1. Metode kueri pencarian dan similarity search pada proyek pengembangan STBI kelompok saya

Query pencarian pada kelompok kami hanya memasukkan sebuah keyword / kata kunci medis yang berkaitan dengan anamnesa dan diagnose.

Similarity search yang digunakan adalah menerapkan TF-IDF, VSM, dan cosine similarity. Pada awalnya kami membuat sebuah matrix TF-IDF untuk masing-masing unique word pada kolom FS\_ANAMNESA dan FS\_DIAGNOSA. Matrix TF-IDF yang terbentuk adalah untuk sumbu x nya merupakan dokumen-dokumen yang ada dan sumbu y nya merupakan unique word. Setelah matrix tersebut terbentuk, maka yang dilakukan adalah mencari unique word yang intersection dengan query. Setelah itu dilakukan penghitungan cosine similarity antara query dengan masing-masing word yang terdapat pada setiap dokumen. Dari penghitungan cosine similarity, didapatkan cosine similarity score yang diinterpretasikan menggunakan vector space model.

1. Metode evaluasi yang digunakan untuk mengevaluasi prototype STBI kelompok saya

Evaluasi pada kelompok saya menggunakan average precision dan average recall.

1. Average precision merupakan sebuah evaluasi yang digunakan untuk melihat ketepatan dari jumlah dokumen yang dapat ditemukan dan dianggap relevan oleh proses pencarian. Precision didapatkan dengan membagi jumlah dokumen relevan yang diambil dengan jumlah dokumen yang terambil. Jumlah dokumen relevan yang terambil merupakan dokumen-dokumen yang mempunyai cosine similarity score > 0. Jumlah dokumen yang terambil merupakan jumlah dokumen yang berhasil diambil sesuai dengan yang sudah ditentukan. Kalau dalam kelompok kami menggunakan 10 dokumen teratas. Untuk setiap query dihitung precisionnya kemudian dijumlahkan. Dari akumulasi precision yang didapatkan dibagi dengan banyaknya query yang diinputkan sehingga idapatkanlah average precision.
2. Average recall merupakan sebuah evaluasi yang digunakan untuk melihat rasio jumlah dokumen relevan yang ditemukan dengan total jumlah dokumen dalam koleksi yang dianggap relevan. Recall didapatkan dengan membagi jumlah dokumen relevan yang diambil dengan jumlah dokumen yang relevan. Jumlah dokumen relevan yang terambil merupakan dokumen-dokumen yang mempunyai cosine similarity score > 0. Jumlah dokumen yang relevan merupakan banyaknya dokumen yang mempunyai cosine similarity score > 0. Untuk jumlah dokumen yang relevan merupakan jumlah dokumen total, tidak hanya yang terambil saja. Jadi semisal dokumen yang diambil 10, tetapi jumlah dokumen relevannya ada 15 maka yang digunakan adalah 15. Dari akumulasi recall yang didapatkan dibagi dengan banyaknya query yang diinputkan sehingga idapatkanlah average recall.

Untuk hasil evaluasi yang didapatkan dari kelompok kami, didapatkan hasil average precision yang masih rendah yakni 54.8% sedangkan average recallnya didapatkan 79.6%. Hasil retrieval dan nilai evaluasi bergantung pada banyaknya dokumen dalam korpus, query yang dimasukkan, dan kualitas data pada dataset yang digunakan. Banyaknya dokumen pada kelompok kami masih cukup rendah, yakni dengan jumlah 1000 data yang digunakan. Untuk preprocessing querynya hanya menggunakan tokenisasi. Dan kualitas data pada dataset terbilang rendah, dikarenakan tidak adanya standar penulisan pada masing-masing kolom, masih terdapat banyak typo, dan banyaknya tag html.

1. Perbedaan system rekomendasi dan system tanya jawab adalah
   1. Sistem rekomendasi adalah sebuah system yang bertujuan untuk memberikan informasi yang menarik dan bagi user dan membantu dalam menentukan pilihannya. Biasanya system rekomendasi ini hanya memberikan pilihan saja, bukan untuk menjawab pertanyaan yang diajukan oleh user. Sedangkan system tanya jawab adalah sebuah system yang memberikan jawaban secara pasti atas pertanyaan yang diajukan oleh user. User tidak perlu mempertimbangkan lagi atas jawaban yang diberikan oleh system.

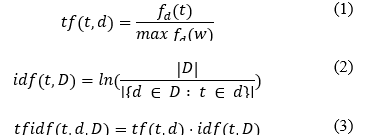
Sistem kami termasuk ke dalam system rekomendasi, dikarenakan system akan memberikan banyak jawaban yang berhubungan dengan query yang dimasukkan. User masih perlu mempertimbangkan atas jawaban/hasil yang diberikan oleh system.

