

BAB III

ANALISIS IDENTIFIKASI BAHASA MENGGUNAKAN METODE N-GRAM

3.1 Identifikasi Bahasa Menggunakan Metode *N-gram*

Ide awalnya adalah menggunakan *n-gram* yang frekuensi kemunculannya dapat menunjukkan apakah teks tersebut menggunakan bahasa tertentu [CAV94]. Salah satu keunggulan menggunakan *n-gram* dan bukan suatu kata utuh secara keseluruhan adalah bahwa *n-gram* tidak akan terlalu sensitif terhadap kesalahan penulisan yang terdapat pada suatu dokumen.

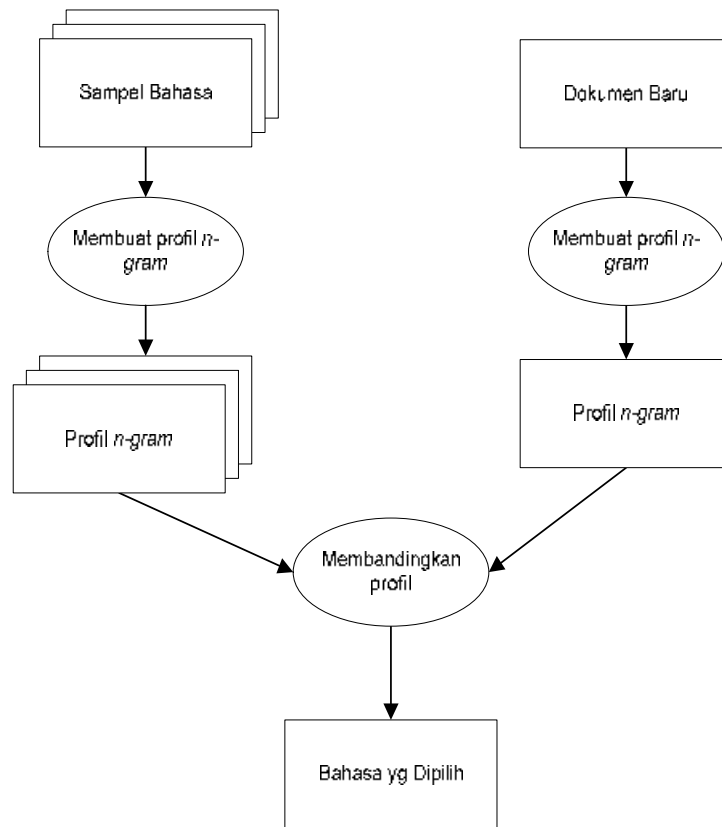
Suatu bahasa pasti akan menggunakan suatu kata lebih banyak dibanding kata lainnya. Hal ini sesuai dengan *Zipf's Law* yang dapat dinyatakan sebagai berikut:

The n-th most common word in a human language text occurs with a frequency inversely proportional to n [ZIP49]

Implikasi dari hukum ini adalah bahwa selalu ada kata yang lebih sering digunakan dibanding kata lainnya dalam suatu bahasa. Selain itu terdapat sifat kontinuitas dalam perbedaan frekuensi penggunaan kata tersebut. Hal ini juga berlaku untuk potongan dari kata tersebut atau dalam kata lain *n-gram* kata tersebut.

Dengan menggunakan *Zipf's Law* sebagai ide dasar maka dapat dibuat suatu metode identifikasi bahasa dengan membandingkan *rank* sebuah *n-gram* dalam dokumen dengan *rank n-gram* yang sama pada dokumen yang telah diketahui bahasanya [CAV94].

Pertama, dari dokumen-dokumen yang telah di ketahui bahasanya (disebut juga sampel bahasa), akan dibangkitkan profil *n-gram*. Profil *n-gram* ini merupakan daftar frekuensi setiap *n-gram* yang muncul dalam dokumen tersebut. Profil *n-gram* ini ini akan menjadi dasar dalam identifikasi bahasa suatu dokumen. Ilustrasi dari ide dasar proses ini dapat dilihat pada Gambar III-1.



