Haifa University

# Big Data Course

Homework 2

## Doing Analytics with Spark

**submission deadline 01.03.2026**

## 1. Abstract

Welcome to your Spark homework assignment! Spark is an open-source data processing engine that is designed for large-scale data processing. It is a popular choice for data scientists and engineers because it is fast, flexible, and easy to use. In this homework assignment, you will be working with Spark to perform various data processing tasks. You will have the opportunity to explore the capabilities of Spark and gain hands-on experience using it to solve real-world problems. Make sure you have a strong understanding of the material covered in the lesson before you begin, as you will be expected to apply that knowledge in this assignment.

After you have practices to implement MapReduce jobs and running them on your laptop, in this exercise you are asked to implement analytics with Spark on the provided files with content.

## 2. Instructions

The assignment has to be done in pairs, unless you received explicit permit from the lecturer. The submission should include:

a. A title page with your names and IDs
b. Implementation of the code
c. If needed instructions and explanations should be submitted within word document file
**d.** Finally, all homework artifacts must be submitted compressed with zip

**NOTE: Please do not submit JAR files**

**e.** Submission deadline is 01/03/2026 23:59

## 3. Project Description

The IMDB dataset is a comprehensive collection of information about movies and TV shows, including details such as titles, release years, genres, ratings, and more. This dataset is widely used for educational and analytical purposes, allowing users to explore trends and patterns in the entertainment industry. Below is a brief explanation of each column in the dataset:

- **title**: The name of the movie or TV show.
- **year**: The release year of the movie or show. For TV shows, it may include a range of years (e.g., "2015–2020").
- **certificate**: The certification rating assigned to the movie or show, indicating suitability for different audiences (e.g., PG-13, TV-MA).
- **duration**: The runtime of the movie or show, typically in minutes.
- **genre**: A list of genres associated with the title, separated by commas (e.g., Action, Drama).
- **rating**: The average user rating for the movie or show on a scale from 1 to 10.
- **description**: A short synopsis or summary of the movie or show's plot.
- **stars**: A list of the main actors or actresses featured in the title.
- **votes**: The total number of user votes the movie or show has received, formatted as a string with commas (e.g., "1,234").

This dataset will serve as the foundation for the tasks in this assignment, where you will apply Spark to clean, analyze, and extract insights from the data. The following tasks are designed to guide you through various aspects of Spark programming, from preprocessing and grouping to advanced analytics.

## 4. Tasks

1. **Data Cleaning and Preprocessing**
   - **Objective**: Prepare the dataset for analysis by handling missing values and ensuring proper formatting.
   - **Steps**:
     1. Identify and handle missing values by filtering or filling them appropriately.
     2. Convert the votes column to numeric by removing commas.
     3. Extract numeric values from the year column for consistency.
   - **Points**: **10**

2. **Top Rated Movies by Genre**

- **Objective**: Identify the top 10 movies for each genre based on ratings.
- **Steps**:
    1. Split the genre column into individual genres.
    2. Group movies by genre and sort them by ratings.
    3. Extract the top 10 movies for each genre.
- **Points**: **10**

3. **Actor Collaboration Network**
    - **Objective**: Analyze actor collaborations by creating a co-occurrence network.
    - **Steps**:
        1. Parse the stars column to extract individual actor names.
        2. Generate pairwise combinations of actors for each movie.
        3. Count the number of collaborations for each pair.
    - **Points**: **10**

4. **High-Rated Hidden Gems**
    - **Objective**: Identify movies with high ratings but relatively low votes.
    - **Steps**:
        1. Filter movies with ratings greater than 8.0 and votes fewer than 10,000.
        2. Sort the filtered results by rating and vote count.
    - **Points**: **10**

5. **Word Frequency in Movie Titles**
    - **Objective**: Find the most frequently used words in movie titles, excluding stop words.
    - **Steps**:
        1. Perform a word count on the title column.
        2. Exclude common stop words using a predefined list.
        3. Identify the top 20 most frequent words.
    - **Points**: **10**

6. **Genre Diversity in Ratings**
    - **Objective**: Measure the variability of ratings across genres.
    - **Steps**:

        1. Group movies by genre.
        2. Calculate the standard deviation of ratings within each genre.
        3. Identify the genres with the highest and lowest variability.
    - **Points**: **10**

7. **Certification Rating Distribution**
    - **Objective**: Analyze the distribution of movies by certification rating.
    - **Steps**:
        1. Count the number of movies for each certification type.
        2. Determine the certification type with the highest average rating.
    - **Points**: **10**

8. **Comparing TV Shows and Movies**
    - **Objective**: Compare TV shows and movies based on ratings and votes.
    - **Steps**:
        1. Separate the dataset into TV shows and movies using the year column.
        2. Calculate average ratings and votes for each category.
        3. Identify trends in popularity over time for both categories.
    - **Points**: **10**

9. **Certification Rating Distribution**
    - **Objective**: Analyze the distribution of movies across different certification ratings. Identify which certification has the highest average rating.
    - **Steps**:
        1. Group movies by certification.
        2. Count the number of movies in each certification category.
        3. Calculate the average rating for each certification type.
        4. Identify the certification with the highest average rating.
    - **Points**: **10**

10. **Comparing TV Shows and Movies**
    - **Objective**: Separate the dataset into movies and TV shows based on the year column and analyze their differences.
    - **Steps**:
        1. Identify entries in the year column that indicate TV shows (e.g., entries with a "-").

2. Compute average ratings and total votes for movies and TV shows.
3. Analyze trends in popularity (total votes) over time for both categories.

- **Points**: **10**

## 5. Grading Criteria

| Criteria | Points | Description |
|---|---|---|
| Accuracy | 30 | Implement tasks correctly and produce the expected results. |
| Code Efficiency | 20 | Write efficient Spark code using transformations and actions effectively. |
| Usage of Spark Concepts | 20 | Demonstrate a clear understanding of DAG, lazy evaluation, and optimization. |
| Insights and Analysis | 20 | Provide meaningful insights and analysis based on the results. |
| Documentation and Style | 10 | Submit clean, well-commented, and well-structured code and reports. |