**Tugas Besar Penambangan Data**

**Data Mining**

Oleh:

Muhammad Hafidh Raditya (NIM 1301184079)

IF-42-GAB02

**Logo, company name

Description automatically generated**

**Program Studi S1 Informatika  
Fakultas Informatika  
Universitas Telkom  
Bandung  
2021**

1. **Deskripsi Tugas Besar**

Tugas besar data mining kali ini berasal dari Data Mining Cup 2021 yang bertujuan untuk membuat sistem rekomendasi pada sebuah toko buku. Pada tugas kali ini diberikan beberapa dataset yang bebas digunakan untuk membuat model rekomendasi. Terdapat 3 dataset yang diberikan yaitu dataset items yang berisi kumpulan data dari buku yang terdapat pada toko buku tersebut, dataset transactions yaitu kumpulan data transaksi pada toko buku tersebut selama 3 bulan kebelakang, dan dataset evaluation yaitu kumpulan data buku yang akan dibuat sistem rekomendasinya. Dataset yang diberikan akan dijelaskan pada tabel berikut.

|  |  |  |
| --- | --- | --- |
| **Dataset** | **Kolom** | **Deskripsi** |
| Items.csv | itemID | ID buku. Berformat integer positif. |
| title | Judul buku. Berformat string. |
| author | Penulis buku. Berformat string. |
| publisher | Penerbit buku. Berformat string. |
| main topic | Topik utama buku. Berformat string. |
| subtopics | Subtopik buku, berformat list of string. |
| Transactions.csv | itemID | ID produk/buku yang terdapat pada suatu transaksi. Berformat integer positif. |
| sessionID | ID transaksi. Berformat integer positif. |
| click | Jumlah suatu produk diklik dalam satu kali transaksi. Berformat integer. |
| basket | Jumlah suatu produk dimasukkan ke dalam keranjang dalam satu kali transaksi. Berformat integer. |
| order | Jumlah suatu produk dipesan dalam satu kali transaksi. Berformat integer |
| Evaluation.csv | itemID | ID buku yang akan dibuat rekomendasinya. Berformat integer positif. |

1. **Proses Pengerjaan**

Tubes ini saya kerjakan menggunakan bahasa pemrograman python dan jupyter lab sebagai *tools*nya. Pendekatan yang saya gunakan adalah *clustering*. Proses pengerjaan tubes ini dibagi menjadi dua tahap besar yaitu praproses data dan tahap *modelling*. Tiap tahap akan dijelaskan pada sub-babnya masing-masing.

1. **Praproses Data**

Pada tahap ini saya melakukan pengolahan data mentah menjadi data siap pakai untuk dilakukan *clustering*. Karena *clustering* merupakan suatu proses pengelompokkan data kedalam beberapa kelompok kecil berdasarkan kemiripan datanya, maka yang akan saya telaah kemiripannya adalah data buku berdasarkan beberapa atribut yang dimiliki oleh suatu buku. Oleh karena itu saya hanya akan menggunakan dan mempraproses dataset items.csv untuk proses pengerjaan tubes kali ini. Sedangkan dataset transactions.csv tidak akan saya gunakan karena dataset tersebut lebih cocok jika digunakan pendekatan analisis asosiasi.

Secara garis besar, praproses data pada pengerjaan tubes kali ini dapat dijelaskan dengan diagram alir berikut.

Diagram

Description automatically generated

Langkah pertama yang saya lakukan tentu saja adalah import dataset kedalam jupyter lab. Dataset mentah memiliki total record data sebanyak 78.030 baris data. Berikut adalah 5 data teratas pada dataset mentah.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **itemID** | **title** | **author** | **publisher** | **main topic** | **subtopics** |
| 21310 | Princess Poppy: The Big Mix Up | Janey Louise Jones | Penguin Random House Children's UK | YFB | [5AH] |
| 73018 | Einfach zeichnen! Step by Step | Wiebke Krabbe | Schwager und Steinlein | AGZ | [5AJ, AGZ, WFA, YBG, YBL, YNA, YPA] |
| 19194 | Red Queen 1 | Victoria Aveyard | Orion Publishing Group | YFH | [5AP, FBA] |
| 40250 | Meine Kindergarten-Freunde (Pirat) | NaN | Ars Edition GmbH | YB | [5AC, 5AD, YBG, YBL, YF] |
| 46107 | Mein großes Schablonen-Buch - Wilde Tiere | Elizabeth Golding | Edition Michael Fischer | WFTM | [WD, WFTM, YBG, YBL, YBLD, YBLN1] |

