Laporan Proyek Machine Learning: Books Recommendation System - Vian Sebastian Bromokusumo

Project Overview

Latar Belakang

Menurut survei oleh Badan Pusat Statistik Indonesia yang dipublikasikan pada 7 Juni 2023, jumlah populasi Indonesia adalah sebesar 278,696 juta [1]. Namun sangat disayangkan, menurut survei UNESCO [2], hanya sekitar 0.001% dari masyarakat Indonesia yang memiliki minat membaca. Hal ini tentunya sangat berpengaruh terhadap tingkat literasi Indonesia, dimana menurut Balai Bahasa Provinsi Sumatera Utara [3], Indonesia masuk dalam 10 negara dengan tingkat literasi terendah. Maka dari itu, diperlukan strategi yang efektif dan efisien untuk meningkatkan minta membaca masyarakat Indonesia, dengan harapan meningkatkan tingkat literasi secara keseluruhan.

Proyek ini difokuskan pada dataset "Book Recommendation Dataset" dengan tujuan untuk menghasilkan suatu Recommendation System yang dapat merekomendasikan buku yang relevan kepada user. Dalam implementasi lanjutannya, harapannya Recommendation System ini dapat memberikan rekomendasi mulai dari literasi ilmiah, buku-buku hiburan, hingga barang-barang dalam bisnis e-commerce dan film-film dalam bisnis perfilman. Jumlah data yang sangat besar dalam dataset ini menjadi salah satu alasan Penulis menggunakan dataset ini, dikarenakan dataset yang besar akan membantu model untuk membangun sistem rekomendasi yang lebih baik dan custom bagi user.

Menurut Jepchumba dari Microsoft [4], *machine learning* merupakan teknik yang menggunakan matematika tingkat tinggi dan ilmu statistika untuk mengenali pola pada data yang tidak ada secara eksplisit, dan dapat memprediksi sesuai dengan hasil pola tersebut. Terdapat dua metode yang banyak digunakan dalam domain sistem rekomendasi ini, antara lain Content Based Filtering, dan Collaborative Filtering.

Menurut Google Machine Learning Developers, Content Based Filtering [5] merupakan teknik mendapatkan rekomendasi item berdasarkan items yang disukai user, dengan cara menghitung Cosine Similarity antar items. Hal ini tentu sangat berat dalam komputasinya, mengingat dalam kasus nyata, akan ada ribuan hingga jutaan items yang perlu dikaji untuk user. Ditambah lagi, menurut sumber yang sama, Collaborative Filtering [6] merupakan teknik yang mengkaji bukan hanya items yang disukai user, namun komunitas user itu sendiri untuk menghasilkan rekomendasi yang lebih baik. Hal ini tentunya membutuhkan komputasi yang lebih rumit lagi, menggunakan Deep Learning seperti Neural Network. Alhasil, dalam membangun Recommendation System, diperlukan Machine Learning untuk mempercepat dan menghasilkan rekomendasi yang efektif bagi user dan sasaran lainnya.

Proyek ini menjadi sarana kecil untuk membantu meningkatkan minat membaca dan literasi Indonesia, dan hasil dari proyek ini diharapkan dapat membantu Pemerintah, instansi literatur, hingga individual untuk mengembangkan minat membaca dan edukasi literasi.

Business Understanding

Stakeholder dan sasaran:

1. Pemerintah Sebagai organisasi tingkat tertinggi dalam negara, tentunya pemerintah dapat menggunakan Recommendation System untuk meningkatkan minta baca masyarakat. Dengan strategi lainnya yang dapat digunakan oleh pemerintah, sistem rekomendasi yang baik diharapkan dapat mendukung strategi pemerintah dalam meningkatkan minat baca dan literasi Indonesia.
2. Perpustakaan Sebagai instansi literatur paling tua di dunia, tentunya perpustakaan-perpustakaan menyimpan ilmu-ilmu yang sudah ada sejak lama. Recommendation System yang baik akan membantu perpustakaan untuk

melayani pendatang dan pembaca yang datang untuk mencari berbagai referensi dan literatur lainnya dengan lebih baik.