Gambar III-1 Gambaran Sistem

Selanjutnya, berdasarkan *Zipf's Law*, untuk tiap bahasa akan terdapat *n-gram frequency rank* yang berbeda. Karena itu, dengan membandingkan *rank* dari dua dokumen yang berbeda, bisa didapatkan kedekatan antara suatu bahasa dengan bahasa yang digunakan oleh dokumen. Kedekatan atau jarak antar bahasa ini dapat digunakan untuk menentukan bahasa yang digunakan oleh suatu dokumen.

3.2 Pembentukan *N-gram*

Pada tugas ini akan digunakan *n-gram* yang memiliki panjang berbeda secara bersamaan. Untuk mencapai hal tersebut, dilakukan *padding* dengan *blank* di awal dan di akhir suatu kata. Sebagai contoh dari kata "TEXT" akan terdapat ("_" merepresentasikan *blank*) :

- Bi-gram* : _T, TE, EX, XT, T_
- Tri-gram* : _TE, TEX, EXT, XT_, T__
- Quad-gram* : _TEX, TEXT, EXT_, XT___, T___

Pada contoh diatas dilakukan penambahan satu buah *blank* di awal kata. Sedangkan untuk penambahan di akhir kata disesuaikan dengan *n-gram* yang dibuat, pada *bi-gram* ditambahkan satu *blank*, pada *tri-gram* ditambahkan dua *blank* dan demikian pula untuk selanjutnya. Dengan demikian untuk suatu kata dengan panjang *k* akan terdapat *k+1 bi-gram*, *k+1 tri-gram*, dan seterusnya.

3.3 Pembentukan Profil Dokumen

Profil dokumen dibentuk berdasarkan frekuensi *n-gram* yang muncul di dalam dokumen. Dokumen akan dibaca kata per kata, dan untuk setiap kata akan dibuat *n-gram* dari kata tersebut. Untuk setiap *n-gram* yang dibangkitkan, akan dicatat dalam sebuah *hash table* dengan *n-gram* sebagai kunci dan jumlah sebagai isi. Apa bila *n-gram* tersebut sudah pernah muncul di dalam dokumen maka frekuensi untuk *n-gram* itu akan ditambah satu, jika belum maka *n-gram* tersebut akan ditambahkan ke dalam *hash table* dengan jumlah kemunculan satu.

Sebagai contoh untuk pembuatan profil yang menggunakan *bi-gram* pada sebuah dokumen yang hanya berisi satu kalimat, “Pengidentifikasi bahasa alami menggunakan metode ngram”. Akan menghasilkan *bi-gram* sebagai berikut:

Tabel III-1 Keluaran Pembangkitan *N-gram* dokumen A

No	Kata	<i>Bi-gram</i>
1	Pengidentifikasi	_P; Pe; en; ng; gi; id; de; en; nt; ti; if; fi; ik; ka; as; si; i_
2	bahasa	_b; ba; ah; ha; as; sa; a_
3	alami	_a; al; la; am; mi; i_
4	menggunakan	_m; me; en; ng; gg; gu; un; na; ak; ka; an; n_
5	metode	_m; me; et; to; od; de e_
6	ngram	_n; ng; gr; ra; am; m_

Sebagai perbandingan berikut ini adalah proses yang sama dilakukan pada dokumen kedua yg berisi satu kalimat: “Natural language identification by using ngram method”. Akan menghasilkan *bi-gram*:

Tabel III-2 Hasil Pembangkitan *n-gram* dokumen B

No	Kata	Bi-gram
1	Natural	_N; na; at; tu; ur; ra; al; l_
2	language	_l; la; an; ng; gg; gu; ua; ag; ge; e_
3	identification	_i; id; de; en; nt; ti; if; fi; ic; ca; at; ti; io; on; n_
4	by	_b; by; y_
5	using	_u; us; si; in; ng; g_
6	ngram	_n; ng; gr; ra; am; m_
7	method	_m; me; et; th; ho; od; d_

Setelah proses tersebut dilakukan pada seluruh kata dalam dokumen, maka *n-gram* akan diurutkan berdasarkan frekuensi kemunculannya dalam dokumen. Daftar *n-gram* yang terurut berdasarkan kemunculannya inilah yang disebut sebagai profil dokumen. Selanjutnya profil dokumen akan digunakan untuk menentukan bahasa yang digunakan oleh dokumen tersebut.

