DATA MANAGEMENT PLAN ON RECOMMENDATION SYSTEMS FOR BOOKS

By Siddhant Singh & Anurag Sundar

Abstract

Recommendation Systems (RS) are an essential part of today's largest websites. Without them, it would be hard for users to find the right content and products. RS have evolved into a fundamental tool for helping users making informed decisions and choices, especially in the era of big data. Content based filtering is one of the most popular methods for recommendations. Its core relies on pre-processing textual data, in order to extract the maximum of relevant information from it. This project represents a pipeline to pre-process textual metadata, train and test a recommendation-based system for books.

Introduction

The goal of the project is to predict similar books and recommend them to the users using the data provided in the dataset which is stored in a SQL relational database. Recommendation Systems (RS) are engines that use algorithms leveraging the interaction between users to generate personalized recommendations. They provide users with recommendations for new content these users might be interested in (music, movies, books, etc). RS is used to predict user's affinity and recommend products/items that quite likely are interesting for them. RS can be divided into three main types: Collaborative Filtering (CF), Content-based Filtering (CBF) and Hybrid systems.

Collaborative filtering systems analyse users' interactions with the items (e.g. users ratings, likes or shares) to create recommendations. Content-based systems use semantic information (frequently called metadata) about the items in the system, and Hybrid Systems combines both methods.

The content also very often requires natural language processing (NLP) techniques to make use of semantic and syntactic characteristics. But most importantly, a recommender system should not be designed without taking into consideration the nature of the items and kind of the recommendation we are looking forward to building.

Dataset Information

The dataset is a file containing the detailed information about the books read by various users identified by a unique book id. The dataset represents a table of 12 attributes: bookID, title, authors, average_rating, isbn, isbn13, language_code, num_pages, ratings_count, text_reviews_count, publication_date & publisher.

Method

The focus here will be more on the pre-processing as it is the first most important step toward building a performant machine-learning system, whatever the task is. A good pre-processing, start with a good understanding both of the problem and of the dataset we have at hand. This ensures and creates a clear vision of the goal we are trying to achieve and the tools needed to do it. Data pre-processing includes cleaning, normalization, transformation, feature extraction and selection, etc. The product of data pre-processing is the final training set.

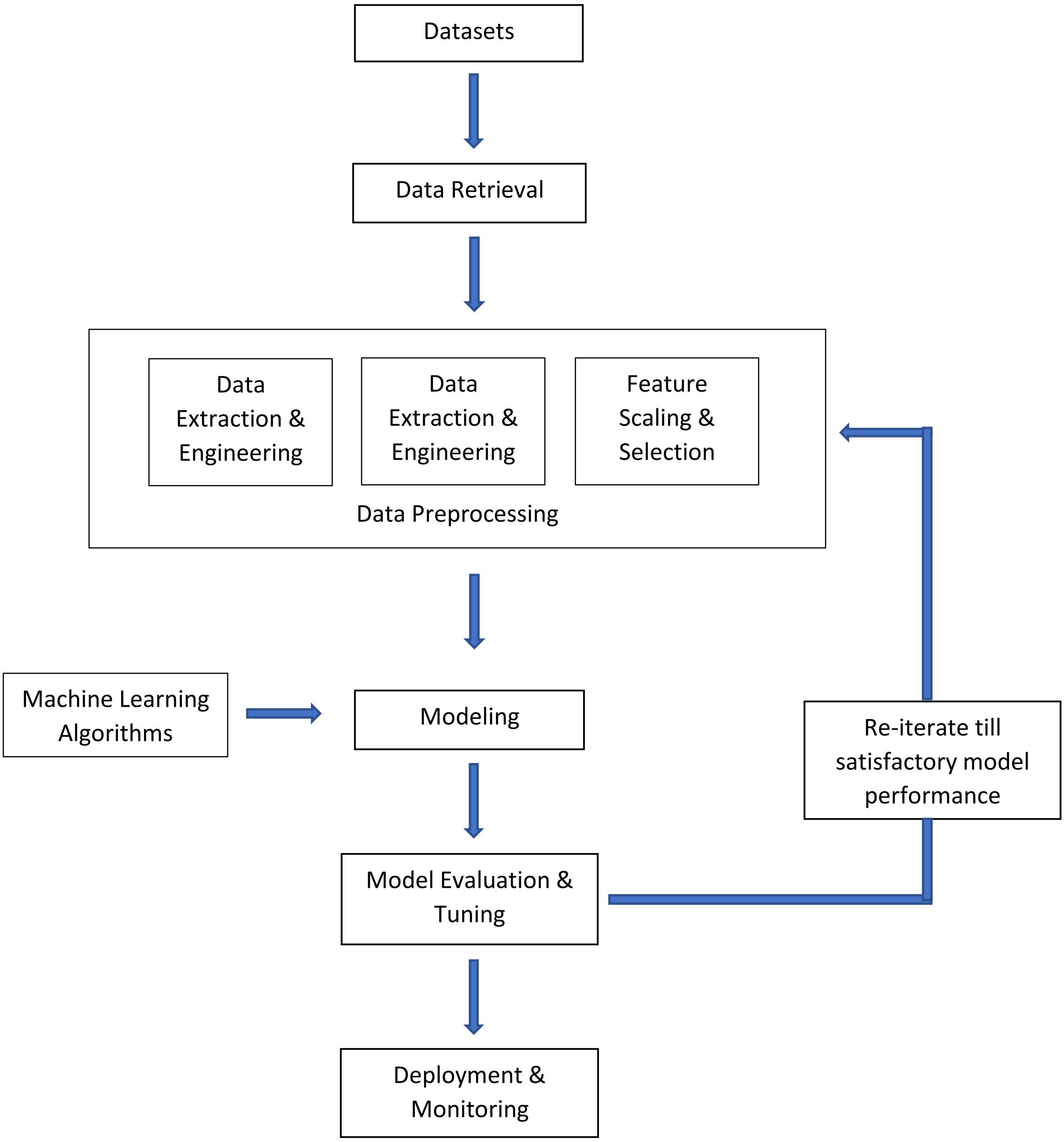
Datasets
Data Retrieval
Data Extraction & Engineering
Data Extraction & Engineering
Feature Scaling & Selection
Data Preprocessing
Machine Learning Algorithms
Modeling
Re-iterate till satisfactory model performance
Model Evaluation & Tuning
Deployment & Monitoring

Fig: Flowchart representing the RS project