PENERAPAN SISTEM PENCARIAN ABSTRAK JURNAL BERDASARKAN KATA KUNCI MENGGUNAKAN ALGORITMA NAIVE BAYES (STUDI KASUS : ABSTRAK JURNAL STMIK BANDUNG)

Dedy Apriadi, M.Si [1], Haris Irnawan, S.T, M.M [2] , Syiffa Nurjayanti [3]

STMIK BANDUNG
Sekolah Tinggi Manajemen dan Informatika Bandung
JL. Cikutra No.113, Bandung 40007, INDONESIA

dedy.apriadi.ssi.msi@gmail.com

Abstrak

Pencarian dokumen merupakan upaya kegiatan pencarian untuk menghasilkan informasi yang dicari, Seiring dengan berkembangnya teknologi yang efektif digunakan serta dalam penyajian informasi secara detail kebutuhan konsumen terhadap informasi dalam bentuk jurnal atau artikel ilmiah semakin meningkat. Naive Bayes merupakan Algoritma penggalian data dalam pencarian pola *text*, algoritma ini dapat membantu pencaria dokumen dengan penerapan metode *k-gram*, ekstraksi dokumen, dan nilai *similarity* serta *extreme programming* sebagai metode pengembangan perangkat lunak, untuk membantu dalam pencarian (abstrak jurnal) berdasarkan kata kunci.

Kata kunci : *Naive Bayes,* Metode k-gram, pencarian, dokumen

Abstract

Document search is a search activity effort to produce the information sought. Along with the development of effective technology used and in presenting detailed information, consumer needs for information in the form of journals or scientific articles are increasing. Naive Bayes is a data mining algorithm in searching for text patterns, this algorithm can help document search by applying the k-gram method, document extraction, and similarity values as well as extreme programming as a software development method, to assist in searching (journal abstract) based on keywords.

Keywords : Naive Bayes, K-gram Method, Search, Doument

1. Pendahuluan

Di era globalisasi ini kebutuhan konsumen terhadap informasi dalam bentuk jurnal atau artikel ilmiah semalin meningkat, sehingga pengelompokan jurnal dibutuhkan untuk mempermudah pencarian informasi. Dalam pencarian jurnal dapat dilakukan berdasarkan topik yang menggambarkan isi pemahaman jurnal tanpa harus membaca secara keseluruhan. Maka dari itu penelitian membangun sebuah sistem pencarian abstrak berdasarkan kata kunci menggunakan algoritma naive bayes.

1.1 *Naive Bayes*

Naive Bayes adalah algoritma yang diterapkan pada sistem yang digabungkan dengan proses rumus k-gram. Langkah-langkah untuk menghitung presentase kecepatan pencarian serta kemiripan abstrak menggunakan Naive bayes dan k-gram :

- Perhitungan presentase tertinggi

P(H | X)=P(X | H)/P(X) P(H)

Keterangan :

X : Data dengan kelas yang belum diketahui
H : Hipotesis data merupakan suatu kelas yang spesifik
P(H|X) : Probabilitas hipotesis H berdasarkan kondisi X (posteriori probabilitas)
P(H) : Probabilitas hipotesis H (prior probabilitas)
P(X|H) : Probabilitas X berdasarkan kondisi pada hipotesis H
P(X) : Probabilitas X

- Perhitungan nilai similarity

$$S = \frac{\text{KC}}{(A + B)}$$

Di mana:
S : Nilai *similarity*
K : 2 (bigrams)
C : Jumlah k-gram yang sama dari teks 1 dan teks 2
A : Jumlah k-gram dari teks 1
B : Jumlah k-gram dari teks 2

2. Pembahasan

Naive Bayes merupakan metode klasifikasi yang didasarkan probabilitas dan statistik. Dalam buku Konsep Data Mining dan Penerapan dijelaskan bahwa Naïve Bayes bekerja sangat baik dibandingkan dengan model classifier yang lain karena memiliki

tingkat akurasi yang lebih baik. Pada teorema Bayes, bila terdapat dua kejadian yang terpisah, maka teorema Bayes seperti pada persamaan dan nilai Bayes yang diambil adalah persentase tertinggi dari semua kemungkinan.

2.1 Metodologi

Pada tahap ini dibahas tentang proses pencarian abstrak jurnal yaitu :

1. Melakukan Ekstraksi Dokumen (*Text Mining*).
2. Perhitungan nilai similarity.

2.2 Pembahasan

Berdasarkan latar belakang dan tujuan dalam penelitian ini penyusun menggunakan beberapa proses(ekstraksi dokumen) sebelum melakukan pencarian abstrak.

