Aplikasi Pencocokan String pada Penyaringan *Email Spam*

Amal Qurany
Program Studi Teknik Informatika
Sekolah Teknik Elektro dan Informatika
Institut Teknologi Bandung, Jalan Ganesha 10 Bandung 40132, Indonesia
13514078@std.stei.itb.ac.id

***Abstract*—Pada makalah ini dibahas mengenai aplikasi pencocokan string untuk menentukan apakah sebuah *email* termasuk *email spam* atau bukan, serta perbandungan kinerja antara metode pencocokan string *Bruteforce*, KMP, dan Boyer Moore dalam pencocokan string pada sebuah *email*.**

Keywords—sring matrching. Knot Moris Path, Boyer Moore, Naïve Bayes Classifier, spam filtering.

I. PENDAHULUAN

1. Latar Belakang

Perkembangan dunia teknologi memiliki dampak yang sangat besar bagi kehidupan manusia. Salah satu dampak yang disebabkan perkembangan teknologi adalah peralihan penggunaan surat kertas menggunakan jasa pos menjadi surat elektronik atau lebih biasa dikenal dengan sebutan *email.* Dewasa ini, surat elektronik (selanjutnya akan disebut *email*) telah menjadi kebutuhan mendasar bagi kebanyakan orang terutama di kalangan anak muda. Layanan-layanan di internet seperti media sosial dan aplikasi chatting pada umumnya mewajibkan pengguna untuk mendaftarkan *email* sebagai salah satu syarat untuk menggunakan layanan tersebut. Komunikasi resmi seperti lamaran pekerjaan juga menggunakan *email* sebagai alat komunikasi.Para pelajar dan mahasiswa menggunakan *email* untuk mengumpulkan tugas kuliah. Tata Usaha pun juga menggunakan *email* sebagai salah satu media dalam menyampaikan pengumuman. Meningkatnya penggunaan *email* memang memiliki dampak positif berupa kemudahan dalam komunikasi. Penggunaan *email* mengambil alih pangsa pasar surat konvensional yang memakan waktu berhari-hari dalam mengirim dan menunggu surat balasan karena dengan menggunakan *email*, pesan dapat dikirim dan dapat langsung dibalas dalam hitungan detik saja. Dalam hal ini tentu keberadaan *email* sangat membantu. Di sisi lain, kemudahan yang diberikan teknologi *email* ini justru dimanfaatkan sebagai media promosi produk bahkan sebagai media untuk melakukan penipuan. *Email* semacam ini sering dikenal dengan istilah *spam* atau *phising* untuk yang kategori penipuan. Para penyedia *email* seperti google dan yahoo harus mengantisipasi hal tersebut agar pengguna tidak dibanjiri dengan *email spam*. Terdapat beberapa teknik yang digunakan untuk melakukan penyaringan *email* spam diantaranya *checksum-based filtering, hybrid filtering, outbound spam protection, keyword blacklist, naïve bayes class,* dan masih banyak lagi.

2. Batasan

Pada makalah ini akan dibahas penerapan pencocokan teks (*string matching*) dalam mendeteksi *email spam*. Serta akan ada sedikit bahasan tentang metode *naive bayes classifier* dan metode *keyword blacklist filtering*. Namun makalah ini menitikberatkan pembahasan pada bagian pencocokan teks. Pembahasan metode *Naïve Bayes Classifier* dan *keyword blacklist filtering* tidak dibahas secara detail.

II. LANDASAN TEORI

A. Pencocokan String

Berikut beberapa terminologi yang harus diketahui sebelum memahami string matching:

1. **Teks (T)**, menyatakan sekumpulan kata atau kalimat yang akan diuji, dalam hal ini teks menyatakan isi email. Panjang teks dilambangkan dengan n.
2. ***Pattern* (P)**, menyatakan kata atau frasa yang akan dicari di dalam teks. Dalam hal ini panjang *pattern* dilambangkan dengan m dan diasumsikan m<<<n (m jauh lebih kecil daripada n).
3. **Prefix**. Jika S adalah sebuah substring, S[1..k-1] adalah prefix dari S.
4. **Suffix**. Jika S adalah sebuah substring, S[k-1 .. m] adalah suffix dari S.
5. **k** merupakan sebuah index diantara 1 sampai m
6. **S[0]** adalah karakter null, disimbolkan dengan ∅

   Misalkan S = amal, maka

   Prefix = {“∅”, “a”, “am”, “ama”}

Sufix = {“∅”, “l”, “al”, “mal”}

Pencocokan String, atau dalam bahasa inggris *String Matching*, merupakan suatu teknik yang digunakan untuk menemukan suatu pola atau *pattern* dalam sebuah teks dengan menentukan lokasi pertama di dalam teks yang bersesuaian dengan *pattern*.

