PENERAPAN METODE COSINE SIMILARITY PADA PERLUASAN PENCARIAN INFORMASI BUAH DAN SAYURAN LOKAL JEMBER

Joko Wahyu Faozhan[1], Wiwik Suharso[2]
Jurusan Teknik Informatika Fakultas Teknik Universitas Muhammadiyah Jember
jokowahyufauzan@gmail.com, wiwiksuharso@unmuhjember.ac.id

ABSTRAK

Aplikasi ensiklopedia buah dan sayuran lokal Jember adalah aplikasi yang menampilkan informasi mengenai nama, jenis, deskripsi, klasifikasi, kandungan dan manfaat buah dan sayuran lokal jember. Pada fitur pencarian informasi di aplikasi ensiklopedia buah dan sayuran lokal Jember, semua *user* dapat mencari informasi buah dan sayuran lokal dengan kata kunci secara bebas. Analisis sistem ensiklopedia pada perluasan pencarian informasi ini menggunakan metode *cosine similarity*. Penelitian dimulai dari *stemming* untuk mencari kata dasar dan TF-IDF untuk mencari nilai bobot di setiap *term* yang akan menghasilkan nilai *cosine* di setiap dokumen. Analisis hasil pengujian rata-rata kinerja *precision* terbaik dari skenario 1, 2, 3 dan 4 sebesar 87% dengan ambang batas 0,09 sehingga digunakan oleh pengguna sebagai default pencarian dengan *precision* tertinggi. Sedangkan hasil pengujian rata-rata kinerja *recall* terbaik dari skenario 1, 2, 3 dan 4 sebesar 100% dengan ambang batas 0,03 sehingga digunakan oleh pengguna sebagai default pencarian dengan *recall* tertinggi. Kemudian hasil pengujian rata-rata kinerja *accuracy* terbaik dari skenario 1, 2, 3 dan 4 sebesar 97% dengan ambang batas 0,07 sehingga digunakan oleh pengguna sebagai default pencarian dengan *accuracy* tertinggi. Akan tetapi nilai rata-rata terbaik dari kinerja *precision*, *recall* dan *accuracy* sebesar 81%, 88% dan 97% dengan ambang batas 0,07 sebagai default pencarian pada aplikasi ensiklopedia buah dan sayuran lokal Jember.
Kata kunci : Buah dan Sayuran lokal Jember, *Cosine similarity*, Ensiklopedia.

ABSTRACT

Jember local fruit and vegetable encyclopedia application is an application that displays information about the name, type, description, classification, content and benefits of local fruits and vegetables Jember. In the information search feature in the Jember local fruit and vegetable encyclopedia application, all users can search for local fruit and vegetable information by keyword freely. Analysis of the encyclopedia system on the expansion of this information search using the cosine similarity method. The research starts from stemming to find the basic word and TF-IDF to find the weight value in each term that will produce a cosine value in each document. Analysis of the results of testing the average best precision performance of scenarios 1, 2, 3 and 4 by 87% with a threshold of 0.09 so that it is used by users as the default search with the highest precision. While the average test results of the best recall performance from scenarios 1, 2, 3 and 4 are 100% with a threshold of 0.03 so that it is used by users as the default search with the highest recall. Then the results of testing the average accuracy of the best performance of scenarios 1, 2, 3 and 4 by 97% with a threshold of 0.07 so that it is used by users as the default search with the highest accuracy. However, the best average value of precision, recall and accuracy performance is 81%, 88% and 97% with a threshold of 0.07 as the default search in the Jember local fruit and vegetable encyclopedia application.
Keywords : *Local Fruits and Vegetables Jember, Encyclopedia, Cosine Similarity.*

1 PENDAHULUAN

Ensiklopedia buah dan sayuran lokal Jember berisi 48 jenis buah lokal dan 51 jenis sayuran lokal yang tersebar dalam 31 Kecamatan di Kabupaten Jember (Komarayanti, 2018). Data tersebut telah digunakan sebagai database pada aplikasi web ensiklopedia buah dan sayuran lokal Jember sehingga dapat diakses pada alamat url http://e-ensiklopedia. 000webhostapp.com/ (Salim, dkk, 2018). Aplikasi web ensiklopedia buah dan sayuran lokal Jember dapat digunakan oleh instansi pendidikan, kesehatan, pertanian, dan masyarakat umum baik sebagai sumber pembelajaran sains, preferensi masyarakat dalam konsumsi buah dan sayuran lokal untuk kesehatan dan pengobatan, serta sumber informasi dalam pengambilan keputusan.

Pengguna membutuhkan informasi yang luas dan spesifik seperti informasi buah dan sayuran lokal jember meliputi nama, sejarah, klasifikasi, kandungan dan manfaat buah dan sayuran lokal jember.

Pada penelitian ini akan dilakukan pengembangan dari penelitian (Salim, dkk, 2018) tentang web ensiklopedia yang menggunakan data buah dan sayuran lokal jember yang meliputi nama dan keterangan buah dan sayuran. Pengembangan yang akan dilakukan pada penelitian ini adalah pengembangan pada perluasan sistem pencarian informasi pada aplikasi web ensiklopedia buah dan sayuran lokal Jember, agar dapat dihasilkan informasi mengenai nama, jenis, deskripsi, klasifikasi, kandungan dan manfaat buah dan sayuran lokal jember menggunakan metode *Cosine Similarity*.

