Kartika Rizqi Nastiti 21/PPA/485636/06220

UAS Sistem Temu Balik Informasi (STBI) Rabu, 14 Desember 2022

- 1. Metode kueri pencarian dan *similarity search* pada proyek kami secara dijelaskan dalam tahap-tahap berikut ini:
 - a. Data kolom FS_ANAMNESA dan FS_DIAGNOSA yang sudah melalui tahap preprocessing kemudian dibentuk menjadi daftar token untuk tiap rekam medis.
 - b. Dilakukan pembentukan *Vector Space Model* dengan representasi dalam matriks TF-IDF untuk menghitung bobot tiap kata. *Term Frequency Inverse Document Frequency* (TF-IDF) merupakan salah satu cara untuk menghitung bobot kata berdasarkan frekuensi kemunculannya di dalam sebuah korpus. Algoritma TF-IDF digunakan untuk mengevaluasi pentingnya kata-kata dalam korpus tekstual. TF mewakili frekuensi kata, menunjukkan berapa kali mereka muncul dalam korpus persamaan. Bobot tiap kata dihitung dari kemunculan kata tersebut dibandingkan dengan jumlah total kata yang ada dalam korpus. Formula untuk mendapatkan nilai tiap kata untuk tiap dokumen rekam medis adalah sebagai berikut.

$$tf(t,d) = \frac{f_d(t)}{\max f_d(w)}$$

$$idf(t,D) = ln(\frac{|D|}{|\{d \in D: t \in d\}|})$$

$$tfidf(t, d, D) = tf(t, d) \cdot idf$$

Di mana:

 $f_d(t) :=$ frekuensi kemunculan kata t dalam dokumen D

D := korpus dokumen

IDF menunjukkan tingkat kepentingan sebuah kata dalam korpus secara keseluruhan. Nilai IDF didapatkan dengan menghitung logaritma dari nilai *inverse* proporsi dokumen yang mengandung kata tertentu. Nilai proporsi diambil dari jumlah dokumen dalam korpus dibagi dengan jumlah

- dokumen yang mengandung kata tertentu. Nilai logaritma dari proporsi tersebut merupakan nilai IDF. Bobot TF-IDF dihitung dengan mengalikan dua nilai tersebut. Semakin besar nilai TF-IDF, menunjukkan semakin signifikan kata yang bersangkutan di dalam korpus.
- c. Setelah tiap kata dari tiap dokumen direpresentasikan dalam bentuk matriks, tahap selanjutnya adalah proses untuk menerima input kueri dari user. User dapat memilih apakah pencarian dilakukan berdasarkan kolom FS ANAMNESA atau FS DIAGNOSA.
- d. Selanjutnya adalah proses pencarian berdasarkan nilai kemiripan antara kueri dan dokumen rekam medis. Proses penghitungan nilai kemiripan ini dihitung menggunakan nilai cosine. Nilai kemiripan untuk tiap term pada masing-masing dokumen didasarkan pada nilai dua buah vektor dalam ruang dimensi yang didapatkan dari cosinus pada sebuah sudut. Cosine score yang bernilai 0 mempunyai makna bahwa kedua buah vektor berada pada sumbu 90 derajat secara ortogonal dan tidak memiliki kecocokan, sedangkan cosine score yang semakin dekat dengan nilai 1 maka semakin kecil sudut dan semakin besar kecocokan antara 2 vektor. Dua buah vektor dikatakan mirip ketika nilai pada cosine similarity-nya adalah 1. Berikut adalah formula cosine similarity yang kami terapkan.

$$Sim(\alpha) = \frac{A \cdot B}{|A||B|} = \frac{\sum_{i=1}^{n} A1 \cdot B1}{\sqrt{\sum_{i=1}^{n} (A1)^{2} \cdot \sum_{i=1}^{n} (B1)^{2}}}$$

Di mana:

A = Vektor Dokumen

B = Vektor Kueri

A . B = Perkalian antara vektor A dan vektor B

|A| = Panjang vektor A

|B| = Panjang vektor B

|A||B| = Cross Product antara |A| dan |B|

 α = Sudut yang terbentuk antara vektor A dan B

e. 10 dokumen rekam medis dengan nilai *cosine* (kemiripan) yang paling tinggi akan dikembalikan kepada user sebagai output dari kueri yang sudah dimasukkan. Berikut adalah contoh hasil keluaran untuk kueri "dermatitis".