Teknik pencocokan teks ini sangat luas digunakan terutama dalam dunia komputer. Beberapa contoh aplikasi pencocokan teks adalah:

- Fitur pencarian pada editor teks
- Teknologi mesin pencari
- Bidang Bioinformatika untuk pencocokan rantai DNA
- Untuk mendeteksi Plagiarisme
- *Digital Forensic*
- *Text Mining*

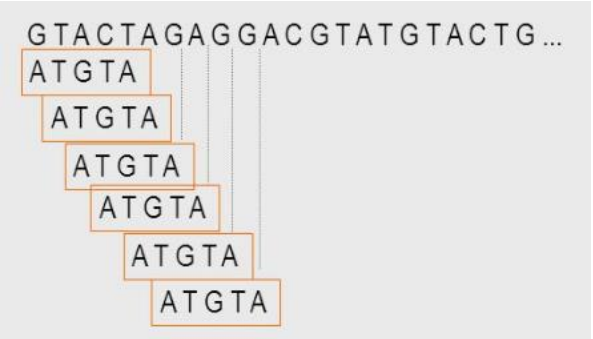
Terdapat beberapa teknik dalam melakukan string matching, diantaranya:

1. Naïve String Matching

   Pencarian jenis ini dinamakan naif karena teknik yang digunakan adalah teknik yang *simple-minded.* Pencocokan string menggunakan *Naïve String Matching* ini menerapkan algoritma *bruteforce* sehingga teknik ini sering juga disebut dengan *bruteforce* string matching.

   Pencocokan string menggunakan *Naïve String Matching* melakukan pencocokan dengan cara membangdingkan setiap substring pada teks dengan panjang m (panjang *pattern*) secara sekuensial dari substring pertama hingga substring terakhir. Setiap substring dilakukan pengujian masing-masing karakternya dari karakter pertama hingga karakter ke m. Jika ditemukan karakter yang tidak cocok dalam pengujian suatu substring, pengujian langsung dilanjutkan dengan menguji kesamaan *pattern* dengan substring berikutnya. Pencarian berhenti ketika ditemukan substring yang bersesuaian dengan *pattern* atau *pattern* tidak ditemukan (tidak ada substring yang sesuai dengan *pattern*).

   Berikut ilustrasi dari Naïve String Matching:



Gambar 1. Ilustrasi *Naïve String Matching*
Sumber:http://images.slideplayer.com/13/4035194/slides/slide_7.jpg

Case	Time Complexity
Best Case	O(n)
Average	O(m+n)
Worst Case	O(mn)

Tabel 1. Kompleksitas *Bruteforce String Matching*

Algoritma *Naïve String Matching* ini cukup mangkus ketika digunakan pada string berukuran besar, missal: A..Z, a..z, 1..9. Sebaliknya, Naïve String Search tidak efisien jika digunakan pada string berukuran kecil misal: bilangan biner.

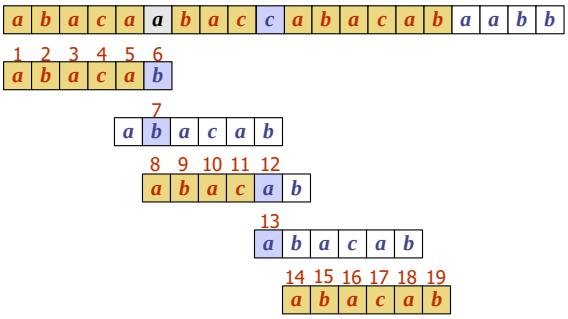
2. Knuth Morris Pratt (KMP)

   Metode pencocokan string jenis ini juga mencocokkan string dari kiri ke kanan seperti halnya metode *Naïve String Matching*, namun penggeseran *pattern* atau pemilihan substring dilakukan secara lebih cerdas.

   Terdapat istilah fungsi pinggiran *(border function)* dalam melakukan pencocokan string dengan metode Knuth Morris Pratt ini. Fungsi pinggiran b(k) didefenisikan sebagai jumlah karakter maksimal pada prefix P[1..k] yang juga merupakan suffix P[1..k].

   Penggeseran *pattern* pada string dilakukan sebanyak b(k)+1 karakter dimana k adalah indeks karakter pada *pattern* yang pengujiannya bernilai *fail*.