Metode *Cosine Similarity* tersebut dilakukan dengan mempertimbangkan frekuensi suatu kata dalam suatu dokumen (*term frequency*), dan penyebaran suatu kata pada sekumpulan dokumen (*Inverse Document Frequency*) serta kemiripan antar dokumen dengan metode *Cosine Similarity* dinyatakan dalam nilai bobot dari TF*IDF. Perhitungan kemiripan antara *Query* dengan *Document* dalam aplikasi Ensiklopedia dengan metode *Cosine Similarity* akan menghasilkan informasi buah dan sayuran lokal Jember secara akurat (Salim, dkk, 2018). Untuk memperoleh hasil pencarian yang maksimal diperlukan sebuah nilai ambang batas (*threshold*) agar sistem dapat memilih mana okumen yang mirip dan mana yang stidak.

1.1 .Rumusan Masalah

1. Berapa nilai ambang batas (*threshold*) untuk mengetahui tingkat kemiripan dokumen dengan *query*?
2. Berapa tingkat *precision*, *recall* dan *accuracy* dalam mengukur kinerja sistem?

1.2 Tujuan Penelitian

1. Menemukan nilai ambang batas (*threshold*) untuk mengetahui tingkat kemiripan dokumen dengan *query*.
2. Menemukan tingkat *precision*, *recall* dan *accuracy* untuk mengetahui kinerja sistem.

1.3 Manfaat Penelitian

1. Kalangan umum khususnya pelajar dapat menggunakan produk ensiklopedia untuk membantu dalam mendapatkan informasi tentang topik buah dan sayuran lokal jember dalam beragam jenis seperti teks, gambar dan video secara mudah dan cepat.
2. Sistem ensiklopedia dapat memberikan informasi secara akurat berdasarkan kata kunci dengan metode C*osine Similarity* .

1.4 Batasan Masalah

2. Jumlah data buah dan sayuran lokal Jember sebanyak 99 jenis.
3. Dataset penelitian menggunakan data buah dan sayuran lokal jember dari buku "Buah dan Sayur Lokal di Kabupaten Jember".
4. Parameter dataset yang digunakan adalah nama, jenis, deskripsi, klasifikasi, kandungan dan manfaat buah dan sayuran lokal jember dalam database ensiklopedia.
5. Untuk mengukur kinerja sistem menggunakan *precision*, *recall* dan *accuracy*.

2. TINJAUAN PUSTAKA

2.1 Data Buah dan Sayuran

Potensi buah lokal di Kabupaten Jember yang di temukan sebanyak 48 jenis buah. Sayur lokal di Kabupaten Jember yang ditemukan sebanyak 51 jenis sayuran. Buah dan sayuran lokal yang berada di Kabupaten Jember tersebar 31 kecamatan, yaitu Kecamatan Kencong, Gumukmas, Puger, Wuluhan, Ambulu, Silo, Mayang, Mumbulsari, Jenggawah, Ajung, Rambipuji, Balung, Umbulsari, Semboro, Jombang, Sumberbaru, Tanggul, Bangsalsari, Panti, Sukorambi, Arjasa, Pakusari, Kalisat, Ledokombo, Sumberjambe, Sukowono, Jelbuk, Kaliwates, Sumbersari dan Patrang. (Komarayanti. 2018)

2.2 Web Ensiklopedia

Ensiklopedia sudah di kenal oleh kalangan pelajar sebagai media untuk mendapatkan informasi tentang topik tertentu yang di inginkan. Kebanyakan produk Ensiklopedia di pasaran

dalam bentuk buku, majalah, atlas dan kartu. Produk ensiklopedia fisik tersebut telah digunakan dalam proses pembelajaran siswa-siswi di sekolah. Akan tetapi produk ensiklopedia fisik memiliki keterbatasan dalam kemudahan akses dan kecepatan penyebaran informasi serta bersifat statis. Sementara Ensiklopedia online seperti wikipedia memiliki kelemahan dalam manajemen penyuntingan, akurasi informasi, dan informasi yang bersifat umum atau tidak spesifik pada konten data lokal. (Salim, dkk, 2018).

2.3 *Cosine Similarity*

Metode *Cosine similarity* merupakan metode yang digunakan untuk menghitung tingkat kesamaan (*similarity*) antar dua buah objek. *Query* dan *Document* di proses dengan memanfaatkan *Text Mining* yang sebagai pra-proses.

Text mining didefinisikan sebagai proses penemuan kembali relasi dan fakta yang terkubur didalam teks dan tidak harus baru. Dalam penelitian ini *text mining* meliputi *tokenizing, filtering, stemming*. (Salim, dkk, 2018)

2.3.1 *Preprocessing*

1. *Tokenizing*

Pemisahan rangkaian term (*tokenization*). *Tokenization* adalah tugas memisahkan deretan kata didalam kalimat, paragraf atau halaman menjadi token atau potongan kata tunggal atau *termmedword.* Tahapan ini juga menghilangkan karakter-karakter tertentu seperti tanda baca dan mengubah semua token ke bentuk huruf kecil (*lower case*).

2. *Filtering*

Tahap filtering adalah tahap pengambilan kata-kata penting dari hasil *tokenizing* menggunakan algoritma *stopword removal, Stopword removal* adalah proses penyaringan (*filtering*) terhadap kata-kata yang tidak layak untuk dijadikan sebagai pembeda atau kata kunci sehingga kata-kata tersebut dapat dihilangkan dari dokumen, seperti, kata sambung, kata depan, kata ganti, kata sifat, dan lain sebagainya.

3. *Stemming*

Kata-kata yang mucul didalam dokumen sering mempunyai banyak varian morfologik. Oleh arena itu setelah menjalani proses *tokenizing* dan *stopword removal,* kata-kata yang tersisa menjalani proses stemming. Stemming bertujuan untuk mengubah atau mengembalikan kata mejadi bentuk kata dasarnya (*root word*) dengan menghilangkan imbuhan. Pada penelitian ini menggunakan algoritma Nazief & Adriani yang disesuaikan atau dikembangkan dalam Bahasa Indonesia.

