PERHITUNGAN DISTRIBUSI PROBABILITAS KARAKTER DALAM BAHASA INDONESIA

SYUKRIYANTO [1]
L2F300564

ABSTRAK *- Perkembangan ilmu pengetahuan dan teknologi saat ini semakin pesat, dimana peranan komputer menjadi hal yang paling mendasar dan mutlak diperlukan dalam karya kreatif di bidang teknik dan ilmu pengetahuan. Institusi akademis semakin menyadari pentingnya penggunaan komputer dan penguasaan software yang dapat digunakan untuk analisis data, visualisasi, presentasi grafik, array dan struktur data, konstruksi bahasa, dan pembuatan aplikasi. Teori probabilitas mempelajari rata-rata gejala massa yang terjadi secara berurutan atau bersama-sama, seperti pancaran elektron, hubungan telepon, deteksi radar, pengendalian kualitas, kegagalan sistem, permainan untung-untungan, mekanika statistik, turbulen, gangguan, laju kelahiran dan kematian serta teori antrian.*

Program Perhitungan distribusi probablitas karakter dalam Bahasa Indonesia ini diperlukan untuk memudahkan menghitung jumlah karakter dan jumlah distribusi probabilitas karakter alphabet dalam karakter ASCII suatu data atau file, pembentukan kode morse dan kode-kode komersial untuk Bahasa Indonesia seperti halnya dalam telegraf, penghematan besarnya kapasitas data (kompresi data) dalam teknologi yang membutuhkan alokasi data, dan sebagai titik awal bagi rancangan pembuatan keyboard (papan ketik) komputer dalam versi Bahasa Indonesia agar lebih mudah dari apa yang telah digunakan sekarang ini yakni letak atau posisi beberapa tombol karakter alphabet amatlah tidak tepat bila mengetik dengan sistem sepuluh jari yang menggunakan teks Bahasa Indonesia.

Jumlah probabilitas distribusi suatu karakter tergantung dalam jenis tulisan (tulisan dalam bentuk Bahasa Ingggris dan Bahasa Indonesia). Probabilitas karakter A dalam teks Bahasa Indonesia dan probabilitas karakter E dalam teks Bahasa Inggris untuk bahasa formal dan non formal lebih sering muncul dibandingkan karakter yang lain.

I. PENDAHULUAN

ILMU pengetahuan dan teknologi saat ini semakin pesat, peranan komputer menjadi hal yang paling mendasar dan mutlak diperlukan dalam karya kreatif di bidang teknik dan ilmu pengetahuan. Institusi akademis semakin menyadari pentingnya penggunaan komputer dan penguasaan *software* yang dapat digunakan untuk analisis data, visualisasi, presentasi grafik, array dan struktur data, konstruksi bahasa, dan pembuatan aplikasi.

Perkembangan teknologi komputer yang sangat pesat ini, memicu perkembangan di berbagai bidang lain yang salah satu diantaranya adalah Teknologi Bahasa Manusia (TBM). TBM pada dasarnya dapat dikelompokkan ke dalam dua kelompok utama yaitu Sistem Bahasa Ucapan (SBU) dan Sistem Bahasa Tertulis (SBT). Dalam kelompok pertama yaitu SBU, kegiatan yang dilakukan adalah pengembangan sistem pengenalan ucapan *(speech recognition*) dengan model *hidden* Markov dan jaringan syaraf terpadu *(neural network)*, pengembangan sistem pembangkit ucapan dari suatu teks *(text-to-speech synthesis system)*, sistem pemahaman bahasa alami dan pemodelan dialog. Kegiatan yang dilakukan dalam SBT diantaranya adalah pengembangan OCR *(Optical Character Reader)*, Pemeriksa Ejaan *(Spell Checker)*, Pemeriksa Tatabahasa *(Grammar Checker)*, Pemenggalan Kata *(hyphenation)*, Sistem Temubalik Informasi *(Information Retrieval)*, sistem Mesin Penerjemah dan yang paling sederhana mungkin adalah sistem Perhitungan Distribusi Probabilitas Karakter dalam Bahasa Indonesia.

[1] Jurusan Teknik Elektro Fakultas Teknik Universitas Diponegoro Semarang.