3. Mesin Pencari Bukan hanya perpustakaan, Mesin Pencari seperti Google Scholar sudah menjadi gudang literasi penelitian bagi para ilmuwan dan pelajar yang ingin meningkatkan ilmunya dalam bidang tertentu. Recommendation System yang baik akan membantu pembaca untuk lebih mudah mencari informasi yang berhubungan, guna meningkatkan literasinya dalam ilmu yang didalami.

4. Individu Dalam konteks individu, tentunya Recommendation System tidak hanya berguna dalam konteks mencari ilmu, namun juga dapat berguna untuk membantu pembaca-pembaca mencari hiburan seperti buku fiksi dan genre-genre lainnya. Harapannya, output proyek ini dapat mendukung kemajuan literasi per individu.

Problem Statements

1. Dengan banyaknya jumlah buku dan genre, apakah Recommendation System yang tepat sasaran (efektif) dapat dibuat?

Recommendation System Goals

1. Membuat Recommendation System yang efektif dan tepat sasaran, guna membantu sasaran (perpustakaan, mesin pencari, maupun individu).

Solution Statements (Metodologi)

1. Melakukan Exploratory Data Analysis untuk mendapatkan informasi berguna dalam data dan mengetahui dinamika fitur-fitur.

2. Membuat model machine learning yang dapat merekomendasikan buku dengan tepat sasaran, menggunakan metode Content Based Filtering dan Collaborative Filtering.

3. Menggunakan metrik evaluasi Precision@k dan Root Mean Squared Error untuk mengevaluasi performa model.

Data Understanding

Dataset: https://www.kaggle.com/datasets/arashnic/book-recommendation-dataset/data

Dataset Overview

Dataset Book Recommendation ini terbagi menjadi tiga dataset, antara lain Books, Ratings, dan Users, dengan deskripsi sebagai berikut.

Books:

1. ISBN (object) : kode identifikasi unik buku
2. Book-Title (object) : judul buku
3. Book-Author (object) : penulis buku
4. Year-Of-Publication (object) : tahun terbit
5. Publisher (object) : penerbit
6. Image-URL-S (object) : tautan gambar kecil buku
7. Image-URL-M (object) : tautan gambar sedang buku
8. Image-URL-L (object) : tautan gambar besar buku

Ratings:

1. User-ID (int64) : ID user yang memberikan rating
2. ISBN (object) : ISBN buku yang di rating
3. Book-Rating (int64) : rating buku

Users:

1. User-ID (int64) : ID user yang terdaftar
2. Location (object) : lokasi domisili user
3. Age (float64) : umur dari user

Deskripsi rinci dari dataset ini adalah sebagai berikut

1. Terdapat 271,360 sampel Books data, dengan 8 fitur object.
2. Terdapat 1,149,780 sampel Ratings data, dengan 2 fitur numerik dan 1 fitur object.
3. Terdapat 278,858 sampel Users data, dengan 2 fitur numerik dan 1 fitur object.
4. Terdapat 6 missing values pada Books data, dan tidak ada missing values pada Ratings data.
5. Terdapat 110,762 missing values pada kolom 'Age' pada User data. Namun umur tidak akan terlalu dipertimbangkan dalam proyek ini.

EDA - Univariate Analysis

Beberapa hasil dari Univariate Analysis pada data ratings, users, dan books adalah sebagai berikut.

1. dari total user yang terdaftar (278,853 users), hanya 105,283 user yang melakukan rating
2. dari total buku yang terdaftar (271,360 buku), terdapat 340,556 buku yang diberikan rating. Hal ini dapat disebabkan oleh buku yang tidak terdaftar pada data Books, namun ada di ratings.
3. terdapat 102,024 penulis yang terdaftar.
4. rating terdiri dari skala 1-10.
5. Terdapat perbedaan jumlah buku (ISBN) dan judul. Hal ini dapat disebabkan oleh judul yang sama, namun berbeda versi.

Data-data tersebut dapat dilihat lebih rinci pada beberapa output di bawah ini.