Bi-gram tersebut akan di urutkan berdasarkan frekuensi kemunculannya sehingga menjadi daftar seperti Tabel III-3:

Tabel III-3 Keluaran Pembangkitan Profil dokumen A

No	Bi-gram	Frekuensi
1	en	3
2	ng	3
3	_m	2
4	am	2
5	as	2
6	de	2
7	i_	2
8	ka	2
9	me	2
10	_a	1
11	_b	1
12	_n	1
13	_p	1

No	Bi-gram	Frekuensi
14	a_	1
15	ah	1
16	ak	1
17	al	1
18	an	1
19	ba	1
20	e_	1
21	et	1
22	fi	1
23	gg	1
24	gi	1
25	gr	1
26	gu	1
27	ha	1
28	id	1
29	if	1
30	ik	1
31	la	1
32	m_	1
33	mi	1
34	n_	1
35	na	1
36	nt	1
37	od	1
38	pe	1
39	ra	1
40	sa	1
41	si	1
42	ti	1
43	to	1
44	un	1

Sedangkan untuk dokumen B akan menghasilkan profil seperti pada Tabel III-1:

Tabel III-4 Keluaran Pembangkitan Profil Dokumen B

No	Bi-gram	Frekuensi
1	ng	3
2	_n	2
3	at	2
4	ra	2
5	ti	2
6	_b	1
7	_i	1
8	_l	1
9	_m	1
10	_u	1
11	ag	1
12	al	1
13	am	1
14	an	1
15	by	1
16	ca	1
17	d_	1
18	de	1
19	e_	1
20	en	1
21	et	1
22	fi	1
23	g_	1
24	ge	1
25	gg	1
26	gr	1
27	gu	1
28	ho	1
29	ic	1

No	Bi-gram	Frekuensi
30	id	1
31	if	1
32	in	1
33	io	1
34	l_	1
35	la	1
36	m_	1
37	me	1
38	n_	1
39	na	1
40	nt	1
41	od	1
42	on	1
43	si	1
44	th	1
45	tu	1
46	ua	1
47	ur	1
48	us	1
49	y_	1

Untuk kemudahan pembacaan, hanya diberikan contoh menyangkut pembangkitan dan pengurutan dengan menggunakan *bi-gram*. Pada prakteknya, *n-gram* yang digunakan dapat memiliki panjang yang berbeda-beda, misalnya profil yang menggunakan *bi-gram* dan *tri-gram*. Efektifitas panjang *n-gram* yang dipilih merupakan salah satu hal yang akan diuji setelah perangkat lunak ini dibuat.

Tabel III-3 dan Tabel III-1 memperlihatkan dua buah profil dokumen yang dibentuk berdasarkan frekuensi kemunculan *bi-gram*. Dapat disimpulkan ada tiga proses utama dalam membentuk suatu profil dokumen. Pertama adalah membangkitkan *n-gram* dan (kedua) menghitung frekuensi kemunculannya, sebagaimana telah disebutkan pada awal bab ini. Kedua proses ini dapat dilakukan

secara hampir bersamaan, dimana setiap *n-gram* yang baru saja di bangkitkan di masukkan kedalam *hash table* dimana setiap kemunculan *n-gram* dapat dicatat. Proses ketiga adalah pengurutan berdasarkan frekuensi kemunculan *n-gram* tersebut. Daftar urutan tingkat kemunculan (*rank*) inilah yang disebut sebagai profil dokumen. Perlu dicatat bahwa proses pengurutan dilakukan setelah proses pembangkitan dan perhitungan frekuensi kemunculan selesai, dengan kata lain pengurutan dilakukan setelah pembangkitan *n-gram* selesai dilakukan. Penggunaan kedua profil ini untuk menentukan bahasa yang dipakai pada suatu dokumen, dapat dilihat pada sub bab selanjutnya.