2.3.2 *Term Frequency* (tf) – *Inverse Document Frequency* (Idf)

Term Frequency (tf) merupakan frekuensi kemunculan suatu kata (*term*) dalam dokumen. Oleh sebab itu, *tf* memiliki nilai yang bervariasi dari satu dokumen ke dokumen lain tergantung dari tingkat kepentingan sebuah *term* dalam sebuah dokumen. Semakin sering suatu *term* muncul dalam suatu dokumen, *term* tersebut akan memiliki nilai *tf* yang lebih besar dari pada *term-term* lain yang jarang muncul. (Sulistyo, dkk, 2015)

Dalam penelitian ini, algoritma pembobotan *Term Frequency (tf) – Inverse Document Frequency (idf)* ditetapkan pada tahap *similarity document*. Nilai tf-idf diperoleh dengan menggunakan persamaan (1) :

$$W_{i,j} = tf_{ij} \times idf_j = tf_{i,j} \times \log(\frac{N}{df_j}) \text{ .....(1)}$$

Keterangan :

$W_{i,j}$ **:** bobot *term* ke-j terhadap dokumen ke-i

tf_{ij} **:** jumlah kemunculan term j ke dalam dokumen i

N **:** jumlah dokumen secara keseluruhan

df_j **:** Jumlah dokumen yang mengandung *term* j

Berdasarkan persamaan (1), berapapun besarnya nilai tf_{ij} apabila N = df_j maka akan didapat hasil nol (0) untuk perhitungan *idf*. Untuk itu, dapat ditambahkan nilai 1 pada sisi *idf*, sehingga perhitungan bobotnya menjadi :

$$W_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_j}\right) + 1 \text{ .....(2)}$$

2.3.3 *Rumus cosine similarity*

$$Cosine \rightarrow \text{sim}(d_j, q) = \frac{\overrightarrow{d_j}.\vec{q}}{|\overrightarrow{d_j}|.|\vec{q}|} = \frac{\sum_{i=1}^{t}(w_{ij} \cdot w_{iq})}{\sqrt{\sum_{i=i}^{t} w_{ij}^2 \cdot \sum_{i=1}^{t} w_{iq}^2}}$$

Keterangan :

d : dokumen

q : *query*

w : bobot *term* dalam sebuah dokumen

2.4 *Threshold*

Untuk memperoleh hasil pencarian dokumen yang maksimal diperlukan sebuah nilai ambang batas *(Threshold)* agar sistem dapat memilih mana dokumen yang mirip dan mana yang tidak mirip. Dokumen dengan nilai ≥ *threshold* dapat dinyatakan mirip, sedangkan dokumen dengan nilai < *threshold* dinyatakan tidak mirip. Untuk mendapatkan nilai batas diperlukan suatu data training untuk melakukan uji coba (Muharromah. 2018).

2.5 *Recall, Precission* dan *Accuracy*

Sistem temu kembali informasi mengembalikan sekumpulan dokumen sebagai jawaban dari *query* pengguna. Terdapat dua kategori dokumen yang dihasilkan oleh sistem temu kembali informasi terkait pemrosesan *query*, yaitu relevant documents (dokumen yang relevan dengan *query)* dan *retrieved* documents (dokumen yang diterima pengguna). Ukuran umum yang digunakan untuk mengukur kualitas dari data *retrieval* adalah kombinasi precision dan *recall*.

Precision mengevaluasi kemampuan sistem temu kembali informasi untuk menemukan kembali data *top-ranked* yang paling relevan, dan didefinisikan sebagai persentase data yang dikembalikan yang benar-benar relevan terhadap *query* pengguna. *Precision* merupakan proporsi dari suatu set yang diperoleh yang relevan.(Herdiawan, 2016) (Melita, dkk, 2018).

Tabel 2.1 Rumus *confusion matrix*

		Nilai Sebenarnya (pakar)	
		Relevant	***Not Relevant***
Nilai Prediksi (sistem)	***Retrieved***	*True Positive* (TP)	*False Positive* (FP)
	Not Retrieved	*False Negative* (FN)	*True Negative* (TN)

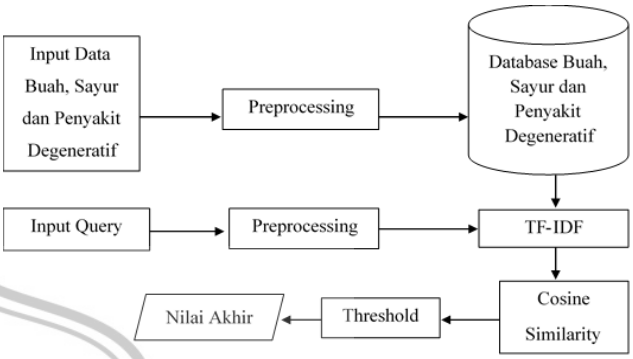
$$\textbf{\textit{Recall}} = \frac{TP}{TP+FN} \quad (1)$$

$$\textbf{\textit{Precision}} = \frac{TP}{TP+FP} \quad (2)$$

$$\textbf{\textit{Accuracy}} = \frac{TP+TN}{TP+FP+TN+FN} \quad (3)$$

3 METODELOGI PENELITIAN

3.1 Rancangan Sistem



Gambar 3.1 Rancangan Sistem

3.2 Skenario Pengujian

Skenario pengujian dilakukan dengan membuat 20 *query* pengguna yang berbeda baik jumlah kata dan jenis data yang diinginkan yang terbagi atas 4 skenario. Ambang batas *(threshold)* nilai *similarity* menggunakan rata-rata keseluruhan nilai *cosine* dari *link-link* yang dihasilkan oleh sistem. Hasil pencarian dari masing-masing *query* akan dilakukan perbandingan kinerja dengan *precision, recall* dan akurasi. Diberikan contoh data perhitungan dalam menghasilkan *link-link* informasi yang dibutuhkan berdasarkan *query* pengguna sebagai berikut :

3.2.1 *Preprocessing Query* dan Dokumen.

