

**Book Recommendation System**

**Authors:**

**Momina Farrukh**

**Meerab Jalil**

**Zainab Jahanzeb**

**Maleeha Asghar**

**Ayesha Tariq**

**Book Recommendation System Recommendation System as a Machine Learning Problem**

|  |
| --- |
| **Book Recommendation System** |

* **Brief Overview**
* **A Book Recommendation System is a data-driven application designed to suggest books to users based on their preferences, reading history, and behavior.**
* **It employs various data science and machine learning techniques to provide personalized book recommendations, enhancing the reading experience for users.**
  + **Main Features**
* **Total Number of Books** 
  + **11128 books**



**Focus**

* **The main focus of this recommendation system is developing a**
  + **Recommendation System which can automatically recommend 10 similar books based on a given book title.**

**Book Recommendation System**

* **Real-world problem**
  + **Similar Book Recommendation**
* **Treated as**
  + **Un-Supervised Machine Learning Problem**
* **Note** 
  + **Book Recommendation System Problem is treated as a**
    - **Content-based Filtering Problem because the**

**Recommendations are made based on the features of items and a profile of the user's preferences.**

* + **Goal**
  + **Learn an Input-Output Function** 
    - **i.e. Learn from Input to predict the Output.**

**Book Recommendation System – Task**

* **Given**
  + **Book Title (Represented as Set of Attributes)**
* **Task**
  + **Automatically recommend similar books**

**Book Recommendation System – Input and Output**

* **Input** 
  + **Title**
  + **Author**
  + **Average Rating**
* **Output**
  + **Similar books**

**Note**

* **In our books Dataset, a Book is represented with many Attributes.**
* **Dataset**
  + **Link:** [**books\_data.csv**](../AppData/Local/Temp/2ef0e011-5f42-4535-9a46-1a75a5ac28cf_Book%20Recommendation%20Model.zip.8cf/books_data.csv)

**Book Recommendation System** **– Attributes**

* **A Book is represented with the following Four Attributes**
* **Attribute 01 – Book Id**
* **Attribute 02 – Title**
* **Attribute 03 – Authors**
* **Attribute 04 – Average rating**

**Book Recommendation System – Summary (Input)**

* **The following Table summarizes the Input Attributes for Books Dataset**

|  |  |  |
| --- | --- | --- |
| **Attribute No.** | **Attribute Names** | **Data Types** |
| **1** | **Title** | **Text** |
| **2** | **Author** | **Text** |
| **3** | **Average Rating** | **Numerical** |

**Table 01: Attributes of Dataset**

|  |
| --- |
| **Steps – Treating Book Recommendation System as an un-Supervised Machine Learning Problem** **using Train-Test Split Approach** |

**Book Recommendation System Problem**

* **Task**
  + **Develop a Book Recommendation System to recommend the books according to the interest/preference of the reader.**
* **Input**
* **Three Attributes**

|  |
| --- |
| 1. **Title** 2. **Author** 3. **Average Rating** |

* **Treated as a**
  + **Un-Supervised Machine Learning Problem**
* **Goal**
  + **Learn an Input-Output Function**
    - **i.e. Learn from Input to predict the Output.**
  + **Book Recommendation System is treated as a**
    - **Content-based Filtering Problem because the**

**Recommendations are made based on the user's preferences.**

**Book Recommendation System Problem – Input and Output**

* **Input**
  + **Text (Authors,Title) /Numerical (Average rating)**
* **Output**
  + **Text**

**Project Focus**

|  |
| --- |
| **Book Recommendation System** |

**Steps – Treating Book recommendation Problem as a Content-based filtering Problem.**

* **I will follow the following steps to treat the Book Recommendation System as a Content-based filtering Problem.** 
  + **Step 1: Decide the Learning Settings**
  + **Step 2: Obtain Sample Data**
  + **Step 3: Understand and Pre-process Sample Data**
  + **Step 4: Represent Sample Data in Machine Understandable Format**
  + **Step 5: Select Suitable Machine Learning Algorithms**
  + **Step 6: Split Sample Data into Training Data and Testing Data**
  + **Step 7: Select Suitable Evaluation Measure(s)**
  + **Step 8: Execute First Two Phases of Machine Learning Cycle**
    - **Training Phase**
    - **Testing Phase**
  + **Step 9: Analyze Results**

|  |
| --- |
| **If (Results are Good)**  **Then**  **Move to the Next Step**  **Else**  **Go to Step 1** |

