least Squares (ALS) algorithm for collaborative filtering to build the recommendation model.
- **Model Training and Evaluation**: Split the dataset into training, validation, and test sets to train and evaluate the model.
- **Performance Metrics**: Calculated Root Mean Squared Error (RMSE) to evaluate the accuracy of the recommendation model.

Debugging Details

Challenges included managing large datasets and optimizing ALS parameters. Modifications involved adjusting dataset splitting ratios and ensuring the system handled new user ratings efficiently. The key achievements included building a scalable recommendation system and achieving a lower RMSE, demonstrating model accuracy improvements. Efficient data handling and parameter tuning were critical coding practices.

Results

The ALS model was trained and validated, achieving an RMSE of 0.918 on the small dataset and 0.827 on the complete dataset. Personalized recommendations were successfully generated, demonstrating the system's ability to cater to user preferences effectively.

When filtering the full dataset to include only movies with at least 25 ratings, the recommendation system was able to provide a list of top movies that not only had high predicted ratings but also had a sufficient number of user ratings to ensure the recommendations were reliable. This filtering helps to remove obscure or less popular movies that might have high ratings from a few users but do not have widespread appeal.

When the filtering threshold is raised to include only movies with at least 100 ratings, the recommendation list tends to feature more well-known and widely-viewed movies. This ensures the recommendations are not only based on high ratings but also on a significant amount of user feedback, making the recommendations more robust and credible

Results for user 1

```
TOP recommended movies (with more than 25 reviews):
('DamNation (2014)', 4.279017974817113, 26)
('Mugabe and the White African (2009)', 4.1474569008214655, 27)
('It Felt Like a Kiss (2009)', 4.124566092862343, 38)
('"Lewis Black: Red', 4.107631241684867, 25)
("Long Night's Journey Into Day (2000)", 4.081662399481523, 36)
('Kizumonogatari II: Passionate Blood (2016)', 4.065885793181304, 71)
('"Last Lions', 4.061899769523185, 51)
('Mushishi: The Shadow That Devours the Sun (2014)', 4.038337697455734, 33)
('Ghost in the Shell Arise - Border 5: Pyrophoric Cult (2015)', 4.007951680519893, 55)
('Norm MacDonald: Me Doing Standup (2011)', 3.984245026827038, 36)
('Heimat - A Chronicle of Germany (Heimat - Eine deutsche Chronik) (1984)', 3.9823197405440744, 37)
('"Come Sweet Death (Komm', 3.945567354007892, 31)
('"Human Condition III', 3.9429487339711464, 145)
('The Web (2013)', 3.9315364495169955, 26)
("Workingman's Death (2005)", 3.8903782005728056, 26)
('The Spy Gone North (2018)', 3.872836084161504, 39)
('Last Ride (2009)', 3.8719793365592743, 25)
('Harakiri (Seppuku) (1962)', 3.871355412678703, 1282)
('Isle of Flowers (1989)', 3.869942074137409, 87)
('"Godfather', 3.869066087785331, 75004)
('The Garden of Sinners - Chapter 5: Paradox Paradigm (2008)', 3.86301676569105, 53)
('I Am So Proud of You (2008)', 3.8605984120110546, 99)
('Pulp Fiction (1994)', 3.8574109729321417, 108756)
('Prohibition (2011)', 3.8565652448095165, 47)
('1987: When the Day Comes (2017)', 3.8548453526012043, 27)
```

```
TOP recommended movies (with more than 100 reviews):
('"Human Condition III', 3.9429487339711464, 145)
('Harakiri (Seppuku) (1962)', 3.871355412678703, 1282)
('"Godfather', 3.869066087785331, 75004)
('Pulp Fiction (1994)', 3.8574109729321417, 108756)
('"Godfather: Part II', 3.8454910361765453, 47271)
('Seven Samurai (Shichinin no samurai) (1954)', 3.8426963737038378, 17120)
('Dr. Strangelove or: How I Learned to Stop Worrying and Love the Bomb (1964)', 3.824202206750776, 34324)
('Cosmos', 3.8109123694417195, 625)
('Yojimbo (1961)', 3.801929559118637, 5208)
('Come and See (Idi i smotri) (1985)', 3.801761734873521, 1501)
('Alone in the Wilderness (2004)', 3.7654654706814874, 422)
('Planet Earth II (2016)', 3.7652897620001937, 2041)
('Ikiru (1952)', 3.761725947092067, 2096)
('Apocalypse Now (1979)', 3.7615098552610586, 34020)
('High and Low (Tengoku to jigoku) (1963)', 3.760244243102754, 1259)
('Star Wars: Episode V - The Empire Strikes Back (1980)', 3.7539379074460335, 80200)
('Paths of Glory (1957)', 3.7462280031998283, 5604)
('"Lord of the Rings: The Fellowship of the Ring', 3.7414881975348067, 79940)
('"Decalogue', 3.7275103426140888, 673)
('Planet Earth (2006)', 3.720412416528591, 3015)
('Ran (1985)', 3.7169871711397455, 6528)
('Blade Runner (1982)', 3.7142912736296303, 47695)
('"Lord of the Rings: The Return of the King', 3.7118338833749185, 75512)
('Band of Brothers (2001)', 3.709614488642451, 2835)
('Rashomon (Rashômon) (1950)', 3.7077988714805663, 6198)
```

