**JURUSAN TEKNIK INFORMATIKA**

**FAKULTAS TEKNOLOGI INFORMASI**

**INSTITUT TEKNLOGI SEPULUH NOPEMBER**

**USULAN TUGAS AKHIR**

# IDENTITAS PENGUSUL

**NAMA : HUSNUN NASHIR**

**NRP : 5108100181**

**DOSEN WALI : SARWOSRI, S.Kom,MT**

# JUDUL TUGAS AKHIR

“Perancangan dan Pembuatan Perangkat Lunak Pengelompokan Email Secara Otomatis memakai Klasifikasi Single Pass Clustering Berbasis Kerangka Kerja Play!”

# LATAR BELAKANG

Perkembangan internet saat ini mengalami kemajuan yang sangat pesat. Seiring dengan itu, muncullah layanan-layanan yang mendukung kegiatan manusia melalui media internet. Salah satu layanan yang populer sejak ditemukannya internet adalah electronic mail (e-mail). E-mail awalnya didefinisikan sebagai surat berbentuk file text yang dikirimkan melalui internet. E-mail sekarang ini sudah berkembang lebih atraktif dengan adanya teknologi HTML e-mail, sehingga e-mail tidak hanya berupa tulisan, namun dapat disisipi gambar maupun file-file lainnya. Dengan demikian, perkembangan e-mail yang awalnya bertujuan untuk saling berkomunikasi menjadi lebih luas karena kemampuannya dapat digunakan sebagai sarana informasi.

Dengan adanya perkembangan teknologi yang semakin pesat dan perkembangan telekomunikasi tersebut sangat memudahkan seseorang berkirim surat melalui e-mail sebab penggunaan e-mail tersebut dianggap murah dan cepat. Penggunaan e-mail juga sangat berperan dalam berbagai kegiatan pendidikan, bisnis, perdagangan, sosial dan berbagai kegiatan lainnya.

Menurut data statistik yang dibuat oleh The Radicati Group Inc.[4] pada 2010 bahwa jumlah akun email diseluruh dunia sekitar 2,9 bilyun pada 2010 dan diperkirakan akan meningkat menjadi 3,8 bilyun pada 2014. Pada tahun 2010, akun email suatu perusahaan mengirim dan menerima sekitar 110 pesan sehari. Sekitar 18% email yang diterima adalah spam ataupun graymail (newsletter yang tidak dinginkan, peringatan, dll).

Statistik diatas menunjukkan bahwa pengguna email sangat banyak. Intensitas penggunaannya pun sangat tinggi. Oleh karena itu, dibutuhkan suatu sistem pengelolaan email yang handal dan dapat mempermudah pengguna email untuk membaca email-email yang diterima. Dalam tugas akhir ini, penulis akan membuat sistem pengelompokkan email berdasarkan kesamaan isinya. Email-email yang masuk akan dikelompokkan menurut kesamaan isinya antara email satu dengan yang lainnya. Dengan ini diharapkan pengguna dapat dengan mudah menentukan mana email yang penting untuk dibaca dan yang tidak.

# RUMUSAN MASALAH

Detil permasalahan yang diangkat dalam tugas akhir ini adalah sebagai berikut:

* 1. Bagaimana membangun sistem aplikasi email *client*.
  2. Bagaimana implementasi pemrosesan awal (*pre-processing*) email.
  3. Bagaimana implementasi metode pengelompokkan (*clustering*) email.
  4. Bagaimana membangun aplikasi menggunakan framework Play!.

# BATASAN MASALAH

Masalah yang dibahas pada tugas akhir ini dibatasi lingkupnya pada:

1. Sistem aplikasi email yang dibangun adalah sistem aplikasi email *client*.
2. Metode pengelompokkan digunakan untuk isi email berbahasa Indonesia.

# TUJUAN PEMBUATAN TUGAS AKHIR

Tujuan dari penyusunan tugas akhir ini adalah:

1. Merancang bangun sistem aplikasi email client.
2. Mengimplementasikan metode pengelompokan email secara otomatis berdasarkan kesamaan isinya.

# MANFAAT TUGAS AKHIR

Manfaat dari tugas akhir ini adalah sebagai media bantu untuk memudahkan pengelolaan email karena sudah dikelompokkan berdasarkan kemiripannya.

# TINJAUAN PUSTAKA

## Rancang bangun perangkat lunak

Rancang bangun perangkat lunak merupakan tahap-tahap teknis untuk membangun perangkat lunak yang melingkupi perencanaan, analisis sistem, implementasi, serta aktivitas pengujian dan pemeliharaan perangkat lunak.