II. DASAR TEORI

2.1 Teori Probabilitias

Teori probabilitas mempelajari rerata gejala massa yang terjadi secara berurutan atau bersama-sama, seperti pancaran elektron, hubungan telepon, deteksi radar, pengendalian kualitas, kegagalan sistem, permainan untung-untungan, mekanika statistik, turbulen, gangguan, laju kelahiran dan kematian serta teori antrian.

Tujuan teori probabilitas adalah menggambarkan dan menaksir rata-rata sedemikian itu dalam bentuk probabilitas peristiwa. Probabilitas peristiwa P(A) ditetapkan bagi peristiwa tersebut. Bilangan ini dapat ditafsirkan bahwa: "Bila suatu ekperimen dilakukan *n* kali dan peristiwa *A* terjadi n_A kali, maka dengan kepastian derajat tinggi, frekuensi relatif n_A / n mendekati *P(A):*

$$P(A) = nA / n \quad \ldots\ldots\ldots\ldots \quad (2.1)$$

asalkan *n* cukup besar".

2.2 Independensi

Dua peristiwa A dan B disebut independen bila:

$$P(AB) = P(A)\,P(B) \quad \ldots\ldots \quad (2.2)$$

Konsep independen adalah konsep pokok. Kenyataanya, konsep ini membenarkan perkembangan matematis probabilitas, tidak hanya sebagai topik dalam teori ukuran, tetapi juga merupakan disiplin terpisah. Hal ini amatlah penting dalam pembahasan selanjutnya. Beberapa sifat sederhana akan dibahas kemudian adalah dalam interpretasi frekuensi, misalkan n_A, n_B, dan n_{AB} masing-masing menyatakan jumlah terjadinya peristiwa A, B, dan AB, maka:

$$P(A) = \frac{n_A}{n} \; ; P(B) = \frac{n_B}{n} ; P(AB) = \frac{n_{AB}}{n_B} \quad \ldots.. \; (2.3)$$

Bila peristiwa A dan B independen, maka frekuensi relatif n_A /n kejadian A pada bagian *n* usaha awal sama dengan frekuensi relatif n_{AB} /n_B kejadian A pada barisan bagian dimana B telah terjadi

2.3 Fungsi Distribusi

Beberapa elemen himpunan *S* yang termuat dalam peristiwa $\{X \leq x\}$ berubah bila nilai *x* mengalami perubahan. Akibatnya, probabilitas $P\{X \leq x\}$adalah bilangan yang bergantung pada *x*. Bilangan ini dinyatakan dengan $F_X(x)$ dan disebut fungsi distribusi (kumulatif) dari variabel random *X* . Jadi,

$$F_X(x) = P\{X \leq x\} \quad \ldots\ldots\ldots \quad (2.4)$$

didefinisikan untuk setiap *x* dari - ∞ sampai ∞.

2.4 Model Distribusi Statistik Karakter

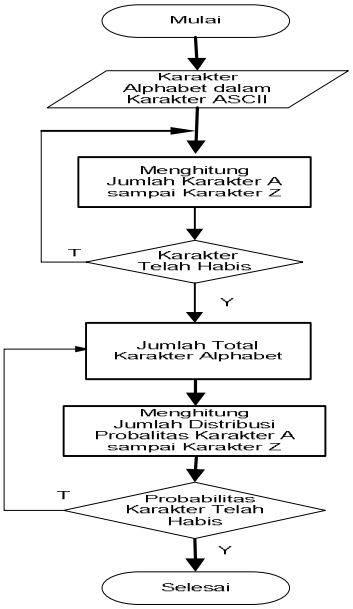
Beberapa cara untuk yang dapat digunakan untuk mengatur model sifat statistik karakter adalah:

a. *Zero-Order-Model*: Setiap karakter secara statistik berdiri sendiri dari semua karakter yang lain dan 26 nilai mungkin sama dengan yang terdapat dalam alfabet ***A***.
b. *First-Order Model*: Dalam Bahasa Indonesia, beberapa huruf terjadi lebih banyak pengulangan dibandingkan dengan huruf yang lain. Sebagai contoh, pada huruf 'a' dan 'e' lebih umum dibanding 'q' dan 'z'. Jadi dalam model ini, karakter masih berdiri sendiri dari yang lain, tetapi distribusi probabilitas dari beberapa karakter adalah menurut *first-order Statistical distribution of English text.*
c. *Second-Order Model:* Dua model sebelumnya diasumsikan bebas secara statistik dari satu karakter berikutnya sebagai contoh, beberapa kali#at i#i hurufnya kehil#ngan, akan tetapi masih dapat dipahami tulisan apa yang dimaksud dengan melihat konteksnya. Ini menyatakan secara tidak langsung bahwa ada beberapa ketergantungan di antara beberapa karakter.
d. *Third-Order Model*: Ini adalah lanjutan model sebelumnya. Disini, karakter sekarang yakni X_i bergantung pada dua karakter sebelumnya: $(X_i, X_2, \ldots, X_{i-3})$, tetapi secara kondisional berdiri sendiri dari semua karakter sebelumnya. Dalam model ini, distribusi karakter dari X_i berubah menurut (X_{i-2}, X_{i-1}).
e. *General Model*: Dalam model ini, buku *X* berubah-ubah secara acak dan stasioner. Sifat statistic dari model ini terlalu rumit bila diaplikasikan. Model ini hanya menarik dari titik pandang teori.

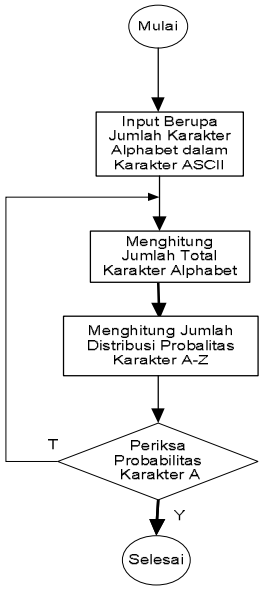
III. PERANCANGAN PROGRAM

3.1 Diagram Alir Program

Gambar 3.1 adalah gambar diagram alir program untuk menghitung distribusi probabilitas karakter alphabet Bahasa Indonesia dalam karakter ASCII.



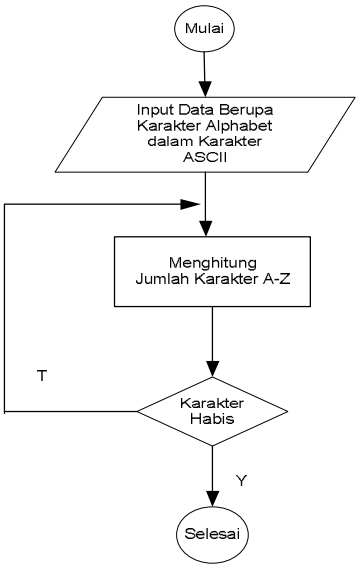
Gambar 3.1 Diagram Perhitungan Distribusi Karakter



Gambar 3.3 Menghitung Probablitas Karakter

3.2 Menghitung Jumlah Karakter

Pada Gambar 3.2 diperlihatkan diagram alir untuk menghitung jumlah Karakter.



Gambar 3.2 Menghitung Jumlah Karakter

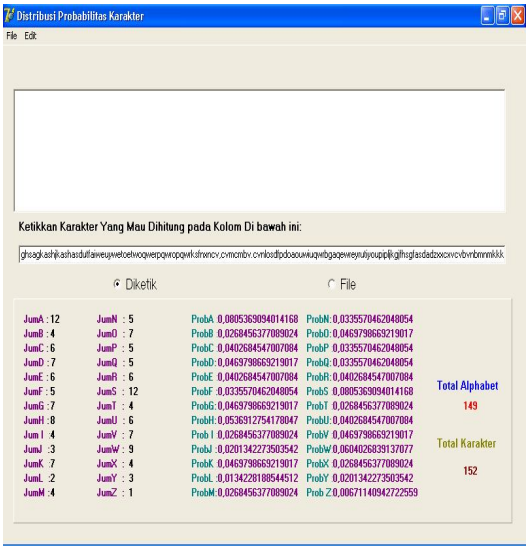
3.3 Menghitung Probabilitas Karakter

Pada Gambar 3.3 diperlihatkan diagram alir untuk menghitung jumlah distribusi probabilitas karakter.