   Pencocokan String jenis KMP ini memperbaiki ketidakefisienan pencocokan string dengan motode *Naive String Matching.*



Gambar 2. Ilustrasi KMP *String Matching*
sumber: Davison, Andrew. *Pattern Matching*, 2006

3. Boyer Moore

Dalam metode *Boyer Moore* dikenal istilah *last occurrence* yang menyatakan posisi terakhir suatu karakter pada *pattern.* Karakter yang dilihat last occurrence nya adalah karakter pada teks yang pengujian kesamaannya bernilai *fail*.

Pencocokan string dengan metode Boyer Moore ini menggunakan dua teknik sebagai berikut:

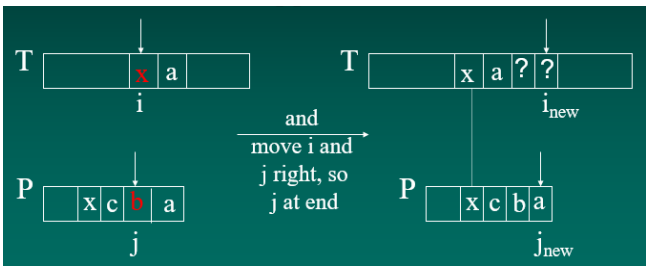
a. *The looking-glass technique*

Teknik ini mencari *pattern* pada teks dengan menguji karakter pada P dari belakang (indeks tertinggi)

b. *The Character-jump*

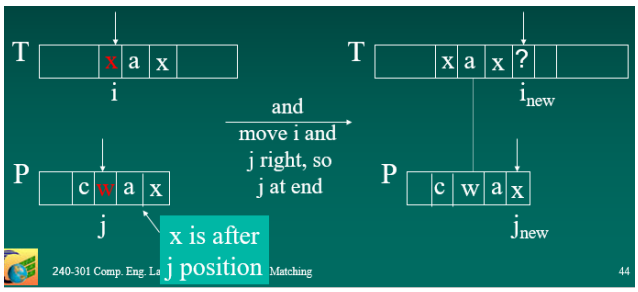
Saat terjadi *mismatch* dalam pengujian karakter, terdapat tiga buah kasus dalam penggeseran *pattern* terhadap teks. Kasus-kasus tersebut berhubungan dengan keberadaan karakter pada *pattern*. Ketiga kasus tersebut dijelaskan sebagai berikut:

- Kasus 1, terjadi ketika *last occurrence* terdapat di sisi kiri karakter yang diuji pada *pattern*. Pada kasus ini, *pattern* digeser ke kanan sehingga karakter yang bersesuaian berada pada posisi sejajar.



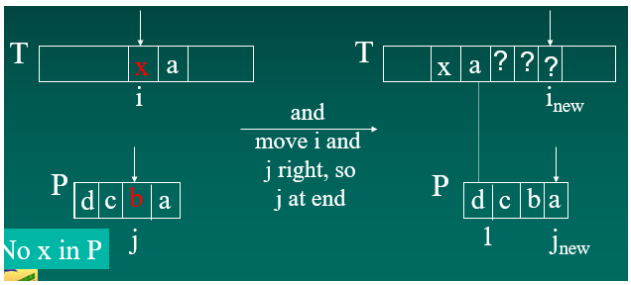
Gambar 3-Kasus pertama aturan Booyer Moore
sumber: Davison, Andrew. *Pattern Matching*, 2006

- Kasus 2, terjeadi ketika posisi *last occurrence* karakter yang diuji pada teks terdapat di sebelah kanan karakter yang diuji pada *pattern*. Pada kasus ini, *pattern* digeser ke kanan satu karakter.



Gambar 4. Kasus ketiga aturan Booyer Moore
sumber: Davison, Andrew. *Pattern Matching*, 2006

- Kasus 3, terjadi ketika karakter yang diuji pada teks tidak terdapat pada *pattern*. Pada kasus ini, *pattern* digeser ke kanan sehingga karakter awal *pattern* sejajar dengan satu karakter sebelah kanan karakter teks yang diuji.



