PROPOSAL PROYEK

12S4054 - DATA MINING

Product Matching Book Analysis using K-means Clustering

PENGUSUL

12S17017 – Yolanda R.M Manurung

12S17027 – Stella Sitinjak

12S17059 – Ekis Naomi Lasma

FAKULTAS INFORMATIKA DAN TEKNIK ELEKTRO

INSTITUT TEKNOLOGI DEL

NOVEMBER 2020

DAFTAR ISI

Daftar Gambar

Daftar Tabel

Pendahuluan

Bagian ini berisi latar belakang dan tujuan pengerjaan proyek.

Latar belakang

Perkembangan teknologi yang pesat pada era industri 4.0 menyebabkan penggunaan digital semakin populer diseluruh golongan masyarakat dengan tawaran efisiensi yang menggiurkan. Hampir seluruh golongan masyarakat menggunakan aplikasi web untuk memudahkan segala pekerjaan yang sebelumnya dilakukan secara offline atau datang ketempat langsung. Aplikasi web yang kerap kali digunakan masyarakat adalah media sosial, e-commerce, kursus online, periklanan hingga internet banking. Salah satu hal yang menjadi populer pada era ini adalah e-commerce atau belanja online, dimana seluruh proses pembelian yang biasanya dilakukan secara langsung dengan cara mengunjungi toko atau store sekarang dapat dilakukan melalui smartphone dengan media aplikasi belanja online. Efisiensi yang diberikan dari kemudahan berbelanja tentunya sangat diminati terutama pada masa pandemi ini. Berbagai produk ditawarkan melalui aplikasi belanja online tersebut dengan keanekaragaman dan variasi serta harga yang bersaing. Namun, terdapat permasalahan besar dalam melakukan proses pencarian sebuah barang atau produk, dimana dalam kasus ini yang akan diteliti oleh penulis adalah produk buku, dimana pembeli bisa menemukan buku dengan kategori, penulis, dan bahkan penerbit yang sama namun dalam toko yang berbeda dalam sebuah aplikasi. Tidak hanya itu, bahkan terkadang aplikasi menawarkan buku yang tidak terkait dengan pencarian buku yang dilakukan oleh customer. Oleh karena itu, penulis ingin membantu menyelesaikan permasalahan tersebut dengan mengimplementasikan product matching buku berdasarkan review buku yang diberikan oleh customer sebelumnya untuk menemukan pencarian kategori buku yang memiliki korelasi dan erat kaitannya antara 1 buku dengan buku yang lain dan buku yang ditawarkan oleh sistem hanyalah buku yang memiliki hasil review positif dan memiliki rating yang tinggi, sehingga customer tidak akan dibingungkan dalam melakukan pembelian buku dalam sebuah aplikasi e-commerce.

Dataset yang akan digunakan yaitu dataset goodreads book reviews 20k yang diambil dari *book readers social network* www.goodreads.com pada tahun 2017 [2]. Kumpulan data tersebut berjumlah 21,6k data yang memiliki 15 atribut, yaitu:

Tabel 1 Atribut Dataset

No.	Nama Atribut	Tipe Data
1.	book_title	string
2.	book_series	string
3.	book_series_url	string
4.	book_image	string
5.	book_image_url	string
6.	book_rating	decimal
7.	book_author	string
8.	author_url	string
9.	genre	string
10.	reviewer_name	string
11.	reviewer_url	string
12.	reviewer_image	string
13.	reviewer_image_url	string
14.	review	string
15.	ID	decimal