IV. PENGUJIAN DAN ANALISIS

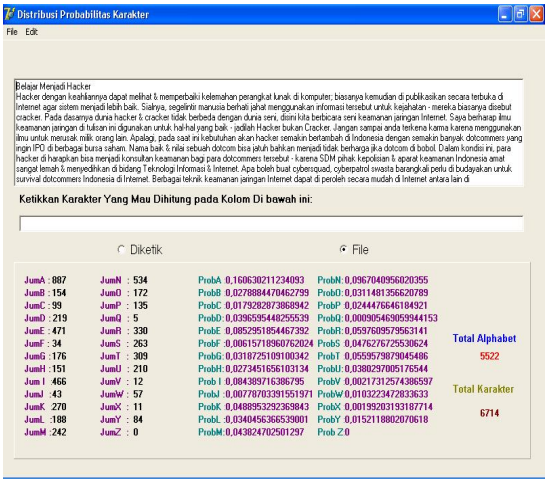
4.1 Menu Program

Gambar 4.2 adalah tampilan program untuk proses manual (diketik langsung) pada kolom yang disediakan.



Gambar 4.2 Menu Program Setelah Dieksekusi (Diketik Manual)

Gambar 4.3 adalah tampilan program setelah dieksekusi diambil dari data yang telah tersedia (*load from file*).



Gambar 4.3 Menu Program untuk Proses *Load from File*

4.2 Distribusi Probabilitas Karakter dalam Bahasa Indonesia

A. Bahasa Non Formal

Tabel 4.1 merupakan tabel jumlah distribusi probabilitas karakter dalam Bahasa Indonesia untuk bahasa non formal, diambil dari salah satu tulisan non formal (majalah) yang terdiri atas 5.522 total karakter alphabet.

Tabel 4.1 Distribusi probabilitas Karakter Bahasa Indonesia (Non Formal).

No.	Karakter	Jumlah Karakter	Probablititas
1.	A	887	0,16063
2.	N	534	0,09670
3.	E	471	0,08529
4.	I	466	0,08438
5.	R	330	0,05976
6.	T	309	0,05595
7.	K	270	0,04889
8.	S	263	0,04762
9.	M	242	0,04382
10.	D	219	0,03965
11.	U	210	0,03802
12.	L	188	0,03404
13.	G	176	0,03187
14.	O	172	0,03114
15.	B	154	0,02788
16.	H	151	0,02734
17.	P	135	0,02444
18.	C	99	0,01792
19.	Y	84	0,01521
20.	W	57	0,01032
21.	J	43	0,00778
22.	F	34	0,00615
23.	V	12	0,00217

Lanjutan Tabel 4.1

No.	Karakter	Jumlah Karakter	Probabilitas
24.	X	11	0,00199
25.	Q	5	0,00090
26.	Z	0	0

B. Bahasa Formal

Tabel 4.2 merupakan tabel jumlah distribusi probabilitas karakter dalam Bahasa Indonesia untuk bahasa formal, diambil dari salah satu tulisan formal (tulisan ilmiah) yang terdiri atas 50.598 total karakter alphabet.

Tabel 4.2 Distribusi probabilitas Karakter Bahasa Indonesia (Formal).

No.	Karakter	Jumlah Karakter	Probabilitas
1.	A	8604	0,17004
2.	N	5055	0,09990
3.	I	4216	0,08332
4.	E	4103	0,08109
5.	T	2900	0,05731
6.	S	2851	0,05634
7.	R	2808	0,05549
8.	U	2604	0,05146
9.	D	2309	0,04563
10.	G	2087	0,04124
11.	K	2045	0,04041
12.	M	1938	0,03830
13.	P	1824	0,03604
14.	L	1804	0,03559
15.	B	1141	0,02250
16.	O	1116	0,02205
17.	Y	757	0,01496
18.	H	721	0,01424
19.	C	448	0,00885
20.	F	423	0,00836
21.	J	373	0,00737
22.	W	252	0,00498
23.	V	101	0,00199
24.	Z	83	0,00164
25.	X	21	0,00041
26.	Q	17	0.00033

Berdasarkan pada Tabel 4.1 dan Tabel 4.2, maka dapat dikatakan bahwa dalam teks Bahasa Indonesia untuk bahasa formal maupun non formal, karakter A, N, I, E merupakan karakter yang paling sering digunakan, sedangkan karakter Q, X, Z, W, dan V merupakan karakter yang jarang sekali digunakan.