Rancang bangun perangkat lunak diperlukan untuk menentukan konsep, strategi, dan praktik yang baik diterapkan untuk menciptakan perangkat lunak yang berkualitas tinggi, sesuai anggaran biaya, mudah dalam pemeliharaannya, serta tidak membutuhkan waktu yang lama dalam pembangunannya [2]. Beberapa model rancang bangun perangkat lunak yang terkenal dan banyak dipakai antara lain model air terjun serta model iterasi.

## Email

Kata email berasal dari electronic mail yang artinya adalah surat elektronik. Email digunakan untuk pertukaran pesan digital antara satu orang dengan orang yang lain. Sistem email memiliki model store-and-forward. Server email menerima, meneruskan, mengirim, dan menyimpan pesan. Email dapat dianalogikan dengan kotak surat yang ada di kantor pos sedangkan server email dapat diibaratkan sebagai kantor pos. Dengan analogi ini sebuah mail server dapat memiliki banyak account email yang ada didalamnya.

Sebuah pesan email memiliki dua komponen yaitu header dan body. Header berisi field-field tentang informasi / identitas email. Sedangkan body adalah isi email / pesan yang akan disampaikan.

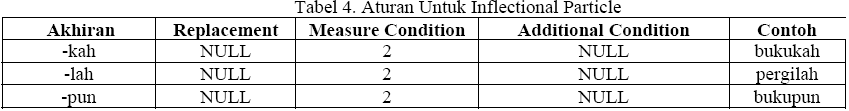
Untuk mengirim sebuah email dari alamat email yang satu ke alamat email yang lain digunakan sebauh protocol (aturan) yaitu Simple Mail Transfer Protocol SMTP. Protocol SMTP telah menjadi aturan dasar yang disepakati untuk pengiriman email. Dengan demikian semua software email server pasti mendukung protokol ini. SMTP merupakan protokol yang digunakan untuk megirim email (komunikasi antar mailserver), dan tidak digunakan untuk berkomunikasi dengan client. Sedangkan untuk client, digunakan protokol imap, imaps, pop3, atau pop3s.

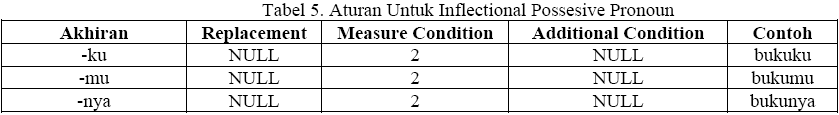
## Text Mining

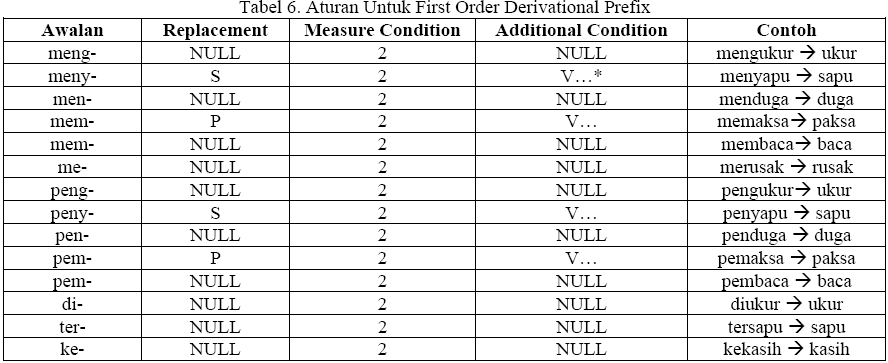
Text Mining adalah semua jenis aktivitas pengolahan teks yang bertujuan untuk mencari, mengorganisir, dan menganalisa informasi. Sedangkan pengertian formal text mining adalah penciptaan informasi baru yang belum dimengerti dalam suatu koleksi dokumen. Informasi baru tersebut didefinisikan sebagai tren, pola, atau hubungan yang tidak dapat dengan mudah dikumpulkan dengan hanya membaca dokumen secara individu. Dokumen yang dianalisis dapat berupa unit teks apapun, seperti halaman web, e-mail, sebuah artikel terformat, slide, atau file teks biasa [1].