Alasan Topik Dipertahankan

Alasan topik dipertahankan oleh penulis dikarenakan dataset yang diperoleh oleh penulis bersumber dari Kaggle (link: https://www.kaggle.com/gapple/goodreads-book-reviews-20k), masih belum memiliki task apapun yang dikerjakan oleh *user* atau pengguna dari Kaggle yang bisa dibuktikan dengan link: https://www.kaggle.com/gapple/goodreads-book-reviews-20k/tasks yang menunjukkan bahwa tidak ada bocoran kode *product matching* buku dengan menggunakan dataset tersebut. Selain kedua link yang sudah dilampirkan, link: https://www.kaggle.com/gapple/goodreads-book-reviews-20k/notebooks juga dapat dijadikan sebagai pembanding bahwa dataset tersebut tidak pernah di kerjakan oleh *user* Kaggle lain dalam membuat sebuah *product matching*. Notebook tersebut, hanya berisikan sebuah proses Exploratory Data Analysis, hanya sampai disitu saja code yang diberikan, tidak sampai ke proses bagaimana untuk membuat sebuah *product matching*.

Sehingga, penulis akan membuat sebuah sistem *product matching* buku dengan menggunakan dataset yang tidak digunakan oleh peneliti lain (dalam paper, tidak ditemukan peneliti menggunakan

dataset tersebut). Hal yang menjadi pembeda proyek penulis dengan penelitian lainnya:

1. Dataset yang digunakan berbeda
2. Clustering dataset akan dilakukan dengan menggunakan K-means
3. Dataset akan dibersihkan terlebih dahulu (review karena atribut tersebut yang memiliki korelasi paling tinggi ke dalam tujuan proyek)

Tujuan Penelitian

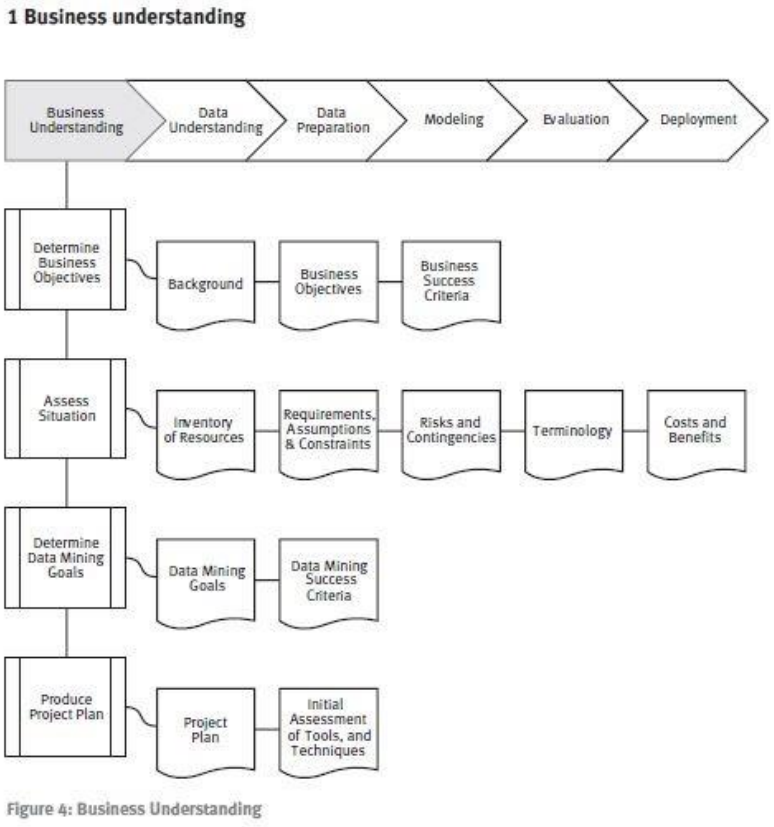
Tujuan penerapan *Data Mining* ini adalah sebagai berikut:

1. Menghasilkan model *product matching book analysis* berdasarkan hasil review yang diberikan oleh *customer* sebelumnya
2. Mengetahui pola (*pattern*) dari data buku sehingga diketahui kriteria atau buku untuk mendapatkan jenis buku yang serupa.
3. Mendukung proses analisis dan pengambilan keputusan dalam pembelian buku pada sebuah aplikasi e-commerce dengan menerapkan CRISP-DM
4. Membuat sistem rekomendasi yang lebih terstruktur sesuai dengan kriteria yang telah ditentukan sebelumnya