Tabel 3.1 *Preprocessing Query*

	Query	***Tokenizing***	***Filtering***	***Stemming***
Q	Buah untuk mencegah kanker	buah untuk mencegah kanker	buah mencegah kanker	buah cegah kanker

***a. Preprocessing* DATA SET D1**

Nama : Alpukat

Jenis : (Persea americana)

Deskripsi :Alpukat berasal dari Amerika dan Meksiko.Alpukat diperkenalkan oleh bangsa

Belanda pada abad ke 19. Spanyol kemudian memasuki wilayah Amerika pada abad 16, kemudian membawa alpukat kepada masyarakat Eropa. Sejak itu, alpukat mulai dikenal luas, bahkan masuk ke Indonesia. Buah alpukat merupakan tanaman tahunan. Buah alpukat termasuk dalam golongan buah sejati tunggal berdaging. Hasil penelitian buah alpukat yang berada di Kabupaten Jember dapat ditemukan di Kecamatan Gumukmas, Puger, Wuluhan, Ambulu, Tempurejo, Silo, Mayang, Mumbulsari, Jenggawah, Rambipuji, Balung, Semboro, Sumberbaru, Tanggul, Panti, Arjasa, Pakusari, Kalisat, Ledokombo, Sumberjambe, Sukowono, jelbuk, Kaliwates, Sumbersari, dan Patrang.

Kandungan : kilojoule, kalori, lemak, protein, karbohidrat, sodium, kalium

Manfaat : 1. Mengurangi risiko kanker 2. Senyawa lain yang ada di dalam alpukat adalah karotenoid. Ini adalah senyawa antioksidan yang memiliki kemampuan sebagai anti kanker . 3. Baik untuk penderita osteoarthtritis 4. Berguna untuk mencegah osteoporosis 5. Menurunkan kolesterol jahat 6. Baik untuk menjaga berat badan dan kesehatan jantung 7. Kebaikan alpukat untuk kesehatan jantung juga mungkin signifikan. 8. Baik bagi kesehatan janin 9. Bagus untuk mata 10. Mengurangi stres

Klasifikasi : Kingdom : Plantae
Divisi : Magnoliophyta
Kelas : Magnoliopsida
Ordo : Laurales
Famili : genus
Genus : Persea
Species : Persea americana

Hasil *Tokenizing, Filtering, Stemming* D1

Nama : alpukat

Jenis : persea americana

Deskripsi : alpukat asal amerika meksiko alpukat kenal bangsa belanda abad spanyol masuk wilayah amerika abad bawa alpukat masyarakat eropa sejak alpukat mulai kenal luas masuk indonesia buah alpukat rupa tanam tahun buah alpukat masuk golong buah sejati tunggal daging hasil teliti buah alpukat kabupaten jember temu gumukmas puger wuluhan ambulu tempurejo silo mayang mumbulsari jenggawah rambipuji balung semboro sumberbaru tanggul panti arjasa pakusari kalisat ledokombo sumberjambe sukowono jelbuk kaliwates sumbersari patrang

Kandungan : kilojoule kalori lemak protein karbohidrat sodium kalium

Manfaat : kurang risiko kanker senyawa alpukat karotenoid senyawa antioksidan milik mampu bagai anti kanker baik derita osteoarthtritis cegah osteoporosis turun kolesterol jahat baik jaga berat badan sehat jantung baik alpukat sehat jantung signifikan baik sehat janin bagus mata kurang stres

Klasifikasi : kingdom plantae
divisi magnoliophyta
kelas magnoliopsida
ordo laurales
famili genus
genus persea
species persea americana

***b. Preprocessing* DATA SET D2**

Nama : Anggur Hitam

Jenis : Vitis Vinifera

Deskripsi : Anggur berasal dari Armenia, tetapi budidaya anggur sudah dikembangkan di Timur Tengah sejak 4000 SM. Buah Anggur merupakan bunga majemuk. Buah anggur yang tumbuh dan dibudidayakan di Kabupaten Jember adalah jenis Vitis vinifera pada varietas anggur ini mampu tumbuh di daerah beriklim kering dan di daerah dataran rendah hingga 300 m di atas permukaan laut. Hasil penelitian buah anggur hitam yang berada di Kabupaten Jember dapat di temukan di Kecamatan Gumukmas, Ambulu, Silo, Rambipuji, Balung, Jombang, Tanggul, Bangsalsari, Arjasa, Jelbuk, Kaliwates, Sumbersari, dan Patrang.

Kandungan : kilojoule, kalori, lemak, protein, karbohidrat, kalium

Manfaat : 1.Mengantur gula darah 2.Meningkatkan fungsi otak 3.Melindungi jantung 4.Meningkatkan penglihatan 5.Mencegah kanker 6.Membuat rambut sehat 7.Meningkatkan

daya tahan tubuh 8.Menjaga kesehatan tulang 9.Menjaga berat badan 10.Menyehatkan kulit

Klasifikasi : Kingdom : Plantae
Divisi : Spermatophyta
Kelas : Magnoliopsida
Ordo : Rhamnales
Famili : genus
Genus : Vitis
Species : Vitis vinifera