Result for user 2

```
TOP recommended movies (with more than 100 reviews):
('Boys Life 2 (1997)', 2.9518789206812457, 101)
('Junior and Karlson (1968)', 2.89748563740741, 108)
('Three from Prostokvashino (1978)', 2.852848159906679, 159)
('Karlson Returns (1970)', 2.844121300859129, 119)
('Formula of Love (1984)', 2.805795037751656, 106)
('Winter in Prostokvashino (1984)', 2.8057230179619745, 131)
('Anne Frank Remembered (1995)', 2.800809529010722, 1058)
('Nobody Loves Me (Keiner liebt mich) (1994)', 2.7982850339351604, 140)
('Boy Meets Girl (2015)', 2.7963783942737344, 101)
('Solas (1999)', 2.7800401898054226, 170)
('The Biggest Little Farm (2018)', 2.779439307713261, 121)
('That Munchhausen (1979)', 2.7583928021932698, 155)
('Ordinary Miracle (1978)', 2.7580537484463363, 152)
("Last Year's Snow Was Falling (1983)", 2.7548471266804118, 244)
('Return to Treasure Island (1988)', 2.753999903474514, 180)
('As it is in Heaven (Så som i himmelen) (2004)', 2.7500127306505195, 290)
('Winnie the Pooh Goes Visiting (1971)', 2.748025706149943, 193)
('Sense & Sensibility (2008)', 2.7465926193751824, 101)
('Prayers for Bobby (2009)', 2.7436421963872775, 147)
('"Chambermaid on the Titanic', 2.73665893887854, 138)
('Office Romance (1977)', 2.7348069586851516, 327)
('Hamlet (2009)', 2.7325480101167257, 101)
('Shirley Valentine (1989)', 2.729706545596928, 539)
('"Marcello Mastroianni: I Remember Yes', 2.7293348019034687, 151)
('"Wooden Man\'s Bride', 2.7248334813519044, 120)
```

```
TOP recommended movies (with more than 25 reviews):
("India's Daughter (2015)", 3.1900949119674493, 34)
('The Magic Ring (1982)', 3.1461716459918954, 27)
('The Mitten (1967)', 3.129949774051118, 29)
('Connections (1978)', 3.059156788369963, 57)
('Crazed Fruit (Kurutta kajitsu) (1956)', 3.0283982419995645, 27)
('The Adventures of Scamper the Penguin (1986)', 3.0233711488814023, 25)
('Thirst (Pyaasa) (1957)', 3.020262430221001, 37)
('Bobik Visiting Barbos (1977)', 3.0188672074759673, 55)
('Salut cousin! (1996)', 2.994001157195015, 35)
('A Plasticine Crow (1981)', 2.9895910057865875, 75)
('The Little World of Don Camillo (1952)', 2.987293687296024, 65)
('After Lucia (2012)', 2.9848709711036747, 40)
('Springsteen On Broadway (2018)', 2.9847495739722394, 28)
('Anne Of Green Gables: The Continuing Story (2000)', 2.983758547776909, 27)
('Rent: Filmed Live on Broadway (2008)', 2.9705508496413913, 41)
('Vovka in the Kingdom of Far Far Away (1965)', 2.9641320772671946, 88)
('Maiden (2019)', 2.9595482568651157, 51)
('Santitos (1999)', 2.9552245400145623, 60)
('Wolf and Calf (1984)', 2.9549452915755747, 32)
('Boys Life 2 (1997)', 2.9518789206812457, 101)
('Family Is Family (2018)', 2.9406263653906652, 26)
('Studio 54 (2018)', 2.936537802565491, 29)
('Dancemaker (1998)', 2.9320338850256267, 46)
('"Seventh Heaven (Septième ciel', 2.9262061124877405, 88)
('Border (1997)', 2.9116411747370083, 38)
```

Insight

This project highlights the efficacy of collaborative filtering and distributed computing in creating scalable recommendation systems. Business implications include enhanced user engagement through personalized recommendations, with potential applications in e-commerce and content streaming. Spark's distributed capabilities significantly improve processing efficiency, enabling the handling of large user bases and extensive datasets.

Quality vs. Quantity: Lowering the rating count threshold to 25 allows for discovering hidden gems that might not be widely known but are appreciated by a niche audience. However, raising the threshold to 100 prioritizes well-known, universally acclaimed films that have stood the test of time and user scrutiny.

Recommendation Reliability: Higher rating thresholds generally result in more reliable recommendations. Movies that have been rated by a large number of users are less likely to be anomalies and more likely to be genuinely appreciated.

Diverse Recommendations: The lower threshold (25 ratings) tends to surface more diverse and eclectic recommendations, which can be particularly useful for users looking to explore beyond mainstream options.

Mainstream Favorites: The higher threshold (100 ratings) tends to highlight mainstream favorites, ensuring that the recommendations are more aligned with widely recognized quality.

Conclusion

In both scenarios, Spark's collaborative filtering algorithm efficiently processes large datasets to provide meaningful and personalized movie recommendations, demonstrating its scalability and effectiveness in handling real-world data. These results illustrate the capability of the recommendation system to cater to different user preferences, whether they seek popular titles or niche films.