Search Rank	Search query >> dermatitis Rank Scores Document number ANAMNESA TINDAKAN TERAPI CATATAN FISIK DIAGNOSA						
	Scores	Document Number	FS_ANAMNESA	FS_TINDAKAN	FS_TERAPI	FS_CATATAN_FISIK	FS_DIAGNOSA
Rankin	g						
1	0.436	98	GATAL DI PAHA makin banyak 1th <	NaN	R/ Hexilon 8 mg tab no X S 1-0-0&nb	Suhu : C,Nadi : x/menit,Respirasi : x/menit	Dermatitis Numular
2	0.436	97	KONTROL sempat membaik	NaN	R/ Hexilon 8 mg tab no XV S 1-0-0&n	Suhu : C,Nadi : x/menit,Respirasi : x/menit	Dermatitis Numular
3	0.436	806	-gatal di kaki	NaN	R/ IKADERM CREAM gr 10 br /> BA	Suhu : - C,Nadi : - x/menit,Respirasi : - x	Dermatitis Numular
4	0.436	805	gatal di kaki kambuh	NaN	R/ IKADERM CREAM gr 10 br /> BA	Suhu : C,Nadi : x/menit,Respirasi : x/menit	Dermatitis Numular
5	0.436	804	gatal2 dikaki sdh 1 bln	NaN	R/ IKADERM CREAM gr 10 br /> BA	Suhu : C,Nadi : x/menit,Respirasi : x/menit	Dermatitis Numular
6	0.436	143	gatal di tangan dan kaki kambuh	NaN	R/ HEXILON 8MG tab no X	Suhu : C,Nadi : x/menit,Respirasi : x/menit	Dermatitis Numular
7	0.436	301	gatal2 di kaki	NaN	R/ Desoxymetason gr 15 Fuladi	Suhu : C,Nadi : x/menit,Respirasi : x/menit	Dermatitis Numular
8	0.410	207	GATAL DI KAKI	NaN	R/ TERMISIL CR gr 10 <	Suhu : C,Nadi : x/menit,Respirasi : x/menit	Dermatitis Atopik &p> Derma
9	0.399	246	KONTROL, membaik	NaN	R/ Metylprednisolon 8 mg tab no X	Suhu : C,Nadi : x/menit,Respirasi : x/menit	>Dermatitis Atopik
10	0.399	247	qatal di badan 2 th terakhir, kambuh2a	NaN	R/ Metylprednisolon 8 mg tab no X	Suhu : C,Nadi : x/menit.Respirasi : x/menit	Dermatitis Atopik

- 2. Metode evaluasi yang kami gunakan dalam proyek STBI adalah dengan penghitungan nilai *precision* dan *recall* yang merupakan salah satu metode evaluasi pada sistem temu balik yang *feasible* untuk dilakukan. Berikut adalah detail proses evaluasi yang kami lakukan.
 - a. Menentukan masing-masing 10 kueri input untuk kolom FS_ANAMNESA dan FS_DIAGNOSA. Berikut adalah rincian banyak dokumen yang dikembalikan untuk tiap kueri.

No	FS_ANAMNESA	Jumlah Dokumen	FS_DIAGNOSA	Jumlah Dokumen
1.	kontrol	10	hamil	10
2.	gigi sakit	10	bronkitis akut	10
3.	ca paru	10	neurosa	10
4.	nyeri bahu kanan	6	dermatitis	10
5.	nyeri	4	et	10
6.	telinga nyeri	4	bph	10
7.	lutut kanan sakit	4	abses gigi	10
8.	lutut kiri sakit	4	meningioma	10
9.	batuk pilek	2	kb	4
10.	batuk	2	rhinitis	4

b. Menghitung nilai *precision* dan *recall* untuk masing-masing kueri dengan rumus sebagai berikut.

$$Precision = rac{Jumlah\ dokumen\ Relevan\ yang\ terambil}{Jumlah\ dokumen\ yang\ terambil}$$

$$Recall = rac{Jumlah\ dokumen\ Relevan\ yang\ terambil}{Jumlah\ dokumen\ Relevan}$$

c. Dilakukan penghitungan nilai rata-rata untuk *precision* dan *recall*. Berikut adalah hasil nilai rata-rata dari keseluruhan kueri.

	Average Precision	Average Recall
FS_ANAMNESA	0.522	0.816
FS_DIAGNOSA	0.574	0.781
Average	0.548	0.796

Precision digunakan untuk mengetahui kemampuan model dalam mendapatkan hasil yang sesuai dengan kueri yang diinputkan. Sedangkan *recall* digunakan untuk mendapatkan apa saja informasi yang mengandung kueri yang dimasukkan. Dari proses *retrieve* yang dilakukan didapatkan nilai *precision* sebesar 54,8% dan *recall* sebesar 79,6%. Hasil evaluasi dengan *low precision* dan *high recall* menunjukkan bahwa model dapat memprediksi lebih banyak t*rue positive*.

3. Sistem rekomendasi merupakan sistem yang bekerja untuk memberikan rekomendasi dengan berbagai pendekatan, dapat berdasarkan tingkat popularitas (popularity-based), preferensi pengguna yang bersifat personal (collaborative filtering), berdasarkan konten pencarian (content-based), atau gabungan dari berbagai pendekatan (hybrid) dari sekumpulan data yang besar. Sementara sistem tanya jawab merupakan sistem untuk menjawab pertanyaan-pertanyaan dari user yang dijawab berdasarkan aturan-aturan yang sudah diprogram sebagai sebuah knowledge base.