1. Output dari Ratings data:
    - rating users: 105283
    - rated books: 340556
    - rating range: [0 5 3 6 8 7 10 9 4 1 2]
2. Output dari Books data:
    - registered books: 271360
    - registered authors: 102024
    - registered titles: 242135
3. Output dari Users data:
    - registered users: 278858

Data Preparation

Secara berurutan, proses Data Preparation yang akan dilakukan adalah

1. *Data Merging*

    Data Merging merupakan prosedur umum dalam proses data Preparation, dimana dalam proyek ini bertujuan untuk menggabungkan data Ratings dengan data Books, untuk mempermudah proses data Cleaning, dan membantu model dalam pengolahan data nantinya.

2. *Missing Values Handling*

Missing Values Handling merupakan prosedur standar Data Cleaning, bertujuan untuk menghilangkan data kosong. Hal ini perlu dilakukan karena data yang kosong berpotensi sangat besar untuk mengacaukan kalkulasi, baik kalkulasi neural-network (deep learning) atau matematis dan stastik.

3. *Duplicates Data Handling*

   Duplicates Data Handling juga merupakan prosedur standar dari Data Cleaning. Mengingat bahwa data yang dimiliki sudah sangat besar, penting untuk menghilangkan data-data duplikat untuk mengurangi jumlah data yang tidak terlalu berguna. Data Duplikat pada hakikatnya dapat meningkatkan potensi overfitting.

4. *Data Selection*

   Data Selection dalam kasus proyek ini ialah memilih jumlah sampel data yang ingin digunakan, kemudian memilah kolom yang ingin digunakan dan memasukannya ke dalam satu DataFrame baru.

Berikut adalah alur dari proyek sesuai dengan penjelasan tahap-tahap di atas.

1. *Data Merging*

   Input dari tahap ini adalah 3 data besar, antara lain Books, Ratings, dan Users. Merge dilakukan pada data Ratings dan Books pada fitur ISBN.

2. *Missing Values Handling*

   Setelah proses Merge, terdapat rata-rata 118,640 missing values pada kolom Book-Title, Book-Author, Year-Of-Publication, Publisher, Image-URL-S, Image-URL-M, dan Image-URL-L. Pada tahap ini pula, data total dari hasil Merge adalah sebanyak 1,149,780 sampel data. Dengan melimpahnya data ini, metode Missing Values Handling yang dilakukan Penulis adalah metode Drop. Hasilnya adalah 1,031,129 sampel data yang tidak memiliki nilai Null.

3. *Duplicates Data Handling*

   Tahap berikutnya adalah Duplicates Data Handling, dimana Penulis melakukan pengecekan pada kolom 'ISBN' dan 'Book-Titles'. Hasilnya adalah terdapat 760,984 nilai duplikat pada kolom 'ISBN', dan 790,063 nilai duplikat pada kolom 'Book-Title'. Menghilangkan duplikat ini penting, untuk mengurangi potensi overfitting, terutama ketika menggunakan Content Based Filtering, data yang duplikat akan memiliki similarity yang sama, sehingga akan direkomendasikan buku yang tidak sesuai.

4. *Data Selection*

   Data Selection dalam kasus proyek ini akan mengambil 30,000 sampel dari total data bersih 241,066 sampel, kemudian memilah fitur yang diinginkan ke sebuah DataFrame baru. Angka 30,000 sampel diambil dikarenakan beban komputasi yang berat. Pada saat proses training, Penulis menemukan beberapa kasus *crash* pada Google Collab, disebabkan oleh dataset yang digunakan terlalu besar. Setelah beberapa percobaan, 30,000 sampel menjadi jumlah data maksimum yang dapat digunakan yang tidak menyebabkan *crash*.

Data Preparation - Collaborative Filtering

---

1. *Encoding and Mapping*

   Encoding dan Mapping adalah proses untuk mengubah data object menjadi numerik. Hal ini penting dilakukan untuk mempermudah model mengolah data. Mapping dilakukan untuk menambah data yang di sudah di-encoded kembali ke DataFrame.

2. *Fetching Random Samples*

Fetching Random Samples pada proyek ini dilakukan agar distribusi data menjadi random. Hal ini baik dilakukan untuk mencegah terjadinya overfitting pada model.

3. *Train Test Split*

   Prosedur standar dalam banyak proyek Machine Learning, Train Test Split adalah proses membagi data menjadi data Train (latih), dan Test (uji), dimana Train untuk latihan model, dan Test untuk uji model.

Berikut adalah alur dari proyek sesuai dengan penjelasan tahap-tahap di atas.