1. Jawaban
   1. Rangkuman dari proyek STBI
      1. Introduction
         1. Pertumbuhan dokumen rekam medis berkembang dengan seiring waktu dan semakin beragam jenis penyakit dan terapi yang diperlukan. Namun hal tersebut kurang diimbangi dengan proses pencarian yang efektif dan efisien. Menurut Peraturan Kementerian Kesehatan Indonesia nomor 129/Menkes/SK/II/2008 menjelaskan bahwa salah satu standar minimum dalam sebuah pelayanan di fasilitas kesehatan adalah dapat menyediakan rekam medis dalam waktu 10 menit dan mempunyai kelengkapan informasi sebesar 100%.
         2. IR adalah sebuah sebuah proses mendapatkan informasi yang relevan dan memanipulasi data besar. Informasi dan data yang dicari ataupun diolah bisa berupa text, gambar, audio, video, dan objek multimedia lainnya.
         3. VSM adalah model aljabar untuk merepresentasikan sebuah objek sebagai vector.
         4. Pengukuran kemiripan pada VSM yang popular adalah cosine similarity
         5. Cara untuk mengkodekan dokumen tekstual ke dalam vektor adalah *Term Frequency - Inverse Document Frequency*
         6. *Precision* adalah rasio jumlah dokumen relevan yang ditemukan dengan total jumlah dokumen yang ditemukan
         7. *Recall* adalah rasio jumlah dokumen relevan yang ditemukan dengan total jumlah dokumen dalam koleksi yang dianggap relevan
      2. Method
         1. Preprocessing

Dilakukan dengan beberapa metode

* + - * 1. Menghapus baris kosong berdasarkan kolom FS\_ANAMNESA dan FS\_DIAGNOSA
        2. Menjadikan lowercase semua
        3. Menghapus tanda baca dan *tag HTML*
        4. Menghapus *stopwords*
        5. Menghapus spasi berlebih antar kata atau karakter
        6. Normalisasi dengan mengganti kata-kata *typo*
      1. Pembentukan matrix TF-IDF

Pembuatan matrix TF-IDF berdasarkan kolom FS\_ANAMNESA dan FS\_DIAGNOSA. Formula yang digunakan



* + - 1. Penghitungan kemiripan

Penghitungan kemiripan dilakukan dengan menggunakan cosine similarity. Formula yang digunakan



* + - 1. Perangkingan

Dilakukan berdasarkan cosine similarity score dengan mengambil 10 dokumen teratas

* + - 1. Evaluasi
         1. Precision

Formula yang digunakan adalah



* + - * 1. Recall

Formula yang digunakan adalah



* + 1. Results

Didapatkan 10 dokumen teratas

Tabel I  
*LIST* *QUERY* KOLOM FS\_ANAMNESA PADA PENGUJIAN MODEL

| **No** | **FS\_ANAMNESA** | **Jumlah Dokumen** |
| --- | --- | --- |
| 1. | kontrol | 10 |
| 2 | gigi sakit | 10 |
| 3. | ca paru | 10 |
| 4. | nyeri bahu kanan | 6 |
| 5. | nyeri | 4 |
| 6. | telinga nyeri | 4 |
| 7. | lutut kanan sakit | 4 |
| 8. | lutut kiri sakit | 4 |
| 9. | batuk pilek | 2 |
| 10. | batuk | 2 |

Tabel II  
*LIST* *QUERY* UNTUK KOLOM FS\_DIAGNOSA PADA PENGUJIAN MODEL

| **No** | **FS\_DIAGNOSA** | **Jumlah Dokumen** |
| --- | --- | --- |
| 1. | hamil | 10 |
| 2. | bronkitis akut | 10 |
| 3. | neurosa | 10 |
| 4. | dermatitis | 10 |
| 5. | et | 10 |
| 6. | bph | 10 |
| 7. | abses gigi | 10 |
| 8. | meningioma | 10 |
| 9. | kb | 4 |
| 10. | rhinitis | 4 |

Dan didapatkan average precision dan average recall pada Tabel III.

Tabel III  
NILAI HASIL PENGUJIAN DENGAN PRECISION DAN RECALL

|  |  |  |
| --- | --- | --- |
|  | *Average Precision* | *Average Recall* |
| FS\_ANAMNESA | 0.522 | 0.816 |
| FS\_DIAGNOSA | 0.574 | 0.781 |
| ***Average*** | **0.548** | **0.796** |

* + 1. Discussion

Didapatkan bahwa terdapat banyak dokumen relevan yang berhasil diretrieve, seperti yang terlihat pada Tabel I dan Tabel II. Hasil yang dihasilkan oleh system kelompok kami dipengaruhi oleh banyak factor, diantaranya banyaknya korpus yang digunakan, kebersihan data pada dataset yang digunakan, dan preprocessing pada query yang diinputkan. Untuk kelompok kami menggunakan 1000 data dari 126483 data pada dataset yang ada. Query yang diinputkan pun menggunakan tokenisasi saja, belum terdapat stemming, casefolding, ataupun proses yang lainnya. Untuk evaluasi dari system kami, terlihat pada Tabel III.