Business Understanding

Business Understanding merupakan tahap pertama dalam CRISP-DM yang secara garis besar digunakan untuk mendefinisikan proyek, tujuan dan kebutuhan dari sudut pandang bisnis, lalu akan menterjemahkan pengetahuan yang sudah diperoleh ke dalam pendefenisian masalah pada *data mining* sehingga dapat dilakukan pencocokan terhadap tujuan bisnis sehingga model terbaik dapat dibangun. Tahap *business understanding* juga merupakan tahap yang digunakan untuk memahami dan selanjutnya akan menentukan rencana dan strategi untuk mencapai tujuan yang sudah didefinisikan diawal. Pada tahap ini diperlukan pengetahuan dari objek bisnis tertentu, yaitu bagaimana membangun atau mendapatkan data, dan bagaimana untuk mencocokkan tujuan pemodelan untuk tujuan bisnis sehingga model terbaik dapat dibangun [1].

Dalam sistem yang akan dikembangkan oleh penulis meliputi 4 tahap dalam *business understanding*, yaitu: *determine business objectives, assess situation, determine data mining goals,* dan *produce project plan* [2].



Gambar 1 Business Understanding

1. *Determine Business Objective*

   Tujuan pertama dari menganalisis data adalah agar dapat benar-benar memahami, dari perspektif bisnis, mengenai apa yang ingin dicapai oleh *customer*. Tujuan dilakukannya analisis adalah untuk mendapatkan faktor-faktor penting yang dapat mempengaruhi hasil proyek sehingga penelitian tidak akan menghasilkan jawaban yang benar atas pertanyaan yang salah.

   Tujuan bisnis dalam pembuatan *product matching* adalah untuk menawarkan buku yang memiliki ciri yang sama dengan buku kepada pembaca. Parameter suksesnya *product matching* adalah untuk mendapatkan buku lainnya dengan ciri dan/atau kriteria yang sama dengan suatu buku. Keseimbangan dalam pengelompokkan ciri buku juga menjadi syarat keberhasilan *product matching* yang dilakukan.

2. *Assess Situation*

   Pada tahapan ini akan dilakukan pencarian fakta yang lebih terperinci mengenai semua sumber daya, kendala, asumsi, dan faktor lainnya yang harus dipertimbangkan dalam menentukan tujuan analisis data dan rencana penelitian. Tahapan ini bertujuan untuk memperluas detail dari analisis yang dihasilkan pada tahapan pertama. Sumber daya yang akan digunakan dalam penelitian ini akan diambil dari data buku-buku yang terdapat pada www.goodreads.com pada tahun 2017 sebanyak 21,6k data buku. Pada saat ini sistem *product matching* sudah bisa memberikan hasil buku-buku dengan ciri yang mirip, namun masih terdapat hasil buku yang tidak memiliki ciri yang mirip sehingga diperlukan penambangan yang lebih mendalam untuk memastikan semua hasil buku memiliki ciri yang sama. Adanya hasil buku yang tidak memiliki ciri yang mirip dapat disebabkan karena aturan validasi (*validation rules*) pada sistem kurang baik, sehingga data tidak konsisten.

   Selain permasalahan yang sudah disebutkan, penulis juga mengalami kendala dalam menyelesaikan tugas secara jarak jauh, namun dikarenakan sudah tersedia Github yang bisa menghubungkan hasil kinerja dari satu anggota dengan anggota lainnya, maka tugas ini akan diusahakan semaksimal mungkin dapat diselesaikan sesuai dengan tenggat waktu yang diberikan.

3. *Determine Data Mining Goals*

   Pada tahapan ini akan ditentukan tujuan dalam terminologi bisnis. Tujuan *data mining* menyatakan tujuan proyek dalam istilah teknis untuk sasaran bisnisnya. Tujuan *data mining* atau tujuan dilakukannya penelitian ini adalah sebagai berikut.