Setelah data saya import saya melakukan proses data cleansing. Lalu saya menghapus data buku yang memiliki judul terlalu pendek, yaitu judul buku yang memiliki karakter dibawah 4 karakter. Hal ini saya lakukan karena jika suatu judul buku terlalu pendek, maka tidak akan bisa dideteksi bahasanya oleh mesin. Setelah itu saya mencari record data yang kosong atau memiliki null value, berikut hasilnya.

|  |  |
| --- | --- |
| **Kolom** | **Jumlah record bernilai null** |
| itemID | 0 |
| title | 0 |
| author | 3237 |
| publisher | 9 |
| main topic | 258 |
| subtopics | 1 |

Setelah melihat jumlah record data yang memiliki null value, saya memutuskan untuk menghapus semua record null value pada kolom publisher, main topic, dan subtopics. Sedangkan null value pada kolom author tidak akan langsung saya hapus karena memiliki terlalu banyak null value. Null value pada kolom author ini akan saya proses nanti pada tahap *feature* *engineering*.

Setelah proses data cleansing selesai, tahap selanjutnya adalah proses deteksi bahasa. Saya menggunakan *package* langdetect yang disediakan oleh bahasa python untuk melakukan proses deteksi bahasa. Alasan saya melakukan proses deteksi bahasa adalah karena saya juga akan mencari kemiripan data buku berdasarkan bahasanya juga. Saya melakukan deteksi bahasa berdasarkan judul buku. Setelah saya melakukan proses deteksi bahasa, saya mendapatkan kalau dari seluruh data buku, terdapat 31 bahasa yang berbeda. Saya membuat satu kolom baru yaitu kolom language untuk menampung data bahasa yang sudah dideteksi.

Yang terakhir adalah proses *feature* *engineering*. *Feature* atau kolom yang akan saya lakukan *engineering* adalah kolom main topic, language, dan author. Yang pertama adalah kolom main topic. Berdasarkan deskripsi tugas yang diberikan oleh pihak data mining cup 2021, kolom main topic berisi string yang terdiri dari 1-4 karakter. Karakter pertama menyatakan topik utama buku. Dua karakter pertama menyatakan sub-topik buku, tiga karakter pertama menyatakan sub-sub-topik buku, dan seterusnya. Untuk menghemat waktu, yang akan saya proses adalah sampai dua karakter pertamanya saja, yaitu untuk topik utama dan sub-topik buku. Sedangkan sub-sub-topik buku tidak akan saya proses. Topik utama saya masukkan ke dalam kolom baru bernama general\_topic, dan sub-topik saya masukkan ke dalam kolom baru bernama sub\_topic.

Kolom selanjutnya yang saya lakukan proses *feature engineering* adalah kolom language. Sebelum saya lakukan proses *engineering* saya akan lihat terlebih dahulu distribusi datanya. Berikut adalah distribusi data pada kolom language.

Chart, bar chart

Description automatically generated

Berdasarkan grafik distribusi datanya, dari 31 bahasa berbeda yang terdapat pada dataset, dapat dilihat kalau bahasa yang mendominasi hanya Bahasa Inggris dan Bahasa Jerman., sedangkan bahasa lain sangat sedikit jumlahnya dan distribusinya merata. Oleh karena ini saya hanya akan menyeleksi 5 bahasa teratas, sedangkan sisanya akan saya labeli ‘otherlanguage’.

Terakhir adalah kolom author. Kolom author memiliki unique value sebanyak 35.765 baris data. Artinya terdapat 35.765 author berbeda. Sama seperti sebelumnya, saya akan melihat dulu distribusi datanya. Berikut adalah distribusi data pada kolom author.

Chart, bar chart

Description automatically generated

Berdasarkan grafik distribusi datanya, dapat dilihat kalau dari sekian banyak data author yang berbeda, hanya terdapat satu author yang mendominasi datanya yaitu author Garcia Santiago, namun untuk author lain dibawahnya, distribusi datanya tersebar sangat merata, yaitu masing-masing author memiliki kisaran data dari 0 sampai 200 baris data. Dengan mempertimbangkan jumlah author berbeda dan distribusi datanya, maka saya memutuskan untuk hanya akan menyeleksi 50 author teratas untuk tetap dimasukkan ke dalam dataset, sedangkan yang tidak termasuk 50 author teratas akan saya labeli dengan ‘otherauthor’.