Gambar 4. Kasus kedua aturan *Booyer Moore*
sumber: Davison, Andrew. *Pattern Matching*, 2006

B. Teorema Bayes

Teorema Bayes merupakan salah satu pokok bahasan dalam ilmu probabilitas dan statistika. Teorema bayes menggambarkan hubungan antara peluang bersyarat dari dua kejadian. Misal dua kejadian tersebut adalah A dan B, maka teorema bayes dituliskan sebagai berikut:

$$P(A \mid B) = \frac{P(B \mid A)\,P(A)}{P(B)}$$

atau

$$P(A \mid B) = \frac{P(B \mid A)\,P(A)}{P(B \mid A)P(A) + P(B \mid \overline{A})P(\overline{A})}$$

P(A) = Peluang kejadian A
P(A|B) = Peluang kejadian A dengan diketahui kejadian B terjadi.

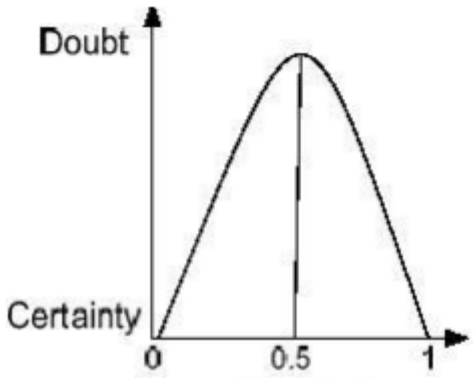
Teorema Bayes sering pula dikembangkan mengingat berlakunya hukum probabilitas total menjadi sebagai berikut:

$$P(A|B) = \frac{P(A)P(B|A)}{\sum_{i=1}^{n} P(Ai|B)}$$

Dimana A_1 U A_2 ..U .. An = S

C. Naive Bayes Classifier

Naïve Bayes Classifier merupakan salah satu teknik klasifikas dokumen yang banyak digunakan pada saat ini. Teknik ini melakukan pencocokkan teks dengan kata yang terdapat dalam daftar kata (dalam hal ini adalah daftar kata yang mengindikasikan sebuah *email* merupakan spam atau tidak). *Naïve Bayes Classifier* menggunakan teorema bayes dalam perhitungannya. Output dari metode klasifikasi ini adalah sebuah nilai probabilitas.



Naïve Bayes Classifier memiliki rumus sebagai berikut

$$P(X \mid C_i,.....,C_n) = \frac{1}{Z} P(X) \prod_{k=1}^{n} P(C_k \mid X)$$

dan peluang sebuah email dikatakan spam dapat ditentukan dengan persamaan sebagai berikut:

$$P(spam) = \frac{\prod_{k=1}^{K} P_k}{\prod_{k-1}^{K} P_k + \prod_{k=1}^{K} (1 - P_k)}$$

dimana K adalah jumlah kata yang terdapat dalam email, dan P_k didefenisaikan sebagai berikut:

$$P_k = P(spam \mid W_k)$$

D. Keyword Blacklist Filtering

Metode *keyword blacklist filtering* merupakan metode primitif dalam penyaringan sebuah dokumen atau teks. Metode ini biasanya digunakan aplikasi pengelola SMS pada smartphone untuk memblacklist SMS promosi yang masuk ke *inbox*. Jika digunakan untuk mendeteksi *email spam*, metode ini memiliki kemungkinan kesalahan yang besar karena jika terdapat satu saja kata atau frasa yang terindikasi, *email* langsung diklasifikasikan sebagai *spam*

.

III. PEMBAHASAN

Meningkatnya penggunaan *email* di internet merupakan peluang tersendiri bagai para pelaku marketing untuk memasarkan produknya. Banyak orang atau perusahaan saat ini memanfaatkan *email* sebagi media untuk mempromosikan produknya. Jika promosi tersebut adalah adalah kehendak *user* atau atas persetujuan pengguna *email* maka hal itu tentu saja bernilai positif karena *user* secara tidak langsung dimudahkan untuk mendapatkan *update* informasi yang ingin diketahuinya.

Di sisi lain, terdapat pihak yang kurang bertanggung jawab yang mengirimkan *email* promosi secara acak kepada pengguna layanan *email*. Jika hal ini terjadi dengan intensitas yang tinggi tentu penerima *email* tersebut akan merasa tidak nyaman dengan masuknya *email* yang tidak diinginkan ke *inbox*-nya. Apalagi *email* penipuan yang mengiming-imingkan *user* dengan suatu hadiah. Telah banyak kasus pengguna awam yang tertipu dengan *email* semacam itu.