   1. Menghasilkan model *product matching book analysis* berdasarkan hasil review yang diberikan oleh *customer* sebelumnya
   2. Mengetahui pola (*pattern*) dari data buku sehingga diketahui kriteria atau buku untuk mendapatkan jenis buku yang serupa.

3. Mendukung proses analisis dan pengambilan keputusan dalam pembelian buku pada sebuah aplikasi e-commerce dengan menerapkan CRISP-DM
4. Membuat sistem rekomendasi yang lebih terstruktur sesuai dengan kriteria yang telah ditentukan sebelumnya

5. *Produce Project Plan*

   Pada tahapan ini akan dijelaskan rencana yang akan dilakukan untuk mencapai tujuan *data mining* dan untuk mencapai tujuan bisnis. Rencana yang dibuat harus dapat menentukan langkah-langkah yang akan dilakukan selama sisa proyek, termasuk pemilihan alat dan teknik awal. Dalam penelitian ini akan dilakukan teknik *k-means clustering* yang akan mengelompokkan buku-buku ke dalam jenis, ciri, dan/atau kriteria yang sama sehingga produk yang akan ditampilkan adalah semua produk pada *cluster* yang sama. Analisis *cluster* digunakan karena teknik tersebut memiliki tujuan utama untuk mengelompokkan objek berdasarkan ciri dan/atau karakteristiknya. Sedangkan penggunaan metode *k-means clustering* digunakan karena algoritma ini memiliki ketelitian yang cukup tinggi dan lebih terukur serta efisien dalam mengelola objek jumlah besar. Selain itu, algoritma *k-means* tidak terpengaruh terhadap urutan objek.

Data Understanding

Data Understanding merupakan tahap dimulainya pengumpulan data yang akan dilanjutkan dengan proses untuk memperoleh pemahaman yang mendalam mengenai data, mengidentifikasi kualitas data, serta memungkinkan untuk melakukan deteksi apabila terdapat sebuah bagian yang menarik dari data yang dapat digunakan sebagai hipotesis untuk informasi yang tersembunyi. Pada tahap ini, *data understanding* juga dapat memberikan fondasi analitik untuk sebuah proyek dengan membuat ringkasan (*summary*) dan melakukan identifikasi potensi masalah dalam data yang harus dilakukan secara cermat dan tidak terburu-buru. Jika ada masalah pada tahapan ini maka akan mengganggu pada tahap selanjutnya. Ringkasan dari data dapat berguna untuk mengkonfirmasi apakah data terdistribusi seperti yang diharapkan, atau mengungkapkan penyimpangan tak terduga yang perlu ditangani pada tahap selanjutnya, yaitu *Data Preperation* [1].

Dalam sistem yang akan dikembangkan oleh penulis akan melakukan 4 proses yang meliputi pengumpulan data atau *Collect Initial Data*, *Describe Data*, *Explore Data*, dan *Verify Data Quality* [2].