1. *Encoding and Mapping*

   Pada proses ini, Encoding dilakukan pada kolom 'User-ID' dan 'Book-Title', dimana Mapping dilakukan pada kolom baru 'user' dan 'books'. Penulis menggunakan 'Book-Title' sebagai fitur yang di-encode karena rekomendasi yang ingin dilakukan berdasarkan judul buku nantinya.

2. *Fetching Random Samples*

   Proses pengambilan sampel random dilakukan dengan menggunakan fungsi sample(), dengan parameter:

   - frac = 1 merupakan parameter yang mengatur berapa persen (fraction) dari baris yang ingin digunakan. frac = 1 berarti kita ingin mengambil seluruh baris dari DataFrame.
   - random_state = 123 dimana random_state merupakan parameter yang mengatur reproduktifitas agar data yang diacak konsisten.

3. *Train Test Split*

   Pada proyek kali ini, nilai x diambil dari kolom 'user' dan 'books', dan y diambil dari kolom 'Book-Rating'. Pembagian data Train dan Test dilakukan sebesar 0.8, atau 80% data Train dan 20% data Test.

Model Development - Content Based Filtering

Tahap-tahap yang akan dilakukan Penulis pada Content Based Filtering adalah sebagai berikut:

1. *Vectorizer Calculations*

   Pada tahap ini, Penulis mengimpor TfidVectorizer, dan menggunakan metode TF-IDF (Term Frequency-Inverse Document Frequency) untuk menghitung kepentingan kata-kata dalam dokumen. Proses ini bertujuan untuk melakukan scaling dan normalisasi dengan membagi jumlah kemunculan kata-kata terhadap panjang dokumen. Tahap-tahap yang dilakukan adalah sebagai berikut: - memanggil TfidVectorizer. - memanggil fungsi fit(). - memanggil fungsi get_features_name_out(). - mentransformasikan data yang telah di fit() ke matrix, dengan memanggil fit_transform(). - memasukkan matrix ke dalam DataFrame, dengan memanggil fungsi todense().

2. *Cosine Similarity Calculations*

   Cosine Similarity merupakan teknik untuk menghitung kesamaan antar dua vektor, dengan cara menghitung sudut cosinus dan dot product antar dua vektor. Semakin kecil sudut cosinusnya, semakin besar nilai Cosine Similarity-nya. Cosine Similarity merupakan teknik yang sangat sering digunakan untuk mengukur kesamaan dalam analisis teks.

   Pada tahap ini, cukup mudah dilakukan, hanya dengan mengimpor cosine_similarity, dan memanggil cosine_similarity() pada matriks hasil TF-IDF.

3. *Recommendation Retrieval Function Initialization*

   Pada tahap ini, dibuat fungsi book_recommendations() yang mengambil parameter judul buku (Titles), similarity_data (sim_df), items, dan k.

Titles merupakan input yang dimasukkan, dimana dibuat agar rekomendasi dibuat berdasarkan judul buku. Similarity_data merupakan DataFrame yang merupakan hasil perhitungan Cosine Similarity, dan items merupakan hasil output yang ingin dikeluarkan yaitu dalam bentuk DataFrame berisi 'ISBN', 'Authors', dan 'Title'. Nilai k merupakan jumlah rekomendasi yang ingin dikeluarkan, dalam hal ini, diatur menjadi 5.

Dalam fungsi ini, parameter yang perlu diisi hanyalah judul. dan melihat dari similarity_data-nya (sim_df) yang merupakan hasil penghitungan, akan menghasilkan output berupa 'items' antara lain ISBN, Authors, dan Titles.

4. *Sample Retrieval*

   Tahap ini cukup mudah dimengerti, yaitu mengambil contoh sampel dari data.

   Tabel 1.0 Sampel Rekomendasi Content Based Filtering Model

ISBN	Authors	Titles
009181460X	Kenton	Sleep Deep

5. *Recommendation Result -- Top-N Recommendations*

   Menggunakan fungsi book_recommendations() yang sudah dibuat, Penulis menginput satu judul buku, dan hasilnya meng-outputkan 5 rekomendasi buku. Selengkapnya dapat dilihat di bagian Evaluasi.