Hasil *Tokenizing, Filtering, Stemming* D2

Nama : anggur hitam

Jenis : vitis vinifera

Deskripsi : anggur asal armenia budidaya anggur kembang timur tengah sejak 4000 buah anggur rupa bunga majemuk buah anggur tumbuh budidaya kabupaten jember jenis vitis vinifera varietas anggur tumbuh daerah iklim kering daerah datar rendah hingga 300 muka laut hasil teliti buah anggur hitam kabupaten jember temu gumukmas ambulu silo rambipuji balung jombang tanggul bangsalsari arjasa jelbuk kaliwates sumbersari patrang

Kandungan : kilojoule kalori lemak protein karbohidrat kalium

Manfaat : atur gula darah tingkat fungsi otak lindung jantung tingkat lihat cegah kanker buat rambut sehat tingkat daya tahan tubuh jaga sehat tulang jaga berat badan sehat kulit

Klasifikasi : kingdom plantae
divisi spermatophyta
kelas magnoliopsida
ordo rhamnales
famil genus
genus vitis
species vitis vinifera

***c. Preprocessing* DATA SET D3**

Nama : Belimbing

Jenis : Averrhoa carambola L.

Deskripsi : Belimbing merupakan tanaman buah berupa pohon yang berasal dari kawasan Malaysia, kemudian menyebar luas ke berbagai negara yang beriklim tropis lainnya di dunia termasuk Indonesia. Hasil penelitian buah belimbing yang berada di Kabupaten Jember dapat ditemukan di Kecamatan Kencong, Gumukmas, Wuluhan, Ambulu, Silo, Mayang, Mumbulsari, Jenggawah, Ajung, Rambipuji, Umbulsari, Semboro, Jombang, Sumberbaru, Tanggul, Panti, Arjasa, Pakusari, Kalisat, Sumberjambe, Sukowono, Jelbuk, Kaliwates, Sumbersari, dan Patrang

Kandungan : kilojoule, kalori, lemak, protein, karbohidrat, sodium, kalium

Manfaat : 1. Menyehatkan pencernaan 2.Pencegahan kanker 3. Kaya antioksidan 4.Meningkatkan kerja enzim 5. Mengatasi tekanan darah tinggi 6. Mengatasi gangguan ginjal 7. Mengurangi kadar kolesterol jahat 8. Belimbing memberikan nutrisi lain yang baik untuk rambut 9. Makan buah belimbing atau menggunakan belimbing langsung pada kulit berjerawat

Klasifikasi : Kingdom : Plantae
Divisi : Magnoliopsyta
Kelas : Magnoliopsida
Ordo : Geraniales
Famili : genus
Genus : Averrhoa
Species : Averrhoa carambola L

Hasil *Tokenizing, Filtering, Stemming* D2

Nama : belimbing

Jenis : averrhoa carambola

Deskripsi : belimbing rupa tanam buah rupa pohon asal kawasan malaysia sebar luas bagai negara iklim tropis dunia masuk indonesia hasil teliti buah belimbing kabupaten jember temu kencong gumukmas wuluhan ambulu silo mayang mumbulsari jenggawah ajung rambipuji umbulsari semboro jombang sumberbaru tanggul panti arjasa pakusari kalisat sumberjambe sukowono jelbuk kaliwates sumbersari patrang

Kandungan : kilojoule kalori lemak protein karbohidrat sodium kalium

Manfaat : sehat cerna cegah kanker kaya antioksidan tingkat kerja enzim atasi tekan darah tinggi atas ganggu ginjal kurang kadar kolesterol jahat belimbing beri nutrisi baik rambut makan

buah belimbing guna belimbing langsung kulit jerawat

Klasifikasi **:** kingdom plantae
divisi magnoliopsyta
kelas magnoliopsida
ordo geraniales
famili genus
genus averrhoa
species averrhoa carambola

3.2.2 Perhitungan TF-IDF

Tabel 3.1 Pembobotan TF*IDF

TERM	Q	TF				IDF	TF-IDF			
		D1	D2	D3	DF		Q	D1	D2	D3
buah	1	4	3	3	4	1	1	4	3	3
cegah	1	1	1	1	4	1	1	1	1	1
Kanker	1	2	1	1	4	1	1	2	1	1
alpukat	0	10	0	0	1	1,6	0	16	0	0
persea	0	3	0	0	1	1,6	0	4,8	0	0
americana	0	2	0	0	1	1,6	0	3,2	0	0
asal	0	1	1	1	3	1,1	0	1,1	1,1	1,1
amerika	0	2	0	0	1	1,6	0	3,2	0	0
meksiko	0	1	0	0	1	1,6	0	1,6	0	0
kenal	0	2	0	0	1	1,6	0	3,2	0	0
bangsa	0	1	1	0	2	1,3	0	1,3	1,3	0
belanda	0	1	0	0	1	1,6	0	1,6	0	0
abad	0	2	0	0	1	1,6	0	3,2	0	0
spanyol	0	1	0	0	1	1,6	0	1,6	0	0
masuk	0	3	0	1	2	1,3	0	3,9	0	1,3
wilayah	0	1	0	0	1	1,6	0	1,6	0	0
bawa	0	1	0	0	1	1,6	0	1,6	0	0
masyarakat	0	1	0	0	1	1,6	0	1,6	0	0
eropa	0	1	0	0	1	1,6	0	1,6	0	0
sejak	0	1	1	0	2	1,3	0	1,3	1,3	0
mulai	0	1	0	0	1	1,6	0	1,6	0	0
luas	0	1	0	1	2	1,3	0	1,3	0	1,3
indonesia	0	1	0	1	2	1,3	0	1,3	0	1,3
rupa	0	1	1	2	3	1,1	0	1,1	1,1	2,2
tanam	0	1	0	1	2	1,3	0	1,3	0	1,3
tahun	0	1	0	0	1	1,6	0	1,6	0	0
golong	0	1	0	0	1	1,6	0	1,6	0	0
sejati	0	1	0	0	1	1,6	0	1,6	0	0
tunggal	0	1	0	0	1	1,6	0	1,6	0	0
daging	0	1	0	0	1	1,6	0	1,6	0	0
hasil	0	1	1	1	3	1,1	0	1,1	1,1	1,1
teliti	0	1	1	1	3	1,1	0	1,1	1,1	1,1
kabupaten	0	1	2	1	3	1,1	0	1,1	2,2	1,1
jember	0	1	2	1	3	1,1	0	1,1	2,2	1,1
temu	0	1	1	1	3	1,1	0	1,1	1,1	1,1
gumukmas	0	1	1	1	3	1,1	0	1,1	1,1	1,1
puger	0	1	0	0	1	1,6	0	1,6	0	0
wuluhan	0	1	0	1	2	1,3	0	1,3	0	1,3
ambulu	0	1	1	1	3	1,1	0	1,1	1,1	1,1
tempurejo	0	1	0	0	1	1,6	0	1,6	0	0
silo	0	1	1	1	3	1,1	0	1,1	1,1	1,1
mayang	0	1	0	1	2	1,3	0	1,3	0	1,3
mumbulsari	0	1	0	1	2	1,3	0	1,3	0	1,3
jenggawah	0	1	0	1	2	1,3	0	1,3	0	1,3
rambipuji	0	1	1	1	3	1,1	0	1,1	1,1	1,1
balung	0	1	1	0	2	1,3	0	1,3	1,3	0
semboro	0	1	0	1	2	1,3	0	1,3	0	1,3