* + **Step 10: Execute 3rd and 4th Phases of Machine Learning Cycle**
    - **Application Phase**
    - **Feedback Phase**
  + **Step 11: Based on Feedback**
    - **Go to Step 1 and repeat all the Steps**

|  |
| --- |
| **Step 1: Decide the Learning Setting** |

**Step 1: Decide the Learning Setting**

* **I will treat the Book Recommendation System as a** 
  + **Un-Supervised Machine Learning Problem**
* **Since Output is based on user’s preference, it will be treated as a**
  + **Content-based filtering Problem**

|  |
| --- |
| **Step 2: Obtain Sample Data** |

**Step 2: Obtain Sample Data**

* **Since I am Treating Book Recommendation System as an un-Supervised Machine Learning Problem, I will need**
  + **Un-Annotated Data**
* **For more accurate learning, I need**
  1. **Large amount of Un-Annotated Data**
  2. **High-quality of Un-Annotated Data**
  3. **Balanced Data**
* **Note**
  + **For simplicity, I will use a toy Corpus / Dataset of 100 instances.**

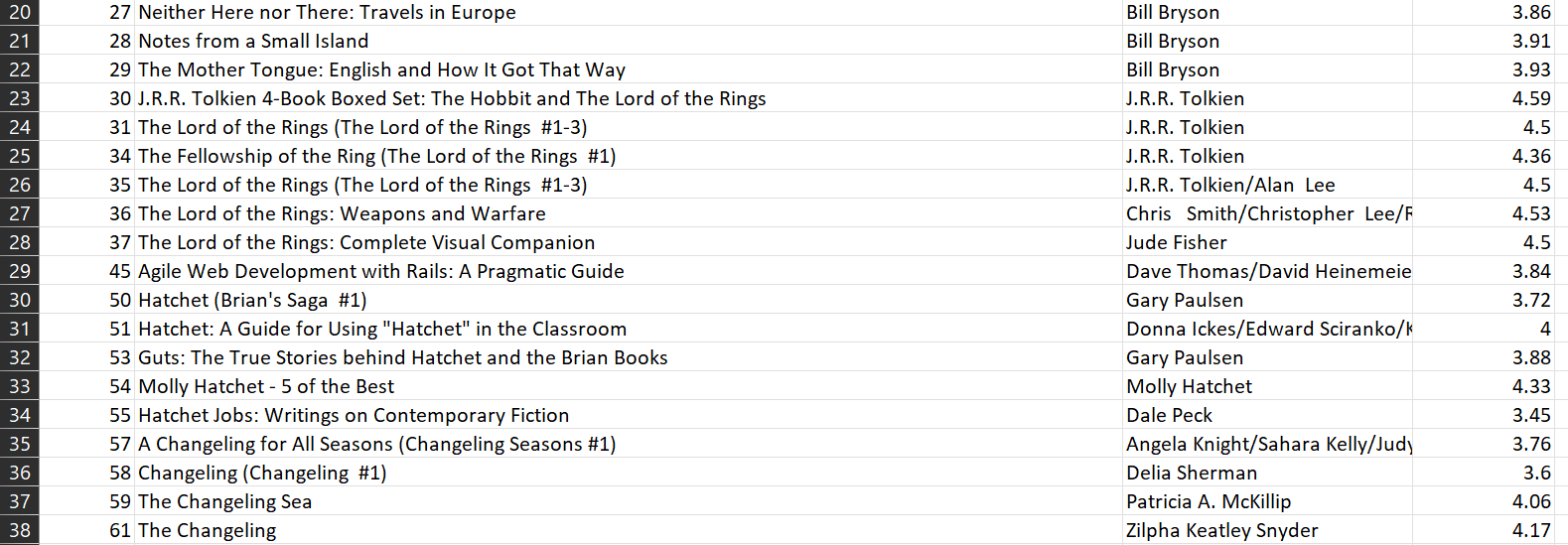
**Step 2: Obtain Sample Data Cont…**

* **Two Main Choices to Obtain Data**
  1. **Use an Existing Corpus**
  2. **Develop Your Corpus**
* **The Dataset to use is a subset of Books Dataset** 
  + **Corpus / Dataset** 
    - **Dataset Link:** [**books\_data.csv**](../AppData/Local/Temp/2ef0e011-5f42-4535-9a46-1a75a5ac28cf_Book%20Recommendation%20Model.zip.8cf/books_data.csv)
* **Obtain Sample Data Cont…**
* **Total Instances in Sample Data = 100**

**Sample Data**

* **We obtained a Sample Data of 100 instances.**
  + **See books\_data.csv File in Supporting Material**
* **The following Table shows the Sample Data**









A screenshot of a computer

Description automatically generated



|  |
| --- |
| **Step 03: Understand and Pre-process Data** |

**Step 3: Understand and Pre-process Sample Data**

* **Understanding Data**
  + **The Sample Data contains Four Attributes** 
    - **Book ID**
    - **Title**
    - **Author**
    - **Average Rating**
  + **Input**