   Tabel 1.1 Hasil Rekomendasi Top-N Content Based Filtering Model

Titles	ISBN	Authors
The deep	0233967931	Peter Benchley
House of Sleep	0140250832	Jonathan Coe
The Little Book of Sleep	0140280693	Paul Wilson
Doctor Sleep	0151261008	Madison Smartt Bell
Deep in the Heart	0061083267	Sharon Sala

Model Development - Collaborative Filtering

Tahap-tahap yang akan dilakukan Penulis pada Collaborative Filtering adalah sebagai berikut:

1. *Model Class Initialization*

   Terinsipirasi dari Studi Kasus Recommendation System dari Dicoding Academy, Penulis akan menggunakan RecommenderNet dengan Keras Model class. Model ini akan menghitung skor kecocokan antar user dan judul buku dengan teknik embedding, melakukan operasi dot product, dan menambahkan bias untuk setiap user dan judul buku. Pada akhirnya akan menghasilkan skala kecocokan dari 0 hingga 1, dengan Sigmoid Activation Function.

2. *Callback Functions Initialization*

   Pada proyek ini, Penulis membuat dua macam callback, antara lain:

   - lr_reduction, sebuah callback yang mengimplementasikan ReduceLROnPlateau, sebuah callback dari tensorflow yang memiliki parameter sebagai berikut:

     1. monitor, diatur untuk memonitor 'root_mean_squared_error'
     2. patience, diatur untuk menunggu 3 epoch sebelum mengurangi learning_rate
     3. verbose, diatur menjadi 1 untuk transparansi

4. factor, diatur menjadi 0.1
5. min_lr, diatur untuk tidak mengurangi learning_rate jika sudah menyentuh 0.000001.

- earlyStop, sebuah callback custom yang dibuat Penulis yang memonitor 'root_mean_squared_error', dan dibuat untuk menghentikan proses training jika sudah menyentuh value RMSE yang diinginkan.

3. *Model Compilation*

Pada tahap ini, dilakukan dua proses.

- model initialization, yaitu memanggil kelas RecommenderNet(), dengan parameter:

  1. num_users, dalam kasus ini 8951 users,
  2. num_books, dalam kasus ini 30.000 buku,
  3. dan embedding size, dalam kasus ini 50.

- model compiling, yaitu melakukan model.compile(), dengan parameter sebagai berikut:

  1. loss, parameter loss function, yang diatur menjadi Binary Crossentropy
  2. optimizer, parameter optimizer, yang diatur menjadi Adam, dengan learning_rate = 0.001
  3. metrics, parameter metrik evaluasi, yang diatur menjadi Root Mean Squared Error (RMSE)

4. *Model Training*

Pada model training, Penulis melakukan model.fit(), dan dimasukkan kedalam variabel 'history' untuk mempermudah proses visualisasi metrik.

Parameter yang dimasukkan pada model.fit() ini antara lain:

- x, yaitu x_train
- y, yaitu y_train
- batch_size, diatur menjadi 8, untuk mempercepat proses training
- epochs, diatur menjadi 40 untuk mempercepat proses training
- verbose, diatur menjadi 1 untuk melihat proses training
- validation_data, yaitu x_val dan y_val
- callbacks, yaitu lr_reduction dan earlyStop yang dibuat sebelumnya

5. *Sample Retrieval*

Sample Retrieval yang dilakukan pada Collaborative Filtering ini merupakan satu user random, kemudian dilakukan ekstraksi buku-buku apa saja yang disukai oleh user tersebut, dan disimpan ke dalam variabel 'rated_books'

Tabel 2.0 Sampel Rekomendasi Collaborative Filtering Model

User-ID	ISBN	Book-Rating	Book-Title	Book-Author	Year-Of-Publication	Publisher	user	books
204359	0060808365	9	Documents in the Case	Dorothy Leigh Sayers	1987	Harpercollins	1074	6508
204359	0299187349	5	The Museum of Happiness: A Novel (Library of A...	Jesse Lee Kercheval	2003	University of Wisconsin Press	1074	25900