sumber baru	0	1	0	1	2	1,3	0	1,3	0	1,3
tanggul	0	1	1	1	3	1,1	0	1,1	1,1	1,1
panti	0	1	1	1	3	1,1	0	1,1	1,1	1,1
arjasa	0	1	1	1	3	1,1	0	1,1	1,1	1,1
pakusari	0	1	1	1	3	1,1	0	1,1	1,1	1,1
kalisat	0	1	1	0	2	1,3	0	1,3	1,3	0
ledokombo	0	1	1	0	2	1,3	0	1,3	1,3	0
sumberjambe	0	1	0	0	1	1,6	0	1,6	0	0
sukowono	0	1	1	1	3	1,1	0	1,1	1,1	1,1
jelbuk	0	1	1	1	3	1,1	0	1,1	1,1	1,1
kaliwates	0	1	1	1	3	1,1	0	1,1	1,1	1,1
sumbersari	0	1	1	1	3	1,1	0	1,1	1,1	1,1
patrang	0	1	1	1	3	1,1	0	1,1	1,1	1,1
kilojoule	0	1	1	1	3	1,1	0	1,1	1,1	1,1
kalori	0	1	1	1	3	1,1	0	1,1	1,1	1,1
lemak	0	1	1	1	3	1,1	0	1,1	1,1	1,1
protein	0	1	1	1	3	1,1	0	1,1	1,1	1,1
karbohidrat	0	1	1	1	3	1,1	0	1,1	1,1	1,1
sodium	0	1	0	1	2	1,3	0	1,3	0	1,3
kalium	0	1	1	1	3	1,1	0	1,1	1,1	1,1
kurang	0	2	0	1	2	1,3	0	2,6	0	1,3
risiko	0	1	0	0	1	1,6	0	1,6	0	0
senyawa	0	2	0	0	1	1,6	0	3,2	0	0
karotenoid	0	1	0	0	1	1,6	0	1,6	0	0
antioksidan	0	1	0	1	2	1,3	0	1,3	0	1,3
milik	0	1	0	0	1	1,6	0	1,6	0	0

mampu	0	1	0	0	1	1,6	0	1,6	0	0
bagai	0	1	0	1	2	1,3	0	1,3	0	1,3
anti	0	1	0	0	1	1,6	0	1,6	0	0
baik	0	4	0	1	2	1,3	0	5,2	0	1,3
derita	0	1	0	0	1	1,6	0	1,6	0	0
osteoarthritis	0	1	0	0	1	1,6	0	1,6	0	0
osteoporosis	0	1	0	0	1	1,6	0	1,6	0	0
turun	0	1	0	0	1	1,6	0	1,6	0	0
kolesterol	0	1	0	1	2	1,3	0	1,3	0	1,3
jahat	0	1	0	1	2	1,3	0	1,3	0	1,3
jaga	0	1	2	0	2	1,3	0	1,3	2,6	0
berat	0	1	1	0	2	1,3	0	1,3	1,3	0
badan	0	1	1	0	2	1,3	0	1,3	1,3	0
sehat	0	3	3	1	3	1,1	0	3,4	3,4	1,1
jantung	0	2	1	0	2	1,3	0	2,6	1,3	0
signifikan	0	1	0	0	1	1,6	0	1,6	0	0
janin	0	1	0	0	1	1,6	0	1,6	0	0
bagus	0	1	0	0	1	1,6	0	1,6	0	0
mata	0	1	0	0	1	1,6	0	1,6	0	0
stres	0	1	0	0	1	1,6	0	1,6	0	0
kingdom	0	1	1	1	3	1,1	0	1,1	1,1	1,1
plantae	0	1	1	1	3	1,1	0	1,1	1,1	1,1
divisi	0	1	1	1	3	1,1	0	1,1	1,1	1,1
magnoliophyta	0	1	0	0	1	1,6	0	1,6	0	0
kelas	0	1	1	1	3	1,1	0	1,1	1,1	1,1
magnoliopsida	0	1	1	1	3	1,1	0	1,1	1,1	1,1
ordo	0	1	1	1	3	1,1	0	1,1	1,1	1,1
laurales	0	1	0	0	1	1,6	0	1,6	0	0
famili	0	1	1	1	3	1,1	0	1,1	1,1	1,1
genus	0	2	2	2	3	1,1	0	2,2	2,2	2,2