Dalam penelitian ini penyusun melakukan beberapa tahap proses text mining yaitu :

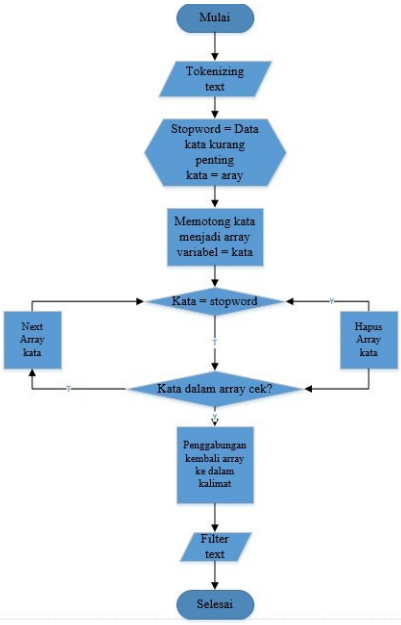
1. *Case Folding*, mengkonversi keseluruhan teks dalam dokumen menjadi suatu bentuk standar

Gambar 1 Flowchart Case Folding

2. *Tokenizing*, pemotongan *string* masukkan karakter selain huruf dihilangkan dan dianggap *delimiter*, delimiter merupakan untuk membatasi atau memisahkan data yang disajikan dalam kalimat. Salah satu contoh dari *delimiter* adalah tanda koma, titik koma atau titik dua.

Gambar 2 Flowchart Tokenizing

3. *Filtering*, mengambil kata-kata penting atau pemacu dari abstrak jurnal yaitu merupakan kata kunci.

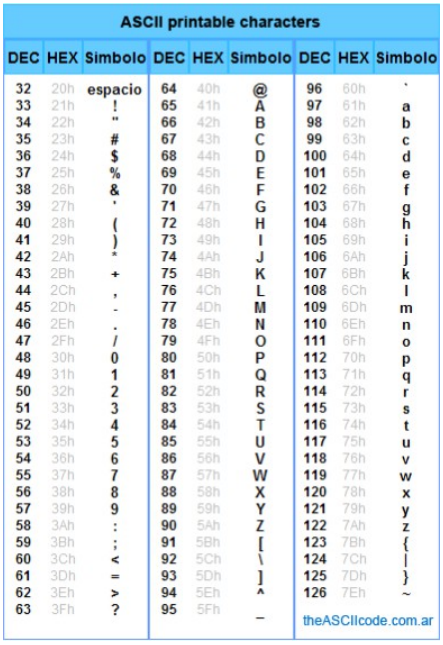


Gambar 3 Flowchart Filtering

2.3 *Hashing*

Hashing adalah suatu cara untuk mentransformasi sebuah *string* menjadi suatu nilai yang unik dengan panjang tertentu *(fixed-length)* yang berfungsi sebagai penanda *string* tersebut. Fungsi untuk menghasilkan nilai ini disebut fungsi *hash*, sedangkan nilai yang dihasilkan disebut nilai *hash*.Berikut contoh perubahan

jenis data keturunan *alphabet* ke dalam bilangan bulat (a = 97, b= 98, dst).Pada laporan yang Penyusun buat urutan *alphabet* menggunakan karakter *code American Standard Code for Information Interchange (ASCII)*.



ASCII printable characters

DEC	HEX	Simbolo	DEC	HEX	Simbolo	DEC	HEX	Simbolo
32	20h	espacio	64	40h	@	96	60h	`
33	21h	!	65	41h	A	97	61h	a
34	22h	"	66	42h	B	98	62h	b
35	23h	#	67	43h	C	99	63h	c
36	24h	$	68	44h	D	100	64h	d
37	25h	%	69	45h	E	101	65h	e
38	26h	&	70	46h	F	102	66h	f
39	27h	'	71	47h	G	103	67h	g
40	28h	(	72	48h	H	104	68h	h
41	29h	)	73	49h	I	105	69h	i
42	2Ah	*	74	4Ah	J	106	6Ah	j
43	2Bh	+	75	4Bh	K	107	6Bh	k
44	2Ch	,	76	4Ch	L	108	6Ch	l
45	2Dh	-	77	4Dh	M	109	6Dh	m
46	2Eh	.	78	4Eh	N	110	6Eh	n
47	2Fh	/	79	4Fh	O	111	6Fh	o
48	30h	0	80	50h	P	112	70h	p
49	31h	1	81	51h	Q	113	71h	q
50	32h	2	82	52h	R	114	72h	r
51	33h	3	83	53h	S	115	73h	s
52	34h	4	84	54h	T	116	74h	t
53	35h	5	85	55h	U	117	75h	u
54	36h	6	86	56h	V	118	76h	v
55	37h	7	87	57h	W	119	77h	w
56	38h	8	88	58h	X	120	78h	x
57	39h	9	89	59h	Y	121	79h	y
58	3Ah	:	90	5Ah	Z	122	7Ah	z
59	3Bh	;	91	5Bh	[	123	7Bh	{
60	3Ch	<	92	5Ch	\	124	7Ch	|
61	3Dh	=	93	5Dh	]	125	7Dh	}
62	3Eh	>	94	5Eh	^	126	7Eh	~
63	3Fh	?	95	5Fh	_			