User-ID	ISBN	Book-Rating	Book-Title	Book-Author	Year-Of-Publication	Publisher	user	books
204359	0025211609	9	Influence	Ramsey Campbell	1988	Simon & Schuster	1074	1691
204359	0140062580	0	The Vendor of Sweets (King Penguin S.)	R.K. Narayan	1983	Penguin Books Ltd	1074	14467
204359	0140230246	10	Middlemarch	George Eliot	1994	Penguin Books Ltd	1074	16354
204359	0060932279	0	Old Man in a Baseball Cap: A Memoir of World W...	Fred Rochlin	2000	Perennial	1074	7149
204359	0151002231	0	Homosexuality In History	Colin Spencer	1996	Harcourt	1074	20403

6. *Prediction Result -- Top-N Recommendations*

Setelah mendapatkan sampel dan data terkait buku-buku yang dimiliki, prediksi dilakukan ke model yang sudah di-train. Hasilnya akan berupa output 10 buku dengan rating terbaik yang belum pernah dirating pembaca (diasumsikan belum pernah dibaca), dengan judul, ISBN, dan penulis buku tersebut. Selengkapnya dapat dilihat di bagian Evaluasi.

Table 2.1 Hasil Rekomendasi Top-N Collaborative Filtering Model

Title	ISBN	Author
On Top of the World: Cantor Fitzgerald, Howard Lutnick, & 9/11	0060510293	Tom Barbash
One Flew over the Cuckoo's Nest	0140043128	Ken Kensey
A Glass of Blessings	0060805501	Barbara Pym
To you with love	0285622714	Terry Rowe
Magnificent Prayer	0310238447	Nick Harrison
A Manual for Writers of Term Papers, Theses, and Dissertations (Manual for Writers of Term Papers, Theses, & Dissertations (Paperback))	0226816214	Kate Turabian
Mother Goose Rhymes (Golden Little Look-Look Book)	0307117561	Golden Books
Heart At Work	0070116431	Jack Canfield
The Callender Papers	0006729835	Cynthia Voight

	Title	ISBN	Author
	7 Days to a Magickal New You	0007123469	Fiona Horne

Model Explanation, pros and cons

Pada proyek ini, Penulis menggunakan dua pendekatan algoritma, antara lain Content Based Filtering dan Collaborative Filtering.

1. **Content Based Filtering**

   Content Based Filtering adalah model sistem rekomendasi berdasarkan konten, yang merekomendasikan konten (items) yang mirip dengan konten yang telah disukai oleh pengguna di masa lalu. Cara kerja Content Based Filtering adalah dengan melihat konten (items) mana yang telah dinilai, artinya disukai, oleh pengguna. Model kemudian akan menyarankan item yang serupa.

   Kelebihan dari model Content Based Filtering adalah besarnya peluang user untuk menyukai hasil rekomendasi, karena rekomendasi tersebut didasarkan dari items yang sudah disukai oleh pengguna tersebut. Di sisi lain, kekurangan dari model ini ialah kecenderungan untuk kurang dapat memberikan rekomendasi item yang unik.

2. **Collaborative Filtering**

   Collaborative Filtering adalah model sistem rekomendasi berdasarkan pendapat komunitas pengguna. Dalam konteks proyek ini, model ini bekerja dengan mengidentifikasi buku-buku dengan rating yang tinggi, dan mirip dengan yang dimiliki user.

   Kelebihan dari model Collaborative Filtering ialah kebalikan dari algoritma Content Based Filtering, yaitu kemampuannya dalam merekomendasikan items unik. Namun, kekurangannya juga merupakan kebalikannya, yaitu terdapat peluang user kurang menyukai hasil rekomendasinya, dikarenakan dasar dari rekomendasi model ini bukan hanya items yang telah disukai user, namun pendapat komunitas, dalam hal ini, buku-buku lain dengan rating tinggi.

Evaluation

Dalam proyek ini, beberapa metrik evaluasi yang digunakan adalah sebagai berikut.