**Input comprises of Three Attributes**

* + - **Title**
    - **Author**
    - **Average Rating**
* **Pre-processing Data**
  + **Corpus is already pre-processed.**
    - **Therefore, no pre-processing is needed 😊**

|  |
| --- |
| **Step 04: Represent Data in Machine Understandable Format** |

**Step 4: Represent Sample Data in Machine Understandable Format**

* **Feature-based Algorithms can understand data in** 
  + **Attribute-Value Pair** 
    - **Values of Attributes / Features must be Numeric.**
* **Problem**
  + **Our Sample Data is not in Attribute-Value Pair form.**
    - **We need to transform our Sample Data into Machine Understandable Format**
* **Solution**
  + **There are many approaches to transform Sample Data into Machine Understandable Format**

**Feature Extraction**

* **Features are already extracted.** 
  + **Therefore, we will skip the Feature Extraction Step 😊**

**Important Note**

* **In our content-based approach, the focus is on transforming textual features (title and author) into TF-IDF vectors and using cosine similarity for recommendations.**

**Transforming Sample Data in Machine Understandable Format**

* **In our Sample Data**
  + **Input is textual (title and authors) & numerical (average\_rating)**
* **Considering Input (Title, Author, Average Rating)** 
  + **Transform Input (Text) into TF-IDF vectors using cosine similarity.**

|  |
| --- |
| **Step 05: Select Suitable Machine Learning Algorithms** |

* **Previous students have shown that Good Starting Points for Content-based Filtering Problems are**
  + **TF-IDF with Cosine Similarity**
  + **Word Embeddings (e.g., Word2Vec or GloVe)**
  + **Doc2Vec**
  + **Feature Extraction with Neural Networks**
  + **Nearest Neighbors Algorithms (e.g., k-NN)**
  + **Term Frequency-Inverse Document Frequency (TF-IDF) with Other Similarity Metrics**

**Model Focus**

* **In this model, we will use.**

|  |
| --- |
| **TF-IDF with Cosine Similarity** |

|  |
| --- |
| **Step 06: Split Sample Data into Training Data and Testing Data** |

**Step 6: Split Sample Data into Training and Testing**

* **We Split the Sample Data using**
  + **Train-Test Split Ratio of**
    - **80% - 20%**
* **Training Data** 
  + **Total Instances = 8902**
* **Testing Data** 
  + **Total Instances = 2226**

**Training Data**

* **The following Table shows first 100 instances of the Training Data**
  + **See train\_books\_data.csv File in Supporting Material**



A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

**Testing Data**

* **The following Table shows first 100 instances the Testing Data**
  + **See test\_books\_data.csv File in Supporting Material**



A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A white background with black lines

Description automatically generated

|  |
| --- |
| **Step 07: Select Suitable Evaluation Measure(s)** |

* **I will use the Mean Reciprocal Rank Evaluation Measure to evaluate the performance of the Model.**
* **Mean Reciprocal Rank:**
* **Mean Reciprocal Rank (MRR) is a metric commonly used in information retrieval and recommendation systems to evaluate the effectiveness of a ranked list of items.**

|  |
| --- |
|  |

* **N is the total number of recommendation scenarios or queries.**
* **RRi is the Reciprocal Rank for the i-th scenario, indicating how well the system ranked the first relevant book in the list of recommended books for that scenario.**
* **MRR is the average of the Reciprocal Ranks across all scenarios, providing a summary measure of the system's overall performance.**
* **Note**

**Error Calculation:**

* + **Inverse of MRR (IMRR)**

**IMRR = 1 / observed MRR**

|  |
| --- |
| **Step 08: Execute First Two Phases of Machine Learning Cycle** |

**Step 8: Execute First Two Phases of Machine Learning Cycle**

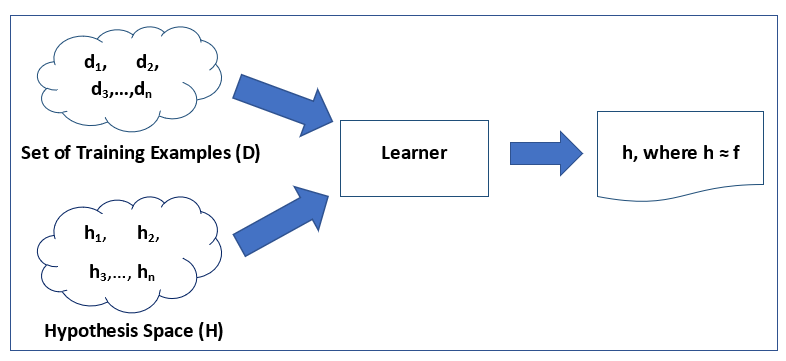
* **Recall the Equation**

|  |
| --- |
|  |

* **Training Phase**
  + **Use Training Data to build the Model.**
* **Testing Phase**
  + **Use Testing Data to evaluate the performance of the Model.**
* **Note that we aim to**
  + **Learn an Input-Output Function**