3.4 Menentukan Bahasa yang Digunakan Dokumen

Proses selanjutnya adalah menentukan bahasa yang digunakan oleh dokumen. Hal ini dilakukan dengan membandingkan profil dokumen yang bahasanya belum diketahui dengan profil dokumen yang bahasanya telah diketahui. Proses perbandingan dilakukan dengan cara menghitung jarak antar kedua profil dokumen tersebut.

Jarak antar profil dihitung berdasarkan jarak antar peringkat *n-gram* di dokumen yang satu dengan peringkat *n-gram* yang sama pada dokumen yang lainnya. Sebagai contoh: misalnya pada dokumen pertama “_T” merupakan *n-gram* yang paling sering muncul sehingga menempati urutan pertama sedangkan pada dokumen kedua, *n-gram* tersebut menempati urutan ke-lima. Maka jarak antar *n-gram* itu adalah empat. Demikian selanjutnya, setelah hal tersebut dilakukan terhadap seluruh *n-gram* yang ada di dokumen maka total jarak yang didapat adalah jarak antar profil dokumen. Secara matematis jika dokumen A dilambangkan dengan D_A dan dokumen B D_B dimana panjang frekuensi pada dokumen A adalah L_A . Maka jarak antar dokumen A dan dokumen B dapat ditulis sebagai

$$Jarak(D_A, D_B) = \sum_{i=1}^{i=L_A} |Rank(D_A, i) - Rank(D_B, i)|$$

(III-1)

Sebagai contoh berikut ini adalah penghitungan jarak dua dokumen yang telah di buat profilnya pada Bab III-3.

Tabel III-5 Penghitungan Jarak Antar Dokumen

Dokumen A		Dokumen B		Jarak
Rank	<i>Bi-gram</i>	Rank	<i>Bi-gram</i>	
1	en	20	en	19
2	ng	-	-	47
3	_m	9	_m	6
4	am	13	am	9
5	as	-	-	44
6	de	18	de	12
7	i_	-	-	42
8	ka	-	-	41
9	me	37	me	28
10	_a	-	-	39
11	_b	6	_b	5
12	_n	-	-	37
13	_p	-	-	36
14	a_	-	-	35
15	ah	-	-	34
16	ak	-	-	33
17	al	12	al	5
18	an	14	an	4
19	ba	-	-	30
20	e_	19	e_	1
21	et	21	et	0
22	fi	22	fi	0
23	gg	25	gg	2
24	gi	-	-	25
25	gr	26	gr	1
26	gu	27	gu	1
27	ha	-	-	22
28	id	30	id	2
29	if	31	if	2
30	ik	-	-	19
31	la	35	la	4
32	m_	36	m_	4
33	mi	-	-	16
34	n_	38	n_	4
35	na	39	na	4
36	nt	40	nt	4
37	od	41	od	4
38	pe	-	-	11
39	ra	-	-	10
40	sa	-	-	9
41	si	43	si	2
42	ti	-	-	7

Dokumen A	
43	to
44	un

Dokumen B		Jarak
-	-	6
-	-	5
Jumlah jarak		671

Seperti terlihat pada Tabel III-5, Pada Dokumen A *rank* ke satu ditempati oleh *bi-gram* “en”. Hasil pencarian pada profil dokumen kedua menyatakan bahwa “en” menempati urutan 20. Hal ini menyebabkan jarak antar kedua *bi-gram* tersebut adalah: $20 - 1 = 19$.