1. **Content Based Filtering**

$$Precision@k = \frac{Rr}{Ra}$$

dimana,

- Precision@k = presisi dari k-prediksi rekomendasi
- Rr = jumlah rekomendasi relevan
- Ra = jumlah seluruh rekomendasi

---

1. **Collaborative Filtering** $$RMSE = \sqrt{\sum{(Y_t - Y_p)^2} \over n}$$

dimana,

- Yt = Y aktual,
- Yp = Y prediksi,
- n = jumlah data

Model Training Results

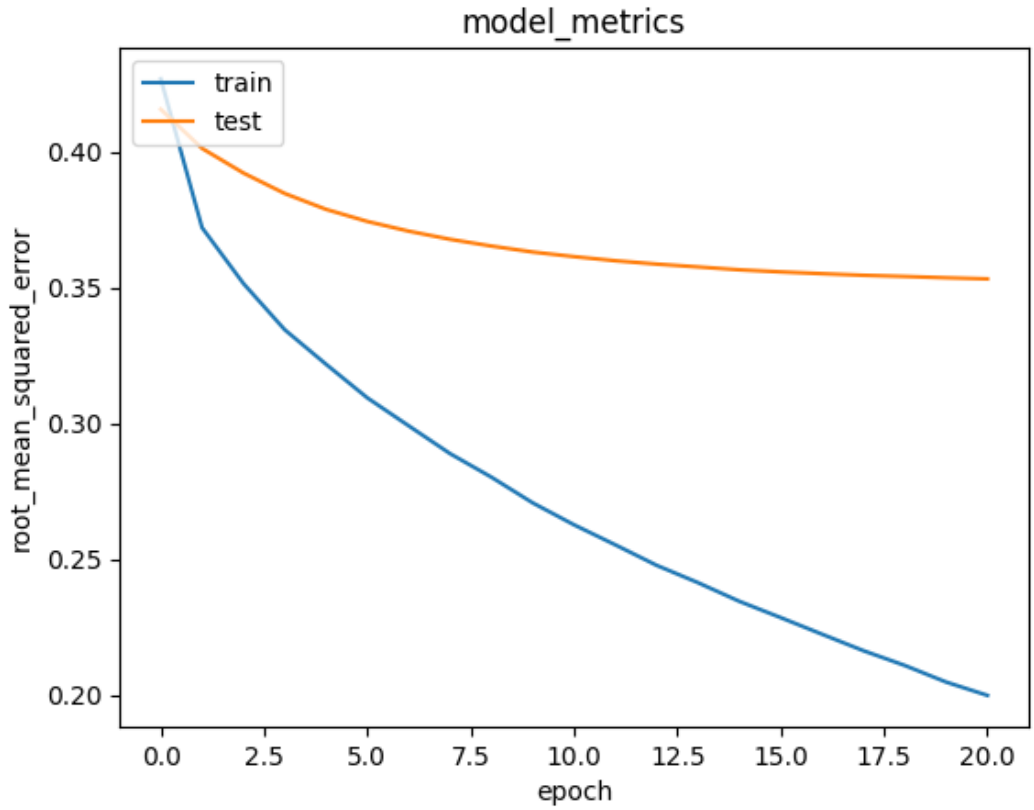
- *Content Based Filtering*

Table 3.0 Cuplikan hasil Cosine Similarity Calculation

Titles \ Titles	Where Peachtree Meets Sweet Auburn: A Saga of Race and Family	Discover Power Within	La Symphonie Pastorale Isabelle	Samurai William: The Englishman Who Opened Japan	Ian Shoales' Perfect World
Switch Bitch	0.000000	0.0	0.0	0.000000	0.0
Man and Superman : A Comedy and a Philosophy	0.041173	0.0	0.0	0.000000	0.0
The Secret Life of the Seine	0.016932	0.0	0.0	0.023209	0.0
CORBIE	0.000000	0.0	0.0	0.000000	0.0
The Stones of the Abbey	0.017753	0.0	0.0	0.024334	0.0
Madam, Will You Talk?	0.000000	0.0	0.0	0.000000	0.0
Marie Bonaparte	0.000000	0.0	0.0	0.000000	0.0
A Return to Love: Relfections on the Principles of a Course in Miracles	0.011489	0.0	0.0	0.007874	0.0
Crossing Brooklyn Ferry : A Novel	0.000000	0.0	0.0	0.000000	0.0
The Last Wanderer	0.000000	0.0	0.0	0.014243	0.0

- *Collaborative Filtering*

Visualisasi dari hasil training model Collaborative Filtering dapat dicermati di bawah ini.