Penyedia layanan *email* telah mengantisipasi hal tersebut. Salah satu metoda yang digunakan penyedia *email* adalah menggunakan teknik *naïve bayes classifier*. Teknik ini cukup terkenal dalam mengelompokkan sebuah data teks karena kemampuannya yang cukup ampuh. Teknik ini menggunakan probabilitas dalam perhitungannya. Jika dibandingkan dengan *spam filtering* yang menggunakan sistem *black list*, metode *Naïve Bayes Classifier* ini jauh lebih efektif. Penyaringan *email* dengan sistim blacklist menghasilkan *false positif* yang tinggi sehingga rentan terjadi kesalahan dalam mengklasifikasikan *email* spam. Kedua teknik *email filtering* yang disebutkan diatas mengaplikasikan ilmu pencocokan string atau lebih dikenal dengan bahasa *string matching*.

Metode Naïve Bayes Classification menggunakan daftar kata dan klasifikasi dari masing-masing daftar kata tersebut. Dalam menentukan apakah sebuah *email* merupakan *email spam* atau bukan, digunakan daftar kata yang mengindikasikan sebuah *email* merupakan spam. Dibawah ini adalah top 50 daftar kata atau frasa pada *email* yang biasanya mengindikasikan *email* tersebut merupakan *email spam*:

1. !!!
2. $$$
3. 100% free
4. Act now!
5. ALL CAPITALS
6. All natural
7. As seen on
8. Attention
9. Bad credit
10. Bargain
11. Best price
12. Billion
13. Certified
14. Cost
15. Dear friend
16. Decision
17. Discount
18. Double your income
19. E.x.t.r.a. Punctuation
20. Eliminate debt
21. Extra income
22. Fast cash
23. Fees
24. Financial freedom
25. FREE
26. Guarantee
27. Hot
28. Increase
29. Join millions
30. Lose weight
31. Lowest price
32. Make money fast
33. Marketing
34. Million dollars
35. Money
36. Money making
37. No medical exams
38. No purchase necessary
39. Online pharmacy
40. Opportunity
41. Partners
42. Performance
43. Rates
44. Satisfaction guaranteed
45. Search engine listings
46. Selling
47. Success
48. T e x t w i t h g a p s
49. Trial
50. Visit our website

Sumber: http://www.leadformix.com/

Saat *email* masuk ke sebuah akun, *email* tersebut akan melewati serangkaian proses, salah satunya penyaringan *spam*. Dalam prosesnya, penyaringan email ini mengaplikasikan ilmu pencocokan string atau *string matching*.

Berikut adalah contoh penyaringan *email* dengan pencocokan string menggunakan metode *Bruteforce* atau *naïve string matching*:

Teks: “Get a 50% discount”

Pattern: “discount”

```
G e t   a   5 0 %   d i s c o u n t
d i s c o u n t
  d i s c o u n t
    d i s c o u n t
      d i s c o u n t
        d i s c o u n t
          d i s c o u n t
            d i s c o u n t
              d i s c o u n t
                d i s c o u n t
                  d i s c o u n t
```

Banyak pembandingan terjadi = 18

Dengan menggunakan metode Knuth Morris Pratt, pencarian dapat diilustrasikan sebagai berikut:

```
G e t   a   5 0 %   d i s c o u n t
d i s c o u n t
  d i s c o u n t
    d i s c o u n t
      d i s c o u n t
        d i s c o u n t
          d i s c o u n t
            d i s c o u n t
              d i s c o u n t
                d i s c o u n t
                  d i s c o u n t
```

Banyak pembandingan terjadi = 18

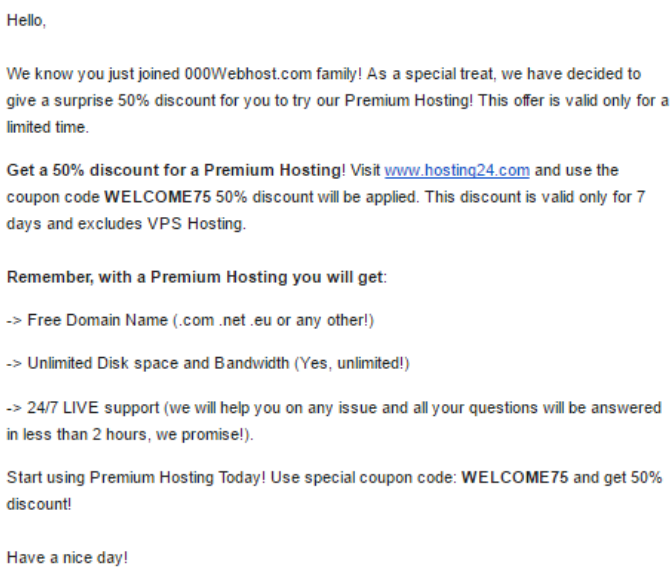
Jika dilakukan pencarian dengan menggunakan metode Booyer Moore, pencocokan string yang terjadi adalah sebagai berikut:

```
G e t   a   5 0 %   d i s c o u n t
d i s c o u n t
                d i s c o u n t
                  d i s c o u n t
```

