Uffi Nadzima
21/484427/PPA/06185
UAS Sistem Temu Balik Informasi

1. Pada proyek kelompok STBI yang dikembangkan kelompok saya dengan menggunakan data RekMed digunakan dua metode kueri pencarian dan similarity search yaitu **TF-IDF dan Cosine Similarity**. Kedua metode tersebut digunakan untuk mengukur kemiripan gejala yang dimasukkan pengguna dengan data gejala-penyakit yang ada pada sistem.

   **TF-IDF**

   $$idf_i = \log\left(\frac{N}{df_i}\right)$$
   $$w_{i,j} = tf_{i,j} \times idf_i$$

   Dimana:

   $i$ : term ke-$i$
   $j$ : dokumen ke-$j$
   $tf_{i,j}$ : term frequency
   $df_i$ : jumlah kemunculan term ke-$i$ pada semua dokumen
   $idf_i$ : inverse document frequency
   $N$ : jumlah semua dokumen
   $w_{i,j}$ : bobot term ke-$i$ pada dokumen ke-$j$

   **Cosine Similarity**

   $$sim(K_1, K_2) = \frac{K_1 \cdot K_2}{||K_1||||K_2||} = \frac{\sum_{i=1}^{n} K_{1i}K_{2i}}{\sqrt{\sum_{i=1}^{n}(K_{1i})^2}\sqrt{\sum_{i=1}^{n}(K_{2i})^2}}$$

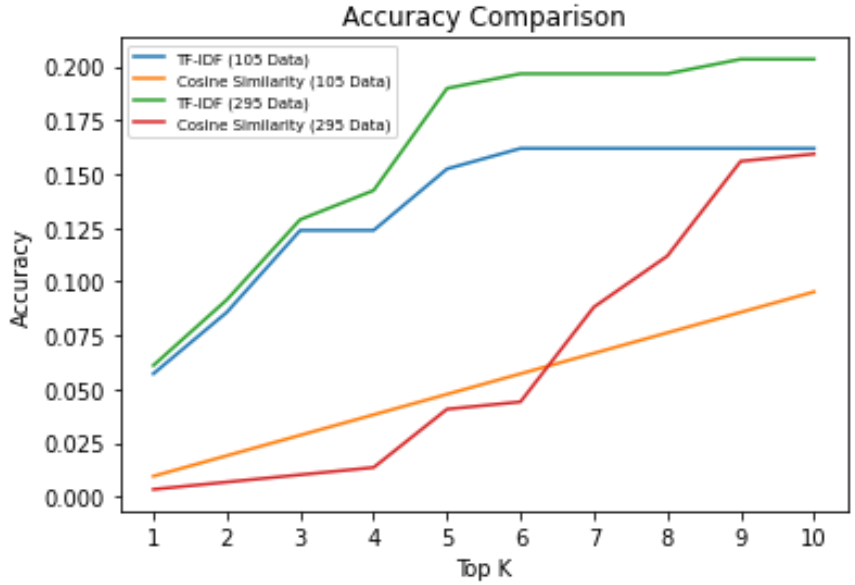
   Dimana:

   $K_1$ : kalimat ke-1
   $K_2$ : kalimat ke-2
   $K_{1i}$ : komponen vektor ke-$i$ pada kalimat ke-1
   $K_{2i}$ : komponen vektor ke-$i$ pada kalimat ke-2

2. Metode evaluasi yang digunakan yaitu **top-$k$-accuracy**, metode ini digunakan untuk mengetahui keakuratan hasil *k* prediksinya dimana ***k* merupakan banyak prediksi penyakit** yang dihasilkan. Jadi apabila diagnosis yang sebenernya termasuk dalam salah satu *k* prediksi maka prediksi dianggap benar/akurat.

   Hasil evaluasi dari sistem temu balik informasi yang kelompok kami kembangkan seperti pada gambar tabel berikut yang divisualisasikan ke dalam grafik yang ditunjukkan di bawahnya:

| $k$ | TF-IDF (105 Data) | Cosine Similarity (105 Data) | TF-IDF (295 Data) | Cosine Similarity (295 Data) |
|---|---|---|---|---|
| 1 | 0.057143 | 0.009524 | 0.061017 | 0.003390 |
| 2 | 0.085714 | 0.019048 | 0.091525 | 0.006780 |
| 3 | 0.123810 | 0.028571 | 0.128814 | 0.010169 |
| 4 | 0.123810 | 0.038095 | 0.142373 | 0.013559 |
| 5 | 0.152381 | 0.047619 | 0.189831 | 0.040678 |
| 6 | 0.161905 | 0.057143 | 0.196610 | 0.044068 |
| 7 | 0.161905 | 0.066667 | 0.196610 | 0.088136 |
| 8 | 0.161905 | 0.076190 | 0.196610 | 0.111864 |
| 9 | 0.161905 | 0.085714 | 0.203390 | 0.155932 |
| 10 | 0.161905 | 0.095238 | 0.203390 | 0.159322 |




Dari tabel dan grafik diketahui bahwa sistem yang kami kembangkan dengan dua metode dan dua macam data uji memiliki perbedaan dimana metode **TF-IDF yang diuji dengan data uji yang memiliki lebih banyak** data (hasil augmentasi data uji yang lebih sedikit) lebih baik akurasinya dibanding metode cosine similarity dan juga jika dibandingkan dengan pengujian dengan data yang lebih sedikit. Namun, dengan **akurasi hanya 20%** tersebut sistem yang kami kembangkan **belum dapat dinilai sebagai sistem yang baik** dalam memprediksi.