Gambar 1.0 Hasil Training model Collaborative Filtering

Dilihat dari gambar 1.0, secara sekilas terlihat bahwa model ini merupakan model overfit, dikarenakan perbedaan besar pada grafik. Namun, jika dicermati lebih lanjut, angka-angka pada epoch terakhir adalah sebagai berikut.

- root_mean_squared_error: 0.1998
- val_root_mean_squared_error: 0.3531

Ketika dihitung, didapatkan bahwa perbedaan RMSE dan validation RMSE hanya ~0.16, sehingga perbedaannya tidak terlalu signifikan. Alhasil, model Collaborative Filtering ini tidak overfit, melainkan sebuah model good fit.

Prediksi Akhir dan Diskusi

Diskusi Hasil Content Based Filtering

Merujuk pada rumus 'Precision@k', dapat disimpulkan bahwa hasil rekomendasi Content Based Filtering sudah sangat baik, dengan 5/5 (100%) rekomendasi memiliki kata kunci kemiripan dengan buku sampel, antara lain 'Sleep' dan 'Deep'. Hasil dari Content Based Filtering ini dikatakan sudah cukup memuaskan untuk menjadi Recommendation System yang efektif bagi user.

Diskusi Hasil Collaborative Filtering

Seperti yang sudah dijelaskan pada bagian Model Training Results, hasil dari training model Collaborative Filtering sudah cukup baik, dengan model yang good fit. Hal ini menjadi indikasi yang baik bahwa model ini sudah bisa menjadi Recommendation System yang efektif.

Hasil dan Kesimpulan Proyek

Berdasarkan dari hasil Data Understanding, Data Preparation, Model Development, dan Evaluation, kesimpulan Proyek ini dapat disimpulkan sebagai berikut.

1. Menjawab Problem Statement 1: Dengan banyaknya jumlah buku dan user yang ada, Recommendation System efektif **dapat dibuat dengan baik.**
2. Jalannya proyek ini sudah sesuai dan memuaskan Penulis, dimulai dari menjawab Problem Statement, mencapai Recommendation System Goals yang dirumuskan, serta dapat mengevaluasi performa dengan baik. Alhasil, harapannya proyek ini dapat menjadi sarana bagi sasaran-sasaran yang sudah diuraikan, dan menjadi referensi bagi pelajar-pelajar Machine Learning, terutama pembuat Recommendation System.

.

Referensi

[Dataset] https://www.kaggle.com/datasets/arashnic/book-recommendation-dataset/data

[1] Badan Pusat Statistika Indonesia. (2023). Diakses dari https://www.bps.go.id/id/statistics-table/2/MTk3NSMy/jumlah-penduduk-pertengahan-tahun--ribu-jiwa-.html

[2] R. Pramudjasi, Juliansyah, D. Lestari. (2019). Effect of population and education and wages on unemployment in paser regency. Journal of Faculty of Economics and Business Universitas Mulawarman. Diakses dari https://journal.feb.unmul.ac.id/index.php/KINERJA/article/download/5284/472.

[2] UNESCO Institute of Statistics. Diakses dari https://uis.unesco.org/en/country/id

[3] Balai Bahasa Provinsi Sumatera Utara, Kemendikbud. (2023). “MANCA” untuk Literasi yang Menyenangkan. Diakses dari https://balaibahasasumut.kemdikbud.go.id/2023/09/07/manca-untuk-literasi-yang-menyenangkan/#:~:text=Dengan%20kata%20lain%2C%20Indonesia%20masuk,2022%20mencapai%2051%2C69%25.

[4] B. Jepchumba. (2020). Getting started with using Visual Machine Learning Tools for building your Machine Learning Models. Microsoft Community Hub. Diakses dari https://techcommunity.microsoft.com/t5/educator-developer-blog/getting-started-with-using-visual-machine-learning-tools-for/ba-p/3578397.

[5] Google Machine Learning Developers. Content-based Filtering. Diakses dari https://developers.google.com/machine-learning/recommendation/content-based/basics

[6] Google Machine Learning Developers. Collaborative Filtering. Diakses dari https://developers.google.com/machine-learning/recommendation/collaborative/basics

Daftar Pustaka

[1] Dicoding Academy. Diakses dari https://www.dicoding.com/ .