theASCIIcode.com.ar

Gambar 4 Tabel ASCII

Berikut merupakan metode yang ada di dalam *hashing*,

1. **K-Gram**
   K-Gram adalah rangkaian *terms* dengan panjang K. Kebanyakan yang digunakan sebagai *terms* adalah kata. K-Gram merupakan sebuah metode yang diaplikasikan untuk pembangkitan kata atau karakter. Metode K-Gram ini digunakan untuk mengambil potongan-potongan karakter huruf sejumlah k dari sebuah kata yang secara kontinuitas dibaca dari teks sumber hingga akhir dari dokumen.
2. **Modulus**
   Sebagai pembagi nilai *hash text* agar memiliki keunikan tersendiri

2.4 Nilai *Similarity*

Pada pembobotan dokumen ini perhitungan nilai *similarity*;

1. ***Dice's Similarity Coeficients***

   Untuk menghitung nilai *similarity* dari dokumen *fingerprint* yang didapat maka digunakan *Dice's Similarity Coeficients* dengan cara menghitung nilai dari jumlah K-Gram yang digunakan pada kedua dokumen yang diuji, sedangkan dokumen *fingerprint* didapat dari jumlah nilai K-Gram yang sama. Nilai *Similarity* tersebut dapat dihitung dengan menggunakan rumus matematis berikut.

$$S = \frac{\mathrm{KC}}{(A+B)}$$

Di mana:
S : Nilai *similarity*
K : 2 (bigrams)
C : Jumlah k-gram yang sama dari teks 1 dan teks 2
A : Jumlah k-gram dari teks 1
B : Jumlah k-gram dari teks 2

2.4 Hasil Penelitian

Berikut hasil penelitian dari proses yang telah dibahas:

1. Ekstraksi Dokumen.
2. Algoritma Naive Bayes dan Metode k-gram

2.4.1 Ekstraksi Dokumen

Setelah melakukan percobaan *trial and error* dari dokumen jurnal uji dan jurnal latih menggunakan model bayesian dengan k-gram = 3, basis = 7 dan *modulus* = 10007 sebagai berikut

Tabel 1 Data Masukan

No	Teks Uji	Teks Latih
1	Penerapan metode Naive Bayes Classifier	Metode Naive bayes digunakan
2	Pengelolaan = lingkungan	Tingkat kepuasan masyarakat
3	Kinerja pemerintah	Kebudayaan terhadap kinerja pemerintah
4	Hutan gundul akibat penebangan liar	Penanaman = menjadikan hutan tetap hijau
5	Setiap orang berusaha	Jangan menyerah !

Tabel 2 Ekstraksi Case Folding

No	Teks Uji	Teks Latih
1	penerapan metode naive bayes classifier	metode naive bayes digunakan
2	pengelolaan = lingkungan	tingkat kepuasan masyarakat
3	kinerja pemerintah	kebudayaan terhadap kinerja pemerintah
4	hutan gundul akibat penebangan liar	penanaman = menjadikan hutan tetap hijau
5	setiap orang berusaha	jangan menyerah !

Tokenizing

Tabel 3 Ekstraksi Tokenizing

Teks Uji	Teks latih
No 1	
Penerapan metode naive bayes classifier	metodenaivebayesdigunakan
No 2	
Pengelolaanlingkungan	tingkatkepuasanmasyarakat
No 3	
kinerjapemerintah	kebudayaanterhadapkinerjapemerintah
No 4	
hutangundulakibatpenebanganliar	penanamanmenjadikanhutantetaphijau
No 5	
setiaporangberusaha	janganmenyerah

Filtering

Tabel 4 Ekstraksi Filtering

No	Teks Uji	Teks Latih
1	naivebayes	metode
2	kelolalingkung	tingkatkepuasanmasyarakat
3	kinerjapemerintah	budayakinerjapemerintah
4	hutangundulakibatpenebanganliar	penanamanmenjadikanhutantetaphijau
5	orangberusaha	menyerah

Hashing

Tabel 5 Hashing

No	Text Uji			Text Latih		
1	5931	5719	8833	6666	4434	9067
	9655	4238	1532	3427		
	5719	84441	6960			
	5191					
2	6639	4788	2463	8523	5583	4668
	8098	7668	4897	7216	4897	9873
	503	4679	4472	6873	1761	8866
	9541	2218	3470	7837	432	405
				8082	2147	9154
				6563	2011	4162
				5659		
3	6416	884	356	133	1761	8817
	3356	5610	7298	7249	7009	3160
	7548	6688	8920	1080	8626	628
	5942	6021		1777	7385	
4	5316	5234	6960	6960	5191	2512
	5485	4962	2621	5289	2610	2474