Berdasarkan metodologi yang kami terapkan, proyek STBI kami termasuk ke dalam sistem rekomendasi, yang mana salah satu pendekatannya adalah content-based filtering dan tidak menggunakan aturan atau knowledge base tertentu untuk hasil keluarannya. Pendekatan ini kami gunakan dalam representasi TF-IDF dan penghitungan cosine similarity. Dokumen rekam medis yang memiliki tingkat kemiripan tertinggi dengan kueri yang dimasukkan oleh pengguna, maka itulah hasil rekomendasi pencarian dokumen rekam medis.

- 4. Rangkuman dan konsep multimedia.
 - a. Berikut adalah rangkuman lengkap dari proyek STBI yang kami kembangkan.

Introduction

Pertumbuhan dokumen rekam medis semakin meningkat seiring waktu dan semakin beragamnya jenis penyakit dan terapi yang diperlukan. Namun, hal ini belum diimbangi dengan proses pencarian yang efektif dan efisien. Penelitian ini bertujuan untuk menangani permasalahan pencarian yang seringkali membutuhkan waktu yang lama dengan hasil pencarian yang belum tentu sesuai harapan, dengan membangun model pencarian dokumen rekam medis dengan metode *Vector Space Model* (VSM) dan TF-IDF. Metode VSM memungkinkan hasil *retrieval* yang tidak harus sama persis dengan kueri pencarian yang dimasukkan oleh pengguna, namun diharapkan tetap memberikan hasil yang relevan dengan kebutuhan yang diinginkan pengguna. Intuisi di balik model ini adalah bahwa setiap vektor mengkodekan arti dari objek yang sesuai dan untuk istilah yang dekat secara semantik, maka ukuran kemiripan antara vektor mereka direpresentasikan dalam nilai yang tinggi [1].

Methods

Metodologi pengembangan proyek STBI kami dimulai dengan mengambil 1000 data rekam medis pada kolom FS_ANAMNESA dan FS_DIAGNOSA, diikuti dengan tahap *preprocessing* yang terdiri dari penghapusan baris kosong, *lowercase*, menghapus tanda baca, *tag* HTML, *stopwords*, spasi berlebih antar kata, serta melakukan normalisasi kata *typo*.

Setelah data dibersihkan, lalu dilakukan penghitungan nilai *Vector Space Model* dengan representasi berupa pembentukan matriks TF-IDF berdasarkan frekuensi kemunculan tiap fitur kata. Algoritma TF-IDF digunakan untuk mengevaluasi pentingnya kata-kata dalam korpus tekstual [2]. Setelah itu dilakukan penghitungan nilai kemiripan kueri pencarian dengan dokumen rekam medis berdasarkan formula *cosine similarity*. Cosine score yang bernilai 0 mempunyai makna bahwa kedua buah vektor berada pada sumbu 90 derajat secara ortogonal dan tidak memiliki kecocokan, sedangkan *cosine score* yang semakin dekat dengan nilai 1 maka semakin kecil sudut dan semakin besar kecocokan antara 2 vektor [3]. Dua buah vektor dikatakan mirip ketika nilai pada *cosine*

similarity-nya adalah 1. Hasil retrieval adalah keseluruhan kolom dari tiap dokumen rekam medis yang sudah ada dan diurutkan berdasarkan 10 baris dengan nilai kemiripan tertinggi.

Proses evaluasi dilakukan dengan terlebih dahulu menentukan masing-masing 10 kueri pencarian untuk kolom FS_ANAMNESA dan FS_DIAGNOSA. Setelah itu dilakukan penghitungan rata-rata nilai precision dan recall untuk kedua kolom tersebut.

Results and Discussion

Dari proses *retrieve* yang dilakukan didapatkan nilai *precision* sebesar 54,8% dan *recall* sebesar 79,6%. Hasil evaluasi dengan *low precision* dan *high recall* menunjukkan bahwa model dapat memprediksi lebih banyak *true positive*.

Hasil *retrieval* dan nilai evaluasi sangat bergantung pada banyak dokumen di dalam korpus dan kueri yang dimasukkan, sehingga sangat mungkin banyak dokumen yang nantinya di-*retrieve* dan nilai *precision* dan *recall* bisa saja berubah jika ada penyesuaian jumlah dokumen dan penggunaan kueri pencarian lainnya.

References

- [1] R. Shaptala, A. Kyselova, and G. Kyselov, "Exploring the vector space model for online courses," 2017 IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON), 2017.
- [2] M. Chiny, M. Chihab, O. Bencharef, and Y. Chihab, "Netflix Recommendation System based on TF-IDF and Cosine Similarity Algorithms," *Proceedings of the 2nd International Conference on Big Data, Modelling and Machine Learning*, 2021.
- [3] J. Han, M. Kamber, and J. Pei, *Data mining: Data mining Concepts and Techniques*, 3rd edition. 225 Wyman Street, Waltham, MA 02451, USA: Morgan Kaufmann, 2012.
- b. Proyek STBI kami bukan merupakan multimedia information retrieval system karena kami hanya memproses data dan mempresentasikan informasi dalam satu jenis media saja, yaitu teks dokumen rekam medis. Hal ini tentunya berbeda dengan definisi "multimedia" yang mengandung lebih dari 1 jenis media, seperti kombinasi antara teks dengan gambar, atau audio dengan gambar, dan seterusnya.