* + 1. References

[1] I. Rosadi and M. I. Purnama, “Analysis Of Time Analysis Of Outstanding Medical Records To Improve The Quality Of Services At Dustira Hospital, Cimahi Imam,” *Int. J. Heal. Sci.*, vol. 1, no. 1, pp. 1–5, 2021.

[2] Menteri Kesehatan Republik Indonesia, *Peraturan Menteri Kesehatan Republik Indonesia Nomor 24 Tahun 2022*, 2022.

[3] Menteri Kesehatan Republik Indonesia, *Peraturan Menteri Kesehatan Republik Indonesia Nomor 129/Menkes/SK/II/2008*, 2008.

[4] E. W. R. P, R. Indrawati, and L. Widjaja, “A Service Quality Review of Medical Record Department In Private Hospital, South Jakarta,” *J. Multidiscip. Acad.*, vol. 05, no. 02, pp. 101–105, 2021, [Online]. Available: <http://www.kemalapublisher.com/index.php/JoMA/article/view/554>.

[5] Indonesia. *Undang-Undang Nomor 29 Tahun 2004 tentang Praktik Kedokteran.* Lembaran Negara RI Tahun 2004 Nomor 116, Tambahan Lembaran RI Nomor 4431. Jakarta: Sekretariat Negara, 2004.

[6] Menteri Kesehatan Republik Indonesia, *Peraturan Menteri Kesehatan Republik Indonesia Nomor 269 Tahun 2008*, 2008.

[7] S. Buttcher, C. L. A. Clarke, and G. V. Cormack, *Information Retrieval Implementing and Evaluating Search Engines*. Cambridge, Massachusetts: The MIT Press, 2010.

[8] N. Hasanah, *Sistem Pencarian Skripsi Berbasis Information Retrieval di FASTIKOM UNSIQ*. Jurnal PPKM, 1(1), 105-113, 2017.

[9] F. Faridah, K. Munadi, and F. Arnia, “Aplikasi Histogram Discrete Cosine Transform (DCT) untuk Sistem Temu Kembali Citra Termal Berbasis Konten,” *Jurnal Nasional Komputasi dan Teknologi Informasi (JNKTI)*, vol. 2, no. 1, 2019.

[10] Y. Kang, J. Li, J. Yang, Q. Wang, and Z. Sun, “Semantic analysis for enhanced medical retrieval,” *2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 2017.

[11] 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.

[12] E. Wahyudi, S. Sfenrianto, M. J. Hakim, R. Subandi, O. R. Sulaeman, and R. Setiyawan, “Information retrieval system for searching JSON files with vector space model method,” *2019 International Conference of Artificial Intelligence and Information Technology (ICAIIT)*, 2019.

[13] P. E. Mas'udia, M. D. Atmadja, and L. D. Mustafa, “Information Retrieval Tugas Akhir dan Perhitungan Kemiripan Dokumen Mengacu pada Abstrak Menggunakan Vector Space Model,” *Simetris: Jurnal Teknik Mesin, Elektro dan Ilmu Komputer*, vol. 8, no. 1, pp. 355–362, 2017.

[14] A. T. Adiyanto and D. H. UN, Information Retrieval Sistem Kearsipan Pencarian Dokumen Di Dinas Pemberdayaan Perempuan Dan Perlindungan Anak Kota Semarang Menggunakan Metode Vector Space Model. JURNAL MAHAJANA INFORMASI, 7(1), pp.9-23, https://doi.org/10.51544/jurnalmi.v7i1.2538, 2022.

[15] R. C. N. Santi,S. Eniyati, R. Retnowati, and H. Yulianton, “PENGGUNAAN SISTEM TEMU KEMBALI DALAM PENCARIAN KATA UNTUK TERJEMAHAN AL QURAN”, 2019.

[16] B. Herwijayanti, D. E. Ratnawati, and L. Muflikhah, Klasifikasi Berita Online Dengan Menggunakan Pembobotan TF-IDF dan Cosine Similarity, Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer, vol. 2, no. 1, p. 306-312, ISSN 2548-964X, Agu. 2017.

[17] M. Widiasri, E. Tjandra and L. M. Chandra, “Peningkatan Kinerja Pencarian Dokumen Tugas Akhir menggunakan Porter Stemmer Bahasa Indonesia dan Fungsi Peringkat Okapi BM25,” TEKNIKA, vol. 6, no. 1, pp. 54-60, 2017.

[18] E. S. Munif, and Y. Kristian, “PENCARIAN TEMA SEJENIS SINOPSIS NOVEL BAHASA INDONESIA DENGAN MENGGUNAKAN GVSM.” Joutica: Journal of Informatic Unisla, 6(2), 492-501, 2021.

[19] 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.

[20] 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.

* 1. Penjelasan terkait projek STBI kami apakah termasuk multimedia information retrieval system

Tidak termasuk multimedia information retrieval system, dikarenakan data yang kami olah dan akan dihasilkan merupakan data text, tidak dalam berbentuk multimedia, seperti suara, gambar, video, dan animasi.