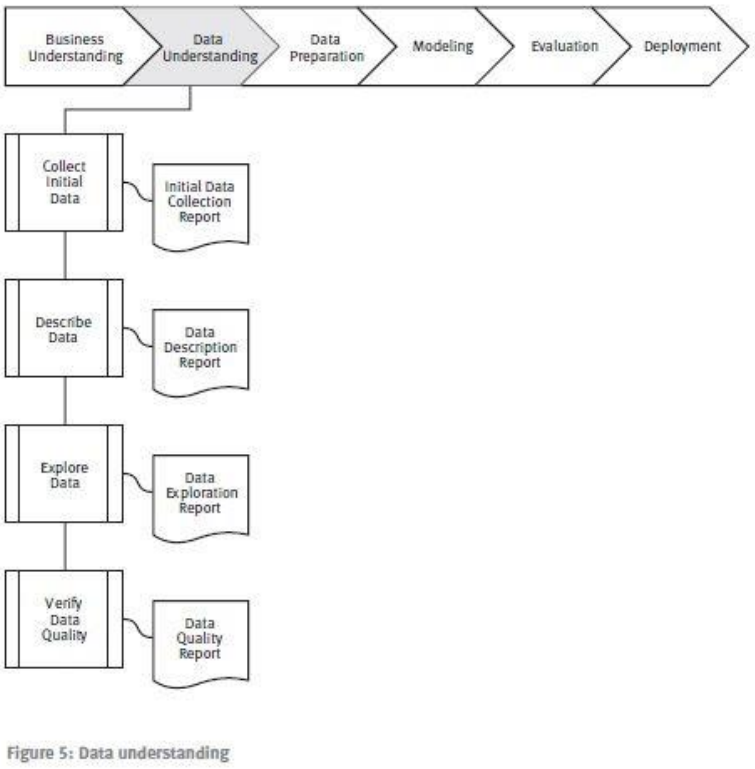


Figure 5: Data understanding

Gambar 2 Data Understanding

1. *Collect Initial Data*

   Tahapan ini merupakan tahap pengumpulan data yang bersumber dari goodreads book reviews 20k yang diambil dari book readers social network www.goodreads.com pada tahun 2017.

2. *Describe Data*

   Pada tahap ini dilakukan deskripsi terhadap data yang akan digunakan dapat berupa penjelasan mengenai format data, jumlah data, jumlah atribut dan fitur yang digunakan pada pengerjaan proyek. Data yang digunakan berjumlah 21,6k data dan terdapat 15 atribut didalamnya dengan 13 atribut bertipe data string dan 2 atribut bertipe data decimal. Pada tahapan deskripsi data ini juga dapat memberikan informasi apa saja yang dapat digunakan untuk melakukan implementasi pada sistem yang dibangun.

3. *Explore Data*

   Pada tahap ini dilakukan eksplorasi terhadap isi dari data yang digunakan. Eksplorasi data dilakukan untuk melihat kesesuaian data dengan tujuan dari dilakukannya proyek ini yaitu untuk melakukan implementasi *product matching* berdasarkan review buku. Untuk itu diperlukan data yang terkait review buku dapat berupa komentar ataupun rating yang diberikan serta judul buku atau mungkin perlu untuk mengikutsertakan *genre* buku tersebut.

4. *Verify Data Quality*

   Pada tahap ini dilakukan evaluasi dan kelengkapan data yang dapat berupa melakukan pemeriksaan terhadap data yang hilang atau atribut yang kosong serta seluruh isi dari data apakah dapat dimengerti dan sesuai dengan atributnya. Hasil yang didapatkan adalah sebagai berikut.

   1. Terdapat data yang *mismatched* atau tidak sesuai dengan jumlah yang besar pada atribut book_series dan book_series_url yang berjumlah sama yaitu 12,8k. Data yang tidak sesuai juga terdapat pada atribut review sejumlah 753.
   2. Terdapat data yang *mismatched* pada atribut reveiwe_name sejumlah 29 serta pada atribut reviewer_image dan reviewer_image_url sejumlah 25.
   3. Terdapat 2 data yang tidak sesuai pada atribut book_image dan book_image_url serta pada atribut genre terdapat 121 data yang tidak sesuai.

Referensi

[1] A. A. P. d. A. Purwarianti, "Prediksi Kinerja Penjualan Karya Musik Menggunakan Framework CRISP-DM (Studi Kasus: X Music Indonesia)," *Jurnal Sarjana Institut Teknologi Bandung bidang Teknik Elektro dan Informatika*, 2011.

[2] T. P. a. Y. C. I. Budiman, "Data Clustering Menggunakan Metodologi CRISP-DM Untuk Pengenalan Pola Proporsi Pelaksanaan Tridharma," *J. Sist. Inf. BISNIS*, 2014.