Untuk *bi-gram* pada dokumen pertama yang tidak ditemukan di dokumen kedua, diberi nilai jarak maksimum yaitu selisih antara jumlah *bi-gram* yang ada pada dokumen kedua dengan rank dari *bi-gram* tersebut di dokumen pertama. Pendekatan ini lebih dipilih dibanding pemberian nilai maksimum secara statis, dikarenakan dengan pendekatan ini rank suatu *bi-gram* di dokumen pertama masih diperhitungkan. Sedangkan dengan pemberian nilai maksimum secara statis, misalnya setiap *bi-gram* yang tidak dapat ditemukan di dokumen ke-dua di beri jarak dengan suatu nilai tertentu, rank *bi-gram* di dokumen pertama tidak lagi diperhitungkan.

Cara perhitungan jarak tersebut juga dapat digunakan pada profil dokumen yang menggunakan berbagai panjang *n-gram* secara bersamaan (misalnya profil dokumen dibuat dengan *bi-gram* dan *tri-gram*). Tentunya hal tersebut akan mengakibatkan profil semakin besar. Dikarenakan jumlah *n-gram* yang dibangkitkan hanya bergantung kepada panjang kata, maka suatu profil yang menggunakan *n-gram* dengan dua panjang karakter yang berbeda (misalkan *bi-gram* dan *tri-gram*) akan memiliki panjang profil dua kali lipatnya. Karena proses pembangkitan profil melibatkan proses pengurutan maka profil yang terlalu panjang dapat memperlama proses identifikasi.

Hal lain yang perlu dicermati adalah jumlah kemunculan tidak perlu dicatat pada profil dokumen. Hal ini disebabkan karena jumlah kemunculan tidak dimasukkan dalam formulasi penghitungan jarak. Walaupun pada Tabel III-3 frekuensi kemunculan suatu *n-gram* masih dimasukkan kedalam suatu profil pada

prakteknya jumlah kemunculan tidak lagi perlu dicatat dalam suatu profile setelah *n-gram* di dalamnya diurutkan. Pada tugas kali ini jumlah kemunculan *n-gram* dicatat sebagai tindakan berjaga-jaga dan akan digunakan sebagai salah satu variabel yang perlu diperhitungkan dalam proses pengujian nanti

Proses menentukan bahasa dokumen dilakukan dengan mencari dokumen yang memiliki jarak paling kecil dengan dokumen yang bahasanya belum diketahui. Dengan kata lain dokumen yang belum diketahui bahasanya akan dihitung jaraknya dengan seluruh koleksi dokumen yang telah diketahui bahasanya. Bahasa dari dokumen yang memiliki jarak paling kecil tersebut yang akan dipilih sebagai bahasa dari dokumen yang belum diketahui.

Misalkan Dokumen A dan Dokumen B yang profilnya telah dibangkitkan pada sub bab 3.3 adalah kamus bahasa yang dimiliki, dimana bahasa kedua dokumen tersebut telah diketahui. Suatu dokumen yang belum diketahui bahasanya (Dokumen C) memiliki isi: “the quick brown fox jump over the lazy dog”. Hasil pembangkitan profil dapat dilihat pada

Tabel III-6 Keluaran Pembangkitan Profil Dokumen C

No	Bi-gram	Frekuensi
1	_t	2
2	e_	2
3	he	2
4	th	2
5	_b	1
6	_d	1
7	_f	1
8	_j	1
9	_l	1
10	_o	1
11	_q	1
12	az	1
13	br	1

No	Bi-gram	Frekuensi
14	ck	1
15	do	1
16	er	1
17	fo	1
18	g_	1
19	ic	1
20	ju	1
21	k_	1
22	la	1
23	mp	1
24	n_	1
25	og	1
26	ov	1
27	ow	1
28	ox	1
29	ps	1
30	qu	1
31	r_	1
32	ro	1
33	s_	1
34	ui	1
35	um	1
36	ve	1
37	wn	1
38	x_	1
39	y_	1
40	zy	1

Hasil penghitungan jarak antara Dokumen C dengan Dokumen A adalah 896. Sedangkan jarak Dokumen C dengan Dokumen B adalah 983. Dikarenakan jarak dengan Dokumen A adalah jarak yang paling kecil maka Bahasa yang dipilih adalah bahasa yang digunakan oleh Dokumen A.