Dengan berakhirnya proses *feature engineering*, maka tahap praproses data sudah selesai dan akan dilanjutkan ke tahap selanjutnya yaitu tahap *modelling*. Jumlah baris data pada dataset hasil praproses sekarang hanya tersisa 77.557 baris data, berkurang 473 baris data dari dataset mentah sebelum dilakukan praproses. Berikut adalah 5 baris teratas pada dataset hasil praproses.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **itemID** | **title** | **author** | **publisher** | **language** | **general\_topic** | **sub\_topic** |
| 21310 | Princess Poppy: The Big Mix Up | Janey Louise Jones | Penguin Random House Children's UK | english | children | childrendiction |
| 73018 | Einfach zeichnen! Step by Step | Wiebke Krabbe | Schwager und Steinlein | german | arts | artstreatment |
| 19194 | Red Queen 1 | Victoria Aveyard | Orion Publishing Group | english | children | childrendiction |
| 40250 | Meine Kindergarten-Freunde (Pirat) | otherauthor | Ars Edition GmbH | german | children | childrenpicture |
| 46107 | Mein großes Schablonen-Buch - Wilde Tiere | Elizabeth Golding | Edition Michael Fischer | german | lifestyle | lifestyledecoration |

1. **Modelling**

Pada tahap ini saya memakai pendekatan *clustering* lalu dilanjutkan dengan proses menyeleksi 5 rekomendasi buku berdasarkan kemiripannya menggunakan *package* sigmoid\_kernel yang sudah disediakan oleh bahasa python. Sebenarnya bisa saja saya langsung menggunakan sigmoid\_kernel untuk menghitung nilai kemiripannya pada seluruh data, namun karena jumlah datanya yang sangat besar yaitu terdapat 77.557 baris data, diperlukan sebesar 30 GB memori untuk menampung nilai kemiripannya ke dalam suatu array. Sedangkan *hardware* yang saya gunakan hanya memiliki memori sebesar 8 GB. Oleh karena itu saya melakukan *clustering* terlebih dahulu menggunakan KMeans *clustering* untuk mengelompokkan data kedalam beberapa *cluster* atau kelompok kecil sehingga nantinya tiap data yang ada pada dataset evaluation.csv hanya akan dibandingkan oleh data lain yang berada dalam satu *cluster* yang sama. Dengan proses seperti ini maka penggunaan memori dapat lebih diefektifkan, dengan besar memori yang dibutuhkan sekarang hanya sebesar 4 GB. Tahap *modelling* pada pengerjaan tubes kali ini dapat dijelaskan pada diagram alur berikut.

Diagram

Description automatically generated

Tahap pertama yang saya lakukan adalah *feature selection*. Saya memutuskan untuk melakukan *clustering* dan penghitungan nilai kemiripan berdasarkan kemiripan beberapa atribut dari suatu buku, yaitu atribut author, language, general\_topic, dan sub\_topic. Lalu atribut title dan publisher akan saya hapus karena tidak akan saya pakai dalam tahap *modelling* ini.

Selanjutnya tahap *clustering*. Pada tahap ini saya menentukan untuk membuat 20 *cluster*. Sebenarnya semakin banyak *cluster* akan semakin bagus karena akan semakin mudah untuk menyeleksi 5 buku rekomendasi. Namun sekali lagi karena keterbatasan *hardware* yang saya miliki maka saya hanya akan menggunakan 20 *cluster*.

Terakhir tahap penghitungan nilai kemiripan menggunakan *package* sigmoid\_kernel yang sudah disediakan oleh bahasa python, lalu seleksi 5 buku paling mirip untuk dijadikan buku rekomendasi. Untuk setiap itemID pada dataset evaluation.csv, akan dicek ke dataset items.csv yang tadi sudah saya lakukan praproses. Jika itemID pada evaluation.csv ditemukan pada items.csv, maka akan dilakukan proses penghitungan nilai kemiripannya berdasarkan *cluster* yang sudah dibuat pada tahap *clustering*. Jika itemID pada evaluation.csv tidak ditemukan pada items.csv, maka 5 buku rekomendasi akan dipilih secara acak. itemID pada evaluation.csv memungkinkan untuk tidak ditemukan pada items.csv dikarenakan pada tahap praproses data sebelumnya saya melakukan penghapusan data sebanyak 473 baris data. Secara garis besar, tahap ini dapat dijelaskan melalui diagram alir berikut.