	33 23	39 33	53 27	21 53	10 15	33 07
	24 90	76 19	45 54	53 16	52 34	69 60
	75 21	26 9	40 64	61 22	86 37	86 26
	81 80	30 29	77 44	68 46	70 58	62 20
	67 2	37 70	54 85	95 19	71 0	30 13
	45 21	92 96	98 73	33 07		
	75 9					
5	14 9	21 58	54 85	61 27	61 60	18 64
	49 62	28 66	48 91			
	42 65	31 33				

2.5 Nilai Similarity

Similarity Dengan K-Gram = 3

Tabel 5 Hasil Similarity

No	Hasil Similarity
1	49.72 %
2	6.45 %
3	52.61 %
4	17.02 %
5	0 %

2.5 Pengujian Pada Dokumen

Penelitian pada pencarian abstrak jurnal berdasarkan kata kunci sebagai bahan uji dengan penyesuaian rumus berikut :

Tabel 6 Dokumen Pengujian

Kata kunci	Waktu pencarian	Total pencarian
No 1 js	1.3830	2
No 2 Sistem Informasi	0.2028	4
No 3 Naive Bayes	2.2325	4
No 4 Kemiripan, dokumen	1.2736	3
No 5 Sistem keputusan	0.3947	7

3. Kesimpulan dan Saran

3.1 Kesimpulan

Dari berbagai penjelasan yang telah diuraikan dalam laporan ini, maka dapat disimpulkan beberapa hal sebagai berikut :

1. Telah menghasilkan sebuah sistem yang mampu melakukan pencarian abstrak jurnal berdasarkan kata kunci secara efektif.
2. Sistem dapat menampilakn hasil pencarian berupa dokumen abstrak jurnal

3.2 Saran

Saran yang diberikan sehingga sistem yang telah Agar sistem yang dibangun dapat digunakan lebih optimal dan dapat berjalan sesuai dengan yang diharapkan, maka ada beberapa saran yang dapat dijadikan bahan pertimbangan, yaitu :

1. Perlunya menerapkan Algoritma pada tahap *stemming* dibanding dengan menerima informasi dari *library*
2. Meningkatkan *visualisasi interface* menjadi lebih interaktif dan menarik.
3. Diharapkan Aplikasi ini tidak hanya dapat melakukan pencarian abstrak jurnal namun dapat mendeteksi dokumen karya-karya yang harus di lindungi guna menjaga integritas suatu karya.

Daftar Pustaka

[8] A. Indranandita, B. Susanto and A. Rachmat, "SISTEM KLASIFIKASI DAN PENCARIAN JURNAL DENGAN MENGGUNAKAN METODE NAIVE BAYES DAN VECTOR SPACE MODEL," *Jurnal Informatika,* vol. 4, no. 2, pp. 14-16, 2008.

[2] I. Gita Anugrah, "Penerapan Metode N-Gram dan Cosine Similarity Dalam Pencarian Pada Repositori Artikel Jurnal Publikasi/ Indra Gita Anugrah.," *Jurnal Building Of Informatics,* vol. 3, no. 3, pp. 5-7, 2021.

[5] N. L. Wiwik Sri Rahayu and N. W. Wardani, "IMPLEMENTASI METODA NAIVE BAYES DAN VECTOR SPACE MODEL DALAM DETEKSI KESAMAAN ARTIKEL JURNAL BERBAHASA INDONESIA," *Jurnal Infomedia,* vol. 4, no. 2, p. 19, 2019.

[6] N. M. Suhendri, Y. Afrilia and R. , "Klasifikasi Karya Ilmiah (Tugas Akhir) Mahasiswa Menggunakan Metode Naive Bayes Classifier (Nbc)," *Jurnal Sistem Informasi,* vol. 10, no. 2, pp. 273-277, 2021.

[4] N. Nilamsari, "Memahami studi dokumen dalam penelitian kualitatif," *Jurnal Wacana,* vol. XIII, no. 2, p. 178, 2014.

[9] W. Rizky, T. Astuti and A. SM, Implementasi Extreme Programming Pada Sistem Reservasi Tiket Travel Berbasis Android Dan Website, 2018.

[10] M. I. Rahayu and F. T. Zhafran, "ANALISIS SENTIMEN LAPORAN PERKEMBANGAN ANAK DIDIK TAMAN KANAK KANAK DENGAN MENGGUNAKAN METODA NAIVE BAYES", *JURTIK STMIK Bandung*, vol. 4, no. 1, pp. 30–36, Jun. 2015.