**General Settings - Learning Input-Output Function**

* **Recall – Our goal is to** 
  + **Learn an Input-Output Function**



**Training Phase**

**Training Phase**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **2** | **1** | **0** | **2** | **0** |
| **1** | **0** | **3** | **2** | **1** |
| **2** | **1** | **3** | **2** | **0** |
| **2** | **0** | **1** | **2** | **1** |
| **2** | **1** | **3** | **1** | **0** |
| **0** | **0** | **1** | **2** | **1** |
| **2** | **1** | **3** | **2** | **0** |
| **2** | **1** | **3** | **2** | **1** |
| **0** | **1** | **3** | **2** | **0** |
| **1** | **1** | **3** | **2** | **1** |

**Set of Training Examples ( D )**

**Hypothesis Space (H)**

**Learner**

**h, Where h ≈ f**

**Testing Phase**

* **Apply Model on the Testing Data**

**Testing Phase**

|  |  |  |  |
| --- | --- | --- | --- |
| **2** | **1** | **2** | **2** |
| **2** | **0** | **3** | **2** |
| **2** | **1** | **3** | **2** |
| **0** | **1** | **0** | **2** |
| **1** | **1** | **3** | **2** |

**Set of Testing Examples ( D )**

**Model (h)**

**Predictions**

|  |
| --- |
| **Step 09: Analyze Results** |

**Step 9: Analyze Results**

* **The assumption for this Example**
  + **Here, I am assuming that the Model.** 
    - **performed well on large Test Data and we can apply it in the real-world 😊**

|  |
| --- |
| **Step 10: Execute 3rd and 4th Phases of Machine Learning Cycle** |

**Step 10: Execute 3rd and 4th Phases of Machine Learning Cycle**

* **Application Phase**
  + **Model is deployed in Real-world to make predictions on Real-time Data**
* **Steps – Make Predictions on Real-time Data**
  + **Step 1: Take Input from User**
  + **Step 2: Convert User Input into Feature Vector** 
    - **The same as Feature Vectors of Sample Data**
  + **Step 3: Apply Model on the Feature Vector of the unseen instance**
  + **Step 4: Return Prediction to the User**

**Application Phase**

**Application Phase**

**Model (h)**

**Prediction**

|  |  |  |  |
| --- | --- | --- | --- |
| **0** | **0** | **2** | **0** |

**Feedback Phase**

* **A Two-Step Process**
* **Step 1: After some time, take Feedback from** 
  + **Domain Experts and Users on deployed Book Recommendation System.**
* **Step 2: Make a List of Possible Improvements based on Feedback received.**

|  |
| --- |
| **Step 11: Improve Book Recommendation System based on Feedback** |

**Step 11: Improve Book Recommendation System based on Feedback**

* **Go to Step 1 and improve the Book Recommendation System based on** 
  + **List of Possible Improvements made in Step 10**

|  |
| --- |
| **TODO and Your Turn** |

**TODO**

* **Task**
  + **Select a project similar to that one and apply all these steps for your project mentioned in the sample document.**
  + **Experiments and Presentation**
* **Import all the libraries that are required for the completion your project.**
* **Read the dataset and analyze it.**
* **Perform preprocessing on your dataset.**
* **Perform exploratory data analysis by visualizing your data.**
* **Perform label encoding on your data.**
* **Split your dataset in training and testing data.**
* **Apply at least one Machine Learning Classifiers**
* **Evaluate the performance of your classifiers.**
* **You have to make a presentation of 10 slides to present your semester project.**
  + **Hand in and Assessment**

**The hand-in consists of two parts: your code and presentation.**

* **Code: Place all files comprising your code into an archive file (a tar, jar or zip file) and submit this file before given deadline. In addition to your source code files(s) this archive should include (1) a README file which explains how to compile/run it/them; (2) the CSV file you have produced by running your program over the training data (3) an electronic copy of your presentation (in .pdf format).**
* **Presentation: Hand in a presentation of not more than 10 slides. in which you must discuss:**
* **Problem you are focusing in your semester project.**
* **Dataset(s) you are using for evaluation.**
* **Preprocessing Steps**
* **Data Visualization**
* **Evaluation measure(s) used.**
* **Summary of Results**
* **Any other thing(s) you feel important.**
  + **Assessment will be based primarily on the quality of your semester project, presentation and viva.**
  + **The due date of assignment submission will be announced in class. Note that there will be deduction of 5% marks (per day) on late submission. Departmental rules concerning plagiarism and collusion will be strictly observed – please refer to the Student Handbook for details of these.**