4.3 Distribusi Probabilitas Karakter dalam Bahasa Inggris

A. Bahasa Non Formal

Tabel 4.3 merupakan tabel jumlah distribusi probabilitas karakter dalam Bahasa Inggris untuk bahasa non formal, diambil dari salah satu contoh tulisan non formal yang terdiri atas 6.928 total karakter alphabet.

Tabel 4.3 Distribusi probabilitas Karakter Bahasa Inggris (Non Formal).

No.	Karakter	Jumlah Karakter	Probabilitas
1.	E	838	0,12095
2.	T	659	0,09512
3.	A	602	0,08689
4.	O	528	0,07621
5.	S	525	0,07577
6.	I	523	0,07549
7.	N	485	0,07000
8.	R	428	0,06177
9.	H	310	0,04474
10.	L	249	0,03594
11.	D	245	0,03536
12.	C	230	0,03319
13.	P	213	0,03074
14.	U	182	0,02627
15.	M	158	0,02280
16.	F	145	0,02092
17.	G	112	0,01616
18.	B	111	0,01602
19.	W	104	0,01501
20.	Y	100	0,01443
21.	V	82	0,01183
22.	K	77	0,01111
23.	X	10	0,00144
24.	J	7	0,00101
25.	Z	4	0,00057
26.	Q	1	0,00014

B. Bahasa Formal

Tabel 4.4 merupakan tabel jumlah distribusi probabilitas karakter dalam Bahasa Inggris untuk bahasa formal, diambil dari salah satu tulisan formal yang telah ada, yang terdiri atas 4.710 total karakter alphabet.

Tabel 4.4 Distribusi probabilitas Karakter Bahasa Inggris (Formal).

No.	Karakter	Jumlah Karakter	Probabilitas
1.	E	549	0,11656
2.	T	410	0,08704
3.	I	401	0,08513
4.	A	391	0,08301
5.	N	364	0,07728

Lanjutan Tabel 4.4

No.	Karakter	Jumlah Karakter	Probabilitas
6.	R	355	0,07537
7.	S	351	0,07452
8.	O	335	0,07112
9.	L	193	0,04097
10.	C	188	0,03991
11.	D	182	0,03864
12.	H	173	0,03673
13.	U	146	0,03099
14.	M	127	0,02696
15.	F	98	0,02080
16.	B	96	0,02038
17.	P	86	0,01825
18.	G	79	0,01677
19.	Y	64	0,01358
20.	W	46	0,00976
21.	V	43	0,00912
22.	J	16	0,00339
23.	K	8	0,00169
24.	X	6	0,00127
25.	Q	3	0,00063
26.	Z	0	0

Berdasarkan pada Tabel 4.3 dan Tabel 4.4, maka dapat dikatakan bahwa dalam Bahasa Inggris baik bahasa formal maupun non formal, karakter E, T, I, A, merupakan karakter yang paling sering digunakan, sedangkan karakter Q, X, Z, W, V merupakan karakter yang jarang sekali digunakan.

4.4. Perbandingan Distribusi Probabilitas Karakter dalam Bahasa Indonesia dan Bahasa Inggris

Tabel 4.5 merupakan tabel hasil perbandingan jumlah distribusi probabilitas karakter dalam Bahasa Indonesia dan Bahasa Inggris untuk karakter yang sering digunakan dan karakter yang jarang digunakan.

Tabel 4.5 Perbandingan Distribusi Probabilitas Karakter dalam Bahasa Indonesia dan Bahasa Inggris

Karakter	Bahasa Indonesia	
	Formal	Non Formal
A	0,17004	0,16063
E	0,08109	0,08529
I	0,08332	0,08438
N	0,09990	0,09670
T	0,05731	0,05595
Q	0,00033	0,00090
V	0,00199	0,00217
W	0,00498	0,01032
X	0,00041	0,00199
Z	0,00164	0