species	0	1	1	1	3	1,1	0	1,1	1,1	1,1
anggur	0	0	7	0	1	1,6	0	0	11	0
hitam	0	0	2	0	1	1,6	0	0	3,2	0
vitis	0	0	4	0	1	1,6	0	0	6,4	0
vinifera	0	0	3	0	1	1,6	0	0	4,8	0
armenia	0	0	1	0	1	1,6	0	0	1,6	0
budidaya	0	0	2	0	1	1,6	0	0	3,2	0
kembang	0	0	1	0	1	1,6	0	0	1,6	0
timur	0	0	1	0	1	1,6	0	0	1,6	0
tengah	0	0	1	0	1	1,6	0	0	1,6	0
4000	0	0	1	0	1	1,6	0	0	1,6	0
bunga	0	0	1	0	1	1,6	0	0	1,6	0
majemuk	0	0	1	0	1	1,6	0	0	1,6	0
tumbuh	0	0	2	0	1	1,6	0	0	3,2	0
jenis	0	0	1	0	1	1,6	0	0	1,6	0
varietas	0	0	1	0	1	1,6	0	0	1,6	0
daerah	0	0	2	0	1	1,6	0	0	3,2	0
iklim	0	0	1	1	2	1,3	0	0	1,3	1,3
kering	0	0	1	0	1	1,6	0	0	1,6	0
datar	0	0	1	0	1	1,6	0	0	1,6	0
rendah	0	0	1	0	1	1,6	0	0	1,6	0
300	0	0	1	0	1	1,6	0	0	1,6	0
muka	0	0	1	0	1	1,6	0	0	1,6	0
laut	0	0	1	0	1	1,6	0	0	1,6	0
jombang	0	0	1	1	2	1,3	0	0	1,3	1,3
bangsalsari	0	0	1	0	1	1,6	0	0	1,6	0
atur	0	0	1	0	1	1,6	0	0	1,6	0
gula	0	0	1	0	1	1,6	0	0	1,6	0
darah	0	0	1	1	2	1,3	0	0	1,3	1,3
tingkat	0	0	3	1	2	1,3	0	0	3,9	1,3
fungsi	0	0	1	0	1	1,6	0	0	1,6	0

otak	0	0	1	0	1	1,6	0	0	1,6	0
lindung	0	0	1	0	1	1,6	0	0	1,6	0
rambut	0	0	1	0	1	1,6	0	0	1,6	0
daya	0	0	1	0	1	1,6	0	0	1,6	0
tahan	0	0	1	0	1	1,6	0	0	1,6	0
tubuh	0	0	1	0	1	1,6	0	0	1,6	0
tulang	0	0	1	0	1	1,6	0	0	1,6	0
kulit	0	0	1	1	2	1,3	0	0	1,3	1,3
spermatophyta	0	0	1	0	1	1,6	0	0	1,6	0
rhamnales	0	0	1	0	1	1,6	0	0	1,6	0
famil	0	1	0	1	2	1,3	0	1,3	0	1,3
belimbing	0	0	0	6	1	1,6	0	0	0	9,6
averrhoa	0	0	0	3	1	1,6	0	0	0	4,8
carambola	0	0	0	2	1	1,6	0	0	0	3,2
pohon	0	0	0	1	1	1,6	0	0	0	1,6
kawasan	0	0	0	1	1	1,6	0	0	0	1,6
malaysia	0	0	0	1	1	1,6	0	0	0	1,6
negara	0	0	0	1	1	1,6	0	0	0	1,6
tropis	0	0	0	1	1	1,6	0	0	0	1,6
dunia	0	0	0	1	1	1,6	0	0	0	1,6
kencong	0	0	0	1	1	1,6	0	0	0	1,6
ajung	0	0	0	1	1	1,6	0	0	0	1,6
umbulsari	0	0	0	1	1	1,6	0	0	0	1,6
cerna	0	0	0	1	1	1,6	0	0	0	1,6
enzim	0	0	0	1	1	1,6	0	0	0	1,6
tekan	0	0	0	1	1	1,6	0	0	0	1,6
tinggi	0	0	0	1	1	1,6	0	0	0	1,6
ganggu	0	0	0	1	1	1,6	0	0	0	1,6

ginjal	0	0	0	1	1	1,6	0	0	0	1,6
kadar	0	0	0	1	1	1,6	0	0	0	1,6
nutrisi	0	0	0	1	1	1,6	0	0	0	1,6
makan	0	0	0	1	1	1,6	0	0	0	1,6
langsung	0	0	0	1	1	1,6	0	0	0	1,6
jerawat	0	0	0	1	1	1,6	0	0	0	1,6
magnoliopsyta	0	0	0	1	1	1,6	0	0	0	1,6
geranial	0	0	0	1	1	1,6	0	0	0	1,6

Keterangan :

Q : Jumlah kemunculan *term* dalam query

D1 : Jumlah kemunculan *term* dalam dokumen 1

D2 : Jumlah kemunculan *term* dalam dokumen 2

D3 : Jumlah kemunculan *term* dalam dokumen 3

DF : Jumlah keseluruhan dokumen yang memuat *term*

IDF :bobot dari jumlah keseluruhan dokumen dibagi jumlah dokumen yang memuat *term.*

3.2.3 Perhitungan *Cosine Similarity*

Berikut langkah *retrieval* menggunakan data pada tabel perhitungan TF-IDF:

a. Mengkalikan bobot antara bobot *query* dengan bobot *term* pada setiap dokumen

Q.D1 = (1*4)+(1*1)+(1*2)= 7

Q.D2 = (1*3)+(1*1)+(1*1)= 5

Q.D3 = (1*3)+(1*1)+(1*1)= 5

b. Menghitung panjang *query* dengan mengakarkan jumlah kuadrat bobot *query.*

$|Q| = \sqrt{1^2 + 1^2 + 1^2}$ = 1,732050808

c. Menghitung panjang dokumen dengan mengakarkan jumlah dari bobot dokumen yang dikuadratkan pada setiap dokumen yang dikalikan dengan Q.