3. Sistem rekomendasi merupakan salah satu sistem penyaringan informasi yang mendukung pemberian saran terhadap item-item tertentu yang paling sesuai dengan yang pengguna ingin/butuhkan, sedangkan sistem tanya jawab merupakan sistem yang akan memberikan satu jawaban yang paling sesuai dengan apa yang ditanyakan oleh pengguna yang diambil dari basis data sistem tersebut dimana datanya merupakan pasangan pertanyaan dan jawabannya contoh yang paling umum digunakan yaitu sistem FAQ.
   Menurut saya, proyek STBI yang dikembangkan kelompok kami merupakan salah satu **sistem rekomendasi** berdasarkan **hasil keluarannya yang berupa prediksi penyakit-penyakit** yang sesuai dengan masukkan gejalanya.
4. Jawaban:
   a) Rangkuman lengkap proyek STBI Kelompok 1 (Stanford) berdasarkan Laporan Proyek:

**Introduction**

Latar belakang masalah dikembangkannya sistem pada proyek kami adalah perlunya deteksi dini pada suatu penyakit berdasarkan gejala-gejalanya. Sistem yang kami kembangkan ditujukan agar dapat bermanfaat untuk pengguna seperti tenaga medis khususnya pemula. Keluaran yang dihasilkan pada sistem ini dapat didetailkan dengan keluaran berupa informasi penyakitnya dan juga kode ICD 10. Sistem yang kami kembangkan dievaluasi dengan top-$k$-accuracy untuk mengetahui keakuratan hasil prediksi sistem.

**Methods**

Dataset yang digunakan pada sistem ada 4 yaitu data yang berisi daftar penyakit dan gejalanya yang telah divektorisasi ke dalam 0 dan 1, data kode ICD 10, data RekMed sebagai data uji pada evaluasi sistem dan data untuk informasi tambahan yang diambil dari Infobox Wikipedia.

Metode yang diimplementasikan pada sistem kami antara lain TF-IDF dan Cosine Similarity sebagai metode untuk mengukur kemiripan serta metode top-$k$-accuracy dengan library sklearn yaitu top_k_accuracy_score untuk mengevaluasi sistem.

Sistem temu balik informasi yang kami kembangkan terdiri dari beberapa komponen utama yaitu (i) Pemrosesan Kueri berupa preprocessing, (ii) Perluasan Kueri dengan menggunakan sinonim dari masukkan, (iii) Seleksi Gejala yaitu pengguna memilih dari daftar gejala hasil perluasan kueri yang diberikan, dan (iv) Prediksi Penyakit dari daftar akhir gejala. Selanjutnya sistem dievaluasi dengan metode top-$k$-accuracy.

**Results**

Sistem dapat berjalan dan memprediksi penyakit berdasarkan masukkan gejalanya namun hasil dari prediksi penyakit memiliki nilai keakuratan yang cukup rendah pada metode yang kami implementasikan. Akurasi paling tinggi pada pengujian yang kami lakukan pada $k = 10$ prediksi penyakit dengan TF-IDF menggunakan 105 data uji $\approx 16\%$, TF-IDF menggunakan 295 data uji $\approx 20\%$, cosine similarity menggunakan 105 data uji $\approx 9\%$ dan cosine similarity menggunakan 295 data uji $\approx 16\%$.

**Discussion**

Meskipun data uji lebih banyak hingga hamper 3 kali lipat keakuratan sistem tidak meningkat sebanding dengan peningkatan jumlah datanya. Dari pengujian tersebut disimpulkan bahwa sistem membutuhkan masukkan berupa gejala-gejala yang memiliki jumlah yang cukup dan juga relevan pada penyakit yang terkait sehingga prediksi penyakitnya dapat lebih akurat.

**References**

[1] M. Mustakim and R. Wardoyo, “Survey model-model pencarian informasi rekam medik elektronik,” JISKA (Jurnal Informatika Sunan Kalijaga), vol. 3, no. 3, p. 132–144, Aug. 2019. [Online]. Available:https://ejournal.uin-suka.ac.id/saintek/JISKA/article/view/33-01

[2] R. Silalahi and E. Sinaga, “Perencanaan implementasi rekam medis elektronik dalam pengelolaan unit rekam medis klinik pratama romana,” Jurnal Manajemen Informasi Kesehatan Indonesia, vol. 7, p. 22, 03 2019.
[3] Commonwealth of Australia, “MBS Telehealth Services from 1 July 2022,” 2022. [Online]. Available: http://www.mbsonline.gov.au/internet/mbsonline/publishing.nsf/Content/Factsheet-telehealth-1July22
[4] V. K and S. Jyothi, “Decision support system for congenital heart disease diagnosis based on signs and symptoms using neural networks,” International Journal of Computer Applications, vol. 19, 04 2011.
[5] C. D. Manning, P. Raghavan, and H. Schutze, ¨ Introduction to Information Retrieval. USA: Cambridge University Press, 2008.
[6] A. R. Lahitani, A. E. Permanasari, and N. A. Setiawan, “Cosine similarity to determine similarity measure: Study case in online essay assessment,” in 2016 4th International Conference on Cyber and IT Service Management, 2016, pp. 1–6.
[7] K. Park, J. Hong, and W. Kim, “A methodology combining cosine similarity with classifier for text classification,” Applied Artificial Intelligence, vol. 34, pp. 1–16, 02 2020.
[8] scikit-learn developers, “Metrics and scoring: quantifying the quality of predictions,” 2022. [Online]. Available: https://scikit-learn.org/stable/modules/model evaluation.html

b) Proyek STBI yang kami kembangkan **bukan merupakan multimedia information retrieval system**. Hal tersebut dikarenakan proyek yang kami kembangkan hanya menggunakan **satu bentuk media yaitu teks** sebagai masukkan maupun keluaran sistemnya.