Banyak pembandingan terjadi = 10

Pencocokan string dapat terjadi lebih dari sekali, tergantung jumlah *pattern* (kata atau frasa) yang akan ditemukan. Pada metode *keyword blacklist filtering*, jika terdapat satu *pattern* yang bersesuaian pada *email*, *email* tersebut langsung diklasifikasikan sebagai *spam*. Lain halnya dengan metode *Naïve Bayes Classifier*, pencarian dilakukan berkali-kali meskipun terdapat sebuah *pattern* yang terindikasi pada *email*. Hal ini disebabkan karena luaran dari kedua metode tersebut berbeda. Metode *keyword blacklist filtering* menghasilkan nilai *boolean* yang menyatakan sebuah email positif atau negatif *spam* sedangkan *Naïve Bayes Calssifier* menghasilkan sebuah nilai probabilitas yang menyatakan kemungkinan email tersebut spam (atau sebaliknya).

Contoh email yang diklasifikasikan sebagai spam:



Hello,

We know you just joined 000Webhost.com family! As a special treat, we have decided to give a surprise 50% discount for you to try our Premium Hosting! This offer is valid only for a limited time.

Get a 50% discount for a Premium Hosting! Visit www.hosting24.com and use the coupon code **WELCOME75** 50% discount will be applied. This discount is valid only for 7 days and excludes VPS Hosting.

Remember, with a Premium Hosting you will get:

-> Free Domain Name (.com .net .eu or any other!)

-> Unlimited Disk space and Bandwidth (Yes, unlimited!)

-> 24/7 LIVE support (we will help you on any issue and all your questions will be answered in less than 2 hours, we promise!).

Start using Premium Hosting Today! Use special coupon code: **WELCOME75** and get 50% discount!

Have a nice day!

Gambar 5. Contoh email spam

IV. KESIMPULAN

Kesimpulan yang didapat dari makalah ini adalah sebagai berikut:

1. Pencocokan string dengan metode *Booyer Moore* dalam penyaringan email spam lebih efektif digunakan daripada metode Knuth Morris Pratt. Hal ini disebabkan karena *body email* pada umumnya terdiri atas string yang berukuran besar (A..Z, a..z, 0..9, ditambah karakter lainnya).
2. Keberadaan setiap ilmu itu saling mendukung satu sama lain.

UCAPAN TERIMAKASIH

Puji syukur penulis ucapkan kepada Allah S.W.T atas segala nikmat yang telah diberikan baik berupa nikmat iman, kesehatan maupun kekuatan dalam menyusun makalah ini. Penulis juga mengucapkan terimakasih kepada kedua orang tua yang berada di kampung halaman yang telah mendidik dan membesarkan penulis dengan penuh kasih sayang. Terimakasih kepada kedua saudara penulis yang selalu memberikan motivasi untuk meraih impian sehingga penulis dapat melanjutkan pendidikan di kampus ini. Selanjutnya penulis ingin menyampaikan terimakasih kepada dosen Strategi Algoritma, Ibu Dr. Nur Ulfa Maulidevi, ST., M.Sc dan Bapak Dr.Ir. Rinaldi Munir, MT., yang telah mencurahkan banyak ilmu kepada kami "warga labtek V". Semoga ilmu yang beliau berikan dapat kami pergunakan dengan semestinya. Semoga Allah membalasi kebaikan orangorang yang namanya disebut diatas dengan kebaikan yang berlipat ganda, amin.

REFERENCES

[1] Munir, Rinaldi. "Diktat Strategi Algoritma". Program Studi Teknik Informatika, 2013.

[2] Davison, Andrew. "*Pattern Matcing*". Pennstate College. 2007

[3] https://www.idomaths.com/id/peluang5.php
Diakses pada 8 Mei 2106

[4] Diwasasri Ratnaningtyas, Dyah. "Aplikasi Teorema Bayes dalam Penyaringan Email" . Program Studi Sistim Dan Teknologi Informasi, 2010

PERNYATAAN

Dengan ini saya menyatakan bahwa makalah yang saya tulis ini adalah tulisan saya sendiri, bukan saduran, atau terjemahan dari makalah orang lain, dan bukan plagiasi.

Bandung, 8 Mei 2016

Amal Qurany

13514078