# Books Recommendation System

**CMPE-256**
**Individual Project**
**Purva Deekshit**
**SID: 013749723**

**Motivation:**

The goal of this project is to recommend useful books, based on students/readers interests or requirements. This can help to save the time spent on searching relevant books from a huge books database.

**Data Collection:**

Since we refer multiple sources before purchasing a new book, this project also collects books metadata from multiple sites, such as Google Books API and Goodreads API.

**1. Books Data:**

**1.1 From Google API**:
I am collecting data from Google Books API using Google API Client Library for Python.
Google Books API's Documentation: https://developers.google.com/books/docs/overview

Google API metadata includes:

1. Various categories of books, such as: Business, Economics, Children, Computer Science, Mathematics, Literature, Education, etc.
2. Information about books: Title, Author, Category, Description, Language, Industry Identifier (ISBN), Website Link, Selling Info, etc.
3. Information about book ratings: Average rating, number of ratings for the book (rating count).

**1.2 From Goodreads API:**
I am also collecting average rating and rating count for books from Goodreads website.
Link: https://www.goodreads.com/

I am taking weighted average of ratings from both sites to get the final rating.

**2. User Data:**

I am sending query to Goodreads with the ISBN received from Google API to get a reviews link for book with that ISBN. This link provides the information of users who have rated the book. This link contains: ISBN, User ID, Username, User Rating and User Text review for the book.

## 1. Below is the sample of collected book metadata from Google API:

| | authors | averageRating | categories | description | industryIdentifiers | infoLink | language | maturityRating | pageCount |
|---|---|---|---|---|---|---|---|---|---|
| 0 | [Mark Lutz] | 3.0 | [Computers] | Get a comprehensive, in-depth introduction to ... | [{'type': 'ISBN_13', 'identifier': '9781449355... | https://play.google.com/store/books/details?id... | en | NOT_MATURE | 1600.0 |
| 1 | [Mark Lutz] | 4.0 | [Computers] | Google and YouTube use Python because it's hig... | [{'type': 'ISBN_13', 'identifier': '9781449379... | http://books.google.com/books?id=1HxWGezDZcgC&... | en | NOT_MATURE | 1216.0 |
| 2 | [Francois Chollet] | 5.0 | [Machine learning] | Summary Deep Learning with Python introduces t... | [{'type': 'ISBN_10', 'identifier': '1617294438... | http://books.google.com/books?id=Yo3CAQAACAAJ&... | en | NOT_MATURE | 384.0 |
| 3 | [Fabrizio Romano] | 5.0 | [Computers] | Learn to code like a professional with Python ... | [{'type': 'ISBN_13', 'identifier': '9781785284... | https://play.google.com/store/books/details?id... | en | NOT_MATURE | 442.0 |
| 4 | [Zed A. Shaw] | 5.0 | [Computers] | You Will Learn Python 3! Zed Shaw has perfecte... | [{'type': 'ISBN_13', 'identifier': '9780134693... | https://play.google.com/store/books/details?id... | en | NOT_MATURE | 320.0 |

## 2. Below is the sample of collected user data from Goodreads API:

| isbn | user_id | user_name | user_ratings | user_review_text |
|---|---|---|---|---|
| 9780321934949 | 753824 | Rolf Häsänen | 5.0 | Scott Kelby aka Mr Photoshop dishes out some n... |
| 9780321934949 | 4728978 | Chris Hlady | 2.0 | It seems everybody loves this series of books,... |
| 9781429957113 | 10519601 | Vicky "phenkos" (semi-hiatus) | 3.0 | Susan Sontag starts her book on photography wi... |
| 9780133856880 | 27610501 | Nazmus | 4.0 | As the title says, it's a recipe book. The aut... |
| 9780133856880 | None | None | NaN | None |
| 9780133856934 | None | None | NaN | None |

## Data Preprocessing:

For categories which have missing values, Google API does not include the specific field in API response. For example, below image has missing average rating and ratings count. Similarly, other fields can also be missing. In data pre-processing, I am reviewing such fields/values and working on filtering them. I am also working on removing uninformative/redundant features.

Below is the sample of unprocessed data:

| | title | averageRating | ratingsCount | authors | categories |
|---|---|---|---|---|---|
| 0 | Learning Python | 3.0 | 4.0 | [Mark Lutz] | [Computers] |
| 1 | Learning Python | 4.0 | 12.0 | [Mark Lutz] | [Computers] |
| 2 | Deep Learning with Python | 5.0 | 2.0 | [Francois Chollet] | [Machine learning] |
| 3 | Learning Python | 5.0 | 1.0 | [Fabrizio Romano] | [Computers] |
| 4 | Learn Python 3 the Hard Way | 5.0 | 1.0 | [Zed A. Shaw] | [Computers] |
| 5 | Learning Python with Raspberry Pi | NaN | NaN | [Alex Bradbury, Russel Winder, Ben Everard] | [Computers] |
| 6 | Python Projects | NaN | NaN | [Laura Cassell, Alan Gauld] | [Computers] |
| 7 | Introduction to Machine Learning with Python | 3.0 | 2.0 | [Andreas C. Müller, Sarah Guido] | [Computers] |
| 8 | Python Machine Learning | 5.0 | 2.0 | [Sebastian Raschka] | [Computers] |
| 9 | Python for Kids | 5.0 | 3.0 | [Jason R. Briggs] | [Juvenile Nonfiction] |
| 10 | Learning Python Application Development | NaN | NaN | [Ninad Sathaye] | [Computers] |

**Approach:**

I am planning to gather statistics about collected data for analysis. I am planning to use Content-based and Collaborative filtering for recommendation. I am also planning to use Hybrid recommendation, to get the best possible results.

I am trying to incorporate recommendation using following approaches:

1. Trending books – which have maximum number of ratings and highest average rating.
2. Similarity between books for a user - Content-based filtering
3. Similarity measures related to ratings of books - Item-based collaborative filtering
4. Similarity between users - User-based collaborative filtering
5. Combination of Content-based and Collaborative filtering – Hybrid recommendation

**Analysis:**

For text analysis, I am planning to use feature extraction techniques like TF-IDF Vectorizer, etc. I am also planning to use Doc2Vec for word embedding.

**Testing / Expected Results:**

Till now, I have collected sample data and doing initial analysis. I am still working on developing recommendation system using above mentioned approaches. I will share the results in final report. I am planning to validate model on training dataset and test it on test dataset. The expected model should recommend most relevant books to the users, considering user similarity, books similarity or both.