Diagram

Description automatically generated

Proses *modelling* seperti ini memiliki kelemahan yaitu proses ini memakan waktu sebanyak 15 jam dalam *running* programnya. Hal ini dikarenakan proses penghitungan nilai kemiripan menggunakan sigmoid\_kernel memakan waktu yang sangat lama karena dimensi datanya yang sangat besar, sebanyak 77.000 baris data.

Dengan ini tahap *modelling* telah selesai dan telah menghasilkan 5 buku untuk direkomendasikan. Berikut adalah 5 data buku teratas dan buku-buku rekomendasinya.

|  |  |
| --- | --- |
| **itemID** | **Book recommandations** |
| 12 | 32729, 2583, 10048, 41535, 69791 |
| 10104 | 10143, 22900, 41320, 47055, 57759 |
| 41371 | 41371, 21540, 49279, 13702, 38622 |
| 14015 | 14015, 47952, 19563, 14015, 40294 |
| 28515 | 28515, 52767, 5900, 73176, 67536 |

Sebagai contoh saya akan melihat salah satu baris data hasil rekomendasi apakah model yang saya buat sudah tepat atau belum. Saya akan mengambil record pertama sebagai contoh. Berikut adalah hasilnya.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **itemID** | **title** | **author** | **language** | **General\_topic** | **Sub\_topic** |
| 12 | Breathtaking | Reva Ryann Thompson | english | fiction | fictionfantasy |
| 32729 | The Angel and the Paintbrush | M. J. Thompson | english | fiction | fictionfantasy |
| 2583 | Letters from America | P. Thompson | english | fiction | fictionfantasy |
| 10048 | Voidoxity | James L. P. Thompson | otherlanguage | fiction | fictionfantasy |
| 41535 | Voidoxity | James L. P. Thompson | otherlanguage | fiction | fictionfantasy |
| 69791 | Sylvow | Eibonvale Press | otherlanguage | fiction | fictionfantasy |

Dari tabel, yang saya *highlight* warna merah adalah buku pada evaluation.csv untuk dicari buku rekomendasinya. Sedangkan yang tidak di*highlight* adalah buku rekomendasinya. Jika dilihat pada tabel diatas semua buku sudah memiliki kemiripan pada topicnya, hanya saja bahasanya masih ada yang berbeda. Selain itu terdapat buku yang identik, namun berbeda itemIDnya. Menurut saya data duplikat ini masih dapat diterima asalkan itemIDnya tidak duplikat juga.

1. **Kesimpulan**

Kesimpulan yang didapat pada pengerjaan tubes kali ini adalah pendekatan *clustering* dan penghitungan nilai kemiripan saya rasa cukup efektif dalam menentukan rekomendasi buku. Hanya saja untuk penelitian selanjutnya saya rasa saya harus menentukan skala prioritas atribut untuk dihitung nilai kemiripannya. Sebagai contoh atribut language seharusnya lebih diprioritaskan dibandingkan dengan atribut general\_topic. Dan atribut general\_topic seharusnya lebih diprioritaskan dibandingkan dengan atribut sub\_topic. Hal ini dikarenakan terdapat buku rekomendasi yang memiliki bahasa yang berbeda dengan buku utamanya. Hal ini tidak dapat diterima jika pada kasus nyata, karena pembeli tidak akan dapat membaca buku rekomendasi tersebut jika bahasanya berbeda.

Selanjutnya soal waktu *running* program yang sangat lama, yaitu sebanyak 15 jam. 15 jam adalah waktu yang sangat lama untuk membuat sistem rekomendasi. Oleh karena itu untuk penelitian selanjutnya saya akan menggunakan pendekatan analisis asosiasi menggunakan data transaksi yang sudah disediakan. Analisis asosiasi menggunakan algoritma apriori, fp growth, dll seharusnya dapat menghasilkan solusi yang lebih cepat jika dibandingkan dengan sigmoid\_kernel, atau algoritma penghitungan nilai kemiripan lain yang membutuhkan proses yang sangat lama karena dimensi datanya yang terlalu besar.