$|D1| = \sqrt{(1^2 + 1^2 + 1^2) * (4^2 + 1^2 + \dots)}$ = 4,405094

$|D2| = \sqrt{(1^2 + 1^2 + 1^2) * (3^2 + 1^2 + \dots)}$ = 4,388376

$|D3| = \sqrt{(1^2 + 1^2 + 1^2) * (3^2 + 1^2 + \dots)}$ = 4,019487

Menghitung *cosine similarity* dengan membagi bobot $Q.D_n$ dengan hasil perkalian antara panjang *query* ($|Q|$) dan panjang dokumen ($|D_n|$).

Cos(Q, D1) = $\frac{7}{42,0174933}$ = 0,166597278

Cos(Q, D2) = $\frac{5}{33,18139481}$ = 0,150686854

Cos(Q, D3) = $\frac{5}{29,2252084}$ = 0,171085179

Ambang batas (threshold) adalah 0,16278977 yang di dapat dari jumlah total cosine di masing-masing dokumen di bagi banyaknya dokumen yaitu 3 Dokumen (D1, D2, D3). Sehingga tingkat similaritas dari skenario di atas yaitu D3, D1 dan D2.

3.3 *Recall, Precission* dan *Accuracy*

Keakuratan sistem dapat di evaluasi dengan cara perhitungan *Recall, Precision* dan *Accuracy* sebagai berikut :

Tabel 3.2 Perhitungan recall dan precision dengan threshold 0,09

	Relevant	***Not relevant***	**Total**
Retrieve	3	0	3
not retrieve	0	96	96
Total	3	96	99

Perhitungan *recall, precision* dan *accuracy* adalah sebagai berikut :

Recall $= \frac{TP}{TP+FN}$

$= \frac{3}{3+0} = \frac{3}{3} \times 100$

= 100%

Precision $= \frac{TP}{TP+FP}$

$= \frac{3}{3+0} = \frac{3}{3} \times 100$

= 100%

$$\textbf{\textit{Accuracy}} = \frac{TP+TN}{TP+FP+TN+FN}$$

$$= \frac{3+96}{3+0+96+0} = \frac{99}{99} \times 100$$

$$= \mathbf{100\%}$$

Dari contoh diatas dapat disimpulkan bahwa dengan *threshold* 0,1 sistem menghasilkan *recall* sebesar 100%, *precision* sebesar 100% dan *accuracy* 100%.

DAFTAR PUSTAKA

Erawati, Indri, dan Yuni Yamasari. 2012. Aplikasi Ensiklopedia Negara Digital untuk Memotivasi Pengguna dalam Mengenal Negara di Dunia. *Jurnal Manajemen Informatika.*

Hasugian, Jonner. 2006. Penelusuran Informasi Ilmiah Secara Online: Perlakuan Terhadap Seorang Pencari Informasi Sebagai Real User. *Jurnal Studi Perpustakaan dan Informasi,* 2(1), 1-13.

Herdiawan. 2016. Analisis Sentimen Terhadap Telkom Indihome Berdasarkan Opini Publik Menggunakan Metoe Improved K-Nearest Neighbor. *Jurnal Ilmiah Komputer dan Informatika (KOMPUTA).*

Komarayanti, Sawitri. 2017. *Ensiklopedia Buah-buahan Lokal Jember Berbasis Potensi Alam Jember*. p-ISSN 2527-7111; e-ISSN 2528-1615. http://jurnal.unmuhjember.ac.id/index.php /BIOMA/article/download/591/470.

Komarayanti, Sawitri. 2018. Buah dan Sayur Lokal di Kabupaten Jember. ISBN : 978-602-6988-63-8.

Melita, Ria dkk. 2018. Penerapan metode *term frequency inverse document frequency* (tf-idf) dan *cosine similarity* pada sistem temu kembali informasi untuk mengetahui *syarah hadits* berbasis web (studi kasus: *syarah umdatil ahkam*). *Jurnal Teknik Informatika.*

Muharromah, Lailiyatul, Ulya Anisatur, Mudafiq Riyan Pratama. 2018. *Penilai Esai Otomatis Ujian Tengah Semester di SMK Asrama Pembina Masyarakat Jatimulyo Jember Menggunakan Metode Cosine Similarity.* Skripsi tidak diterbitkan. Jember: Program Studi Teknik Informatika.

Salim, Agus, Wiwik Suharso, Hardian Oktavianto. 2018. *Pencarian Link Informasi Pada Aplikasi Ensiklopedia Buah dan Sayuran Lokal Dengan Metode Cosine Similarity.* Skripsi tidak diterbitkan. Jember: Program Studi Teknik Informatika.

Sulistyo, Meiyanto Eko, Ristu Saptono, Adam Asshidiq. 2015. Penilaian Ujian Bertipe Essay Menggunakan Metoe *Text Similarity*. Jurnal Telematika.

Wahyudi, Dwi, Teguh Susyanto, Didik Nugroho. 2017. Implementasi dan Analisis Algoritma Stemming Nazief & Adriani dan Porter pada Dokumen Berbahasa Indonesia. *Jurnal Ilmiah Sinus* 15 (2), 49-56.