Karakter	Bahasa Inggris	
	Formal	Non Formal
E	0,11656	0,12095
T	0,08704	0,09512
I	0,08513	0,07549
A	0,08301	0,08689
N	0,07728	0,07000
V	0,00912	0,01183
X	0,00127	0,00144
J	0,00339	0,00101
Q	0,00063	0,00014
Z	0	0,00057

Berdasarkan Tabel 4.5, dapat dijelaskan bahwa dalam Bahasa Indonesia, baik bahasa formal maupun non formal, karakter A merupakan karakter yang paling sering digunakan, sedangkan dalam Bahasa Inggris, untuk bahasa formal maupun non formal karakter E merupakan karakter yang paling sering digunakan. Untuk beberapa karakater yang lain, baik itu dalam Bahasa Inggris maupun Bahasa Indonesia (formal dan non formal), jumlah distribusi probabilitas karakternya relatif hampir sama.

V. KESIMPULAN

1. Probabilitas suatu karakter amat sangat bergantung pada jenis tulisan (tulisan dalam bentuk Bahasa Inggris atau Bahasa Indonesia).
2. Dalam tulisan dengan menggunakan Bahasa Indonesia untuk bahasa formal dan bahasa non formal, jumlah probabilitas distibusi karakter A adalah yang paling banyak.
3. Dalam tulisan dengan menggunakan Bahasa Inggris untuk bahasa formal dan bahasa non formal, jumlah probabilitas distibusi karakter E adalah yang paling banyak.
4. Adanya distribusi probabilitas karakter ini, akan memberikan suatu penghematan pada waktu atau kapasitas saluran dengan penyandiaan, kompresi secara tepat deretan pesan ke dalam deretan sinyal.

V. DAFTAR PUSTAKA

[1] H., Nyquist. *Certain Factors Effecting Telegraph*, Bell System Technical Journal. April 1942, p.324, Certain Topics in Telegraph Transmission Theory, A.I.E.E Trans., volume 47, April 1928, p.617.

[2] Kadir, Abdul. *Dasar Pemrograman Delphi Versi 5,0*. Jilid 2. ANDI. Yogyakarta. 2001.

[3] Liu, C, L., *Dasar-dasar Matematika Deskret* dialihbahasakan oleh Bambang Sumantri. Edisi ke-2. PT. Gramedia Pustaka Utama. Jakarta. 1995.

[4] Miller, Irwin., Freund, E John. *Probability and Statistics for Engineers*. Prentice Hall, Englewood Cliffs. New Jersey. 1997.

[5] Nelson, Mark., Loup – Gailly, Jean. *The Data Compression Book*. Second Edition. M & T Book. New York 1995.

[6] Papoulis, Athanasios. *Probabilitas, Variabel Random, dan Proses Stokastik* diterjemahkan oleh Gadjah Mada University Press. Yogyakarta. 1992.

[7] Pranata, Anthony. *Pemrograman Borland Delphi Edisi 3*. ANDI. Yogyakarta. 2000.

[8] Peebles, Peyton Z., Jr. *Probability, Random Variables, and Random Signal Principles*. Fourth Edition. McGraw-Hill. New York. 2001.

[9] Roman, Steve., *Coding and Information Theory*. Springer-Verlag. New York, Berlin, London, Paris. 1992.

[10] R.V.L, Hartley. *Transmission of Information*. Bell System Technical Journal, July 1928, p. 534.

[11] Shannon, Claude, E., *A Mathematical Theory of Communication*. Paper Volume 27. 1948

[12] ---------- , WAHANA Komputer – Semarang. *Panduan Praktis Pemrograman Borland Delphi 5,0*. ANDI. Yogyakarta. 2001.

[13] ---------, www. Data-compression.com

[14] ---------, www. Epusbs.siam.org

[15] ---------, www. The Jakarta Post.com

[16] --------, LPKBM Madcoms - Madiun. *Paduan Lengkap Pemrograman Borland Delphi 5,0*. ANDI. Yogyakarta. 2001

Syukriyanto, lahir di Buton Sulawesi Tenggara pada tanggal 18 Desember 1977 dan sekarang masih berstatus Mahasiswa Jurusan Teknik Elektro Universitas Diponegoro Semarang

Mengetahui/Mensahkan,
Dosen Pembimbing II

Achmad Hidayatno, ST, MT
NIP: 132 137 933