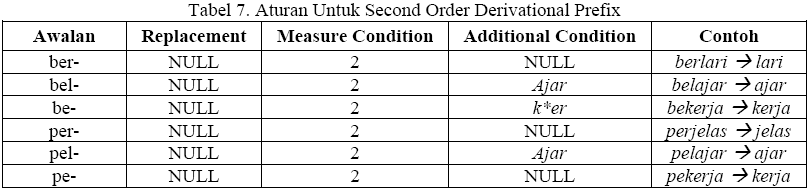
## Stemmer Bahasa Indonesia

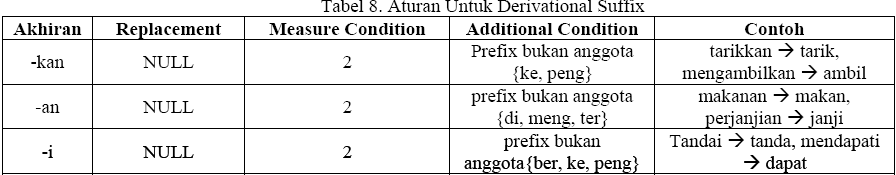
Stemming adalah suatu proses pengembalian suatu kata berimbuhan ke bentuk dasarnya (stem/root). Sebagai contoh, stemming pada kata “mempermainkan” akan menghasilkan kata “main”. Stemming adalah alat pemrosesan teks dasar yang sering digunakan untuk mendapatkan kinerja yang efektif dan efisien pada text retrieval dan text classification. Tabel aturan pembentukan kata dasar tersebut dapat dilihat pada Tabel 4 sampai Tabel 8.











## Single Pass Clustering

Single Pass Clustering[3] merupakan suatu tipe clustering yang berusaha melakukan pengelompokan data satu demi satu dan pembentukan kelompok dilakukan seiring dengan pengevaluasian setiap data yang dimasukkan ke dalam proses cluster. Pengevaluasian tingkat kesamaan antar data dan cluster dilakukan dengan berbagai macam cara termasuk menggunakan fungsi jarak, vectors similarity, dan lain-lain. Algoritma umum Single Pass Clustering adalah sebagai berikut :

1. Masukkan (dokumen pertama) D1 representasi (Cluster pertama) C1
2. Untuk (dokumen ke-i) Di hitung kesamaan (similarity) dengan setiap wakil dari masing-masing cluster.
3. Jika (Maximum Similarity) Smax lebih besar dari batas nilai (threshold value) ST, tambahkan tambahkan item kepada cluster yang bersesuaian dan hitung kembali representasi cluster, sebaliknya gunakan Di untuk inisialisasi cluster baru.
4. Jika masih ada sebuah item Di yang belum dikelompokkan, kembali ke langkah ke-2



## Play! Framework

Play! Framework[7] adalah framework open source untuk pengembangan aplikasi web yang berbasis Java. Play! Framework dibuat oleh pengembang software bernama Guillaume Bort. Play! Framework dirilis pertama kali pada tahun 2007. Versi stabil terbarunya bernomor versi 1.2.4 yang dirilis pada 2 Desember 2011 sedangkan versi 2.0 masih Beta. Beberapa keunggulan Play! Framework dibandingkan dengan framework lain diantaranya :

1. Framework Play! secara jelas menerapkan arsitektur model, view, controller (MVC).
2. *Fix the bug and hit reload*. Edit file java, simpan, refresh browser dan dapatkan hasilnya secepatnya. Tanpa perlu compile, deploy atau restart server.
3. *Stateless model*. Play adalah 'Share nothing' sistem. Dapat dengan mudah menjalankan beberapa aplikasi yang sama pada beberapa server yang berbeda.
4. *Resolve errors quickly*. Saat error muncul, Play menampilkan source code dan baris code yang bermasalah. Meskipun code tersebut berada di template.
5. Menyediakan integrasi dengan Hibernate, OpenID, Memchached.
6. *Pure Java*. Code dengan Java, library yang digunakan semua dengan Java. Dapat dengan mudah diintegrasikan dengan eclipse dan netbeans.
7. *Really fast*. Dimulai dan dijalankan dengan cepat.
8. View yang digunakan sangat fleksibel. jQuery ataupun *stylesheet* CSS dapat digunakan pada Play

# RINGKASAN ISI TUGAS AKHIR

Dalam tugas akhir ini akan dibuat sebuah Sistem Pengelompokan Email secara Otomatis memakai Klasifikasi Single Pass Clustering Berbasis Kerangka Kerja Play!. Deskripsi sistem adalah sebagai berikut:

* Sistem dapat mengunduh email dari email server dan menyimpannya sebagai file teks.
* Sistem melakukan pemrosesan awal terhadap file tersebut (menghilangkan kata hubung, melakukan stemming, menghitung frekuensi kata, dll.).
* Sistem melakukan perhitungan tingkat kemiripan (similarity) tiap file.
* Sistem membentuk kelompok (cluster) sesuai dengan hasil perhitungan. Untuk proses pengelompokan dilakukan secara unsupervised.
* Sistem diuji coba dengan beberapa kondisi tertentu untuk melihat efektifitas proses pengelompokan.
  1. Arsitektur Sistem



Penjelasan sistem :

Aplikasi secara terus menerus akan melakukan listening terhadap server email. Jika ada email yang baru maka email tersebut akan langsung diunduh ke dalam filesystem lokal dimana aplikasi itu berada menggunakan protokol IMAP. Kemudian aplikasi akan mengambil data email dari filesystem lokal nya tersebut untuk diolah. Pengolahan data email ini nantinya menghasilkan output pengelompokan/clustering yang ditampilkan pada antarmuka aplikasi. Alur proses ini akan dilakukan secara terus-menerus selama ada koneksi jaringan/internet.

* 1. Algoritma/Prosedur Pengelompokan Email



1. Penyaringan Term

Penyaringan term maksudnya adalah memproses file email mentah sehingga menjadi daftar term yang sesuai dengan kamus. Di dalam proses ini juga dilakukan proses *tokenizing* dan *stemming*.

1. Menghitung frekuensi term

Menghitung jumlah masing-masing term pada satu file

1. Indexing

Indexing adalah proses pemberian bobot masing-masing term. Perhitungan indexing menggunakan TF/IDF. Dengan rumus sebagai berikut :

Bobot (i) = frekuensi term \* (1 + log(N) - log(di))

Dimana :

Bobot = nilai IDF term i.

N = jumlah file keseluruhan

di = jumlah file yang mengandung term i

1. Perhitungan similarity

Adalah proses perhitungan similarity antara satu file dengan file lainnya. Perhitungan similarity ini menggunakan metode Jaccard Similarity dengan rumus sebagai berikut :

similarity(A,B) = n(A ∩ B) / n(A ∪ B)

dimana

similarity(A,B) = nilai similarity antara A dan B

n(A ∩ B) = Σ min(Ai, Bi)

n(A ∪ B) = Σ max(Ai, Bi)

i = term i

1. Pengelompokan email

Prosedur pengelompokan email dengan metode Single Pass Clustering :

1. Membaca email yang masuk
2. Membaca cluster yang tersedia
3. Menghitung similarity antara email masuk dengan vektor term yang merepresentasikan masing-masing cluster.
4. Ambil cluster dengan nilai rata-rata similarity terbesar
5. Jika nilai rata-rata similarity lebih besar dari nilai batas maka masukkan email masuk ke dalam cluster tersebut.
6. Sebaliknya, buat cluster baru dan masukkan email masuk ke dalam cluster baru tersebut.
7. Memperbarui cluster

# METODOLOGI

## Penyusunan proposal tugas akhir

Pada tahap ini penulis menyusun proposal tugas akhir sebagai langkah awal dalam pengerjaan tugas akhir. Pada proposal ini penulis menggagas penulisan tugas akhir untuk merancang bangun sistem pengelompokan email secara otomatis.

## Studi literatur

Pada tahap ini dilakukan pengumpulan informasi yang diperlukan untuk perancangan sistem. Informasi tersebut dapat diperoleh dari literatur, paper, maupun buku-buku terkait dengan perancangan maupun pembangunan perangkat lunak.

## Analisis dan desain perangkat lunak

Analisis kebutuhan dan perancangan sistem dilakukan untuk merumuskan solusi yang tepat dalam pembuatan aplikasi serta kemungkinan yang dapat dilakukan untuk mengimplementasikan rancangan tersebut. Fase desain meliputi arsitektur perangkat lunak yang digunakan, desain class-class yang terlibat dalam aplikasi, desain basis data, dan lain-lain.

## Implementasi perangkat lunak

Pada fase ini desain perangkat lunak diwujudkan ke dalam bentuk kode program. Pembangunan aplikasi dilakukan menggunakan platform Java. Pada fase ini, penulis menggunakan *Netbeans 6.8* sebagai alat bantu untuk implementasi pengkodean program.

## Uji coba dan evaluasi

Pada tahap ini dilakukan uji coba terhadap perangkat lunak yang dibuat. Tujuan uji coba perangkat lunak adalah untuk menemukan kesalahan-kesalahan (*bug*) sedini mungkin sehingga dapat diperbaiki sesegera mungkin.

## Penyusunan Buku Tugas Akhir

Pada tahap ini dilakukan penyusunan laporan yang menjelaskan dasar teori dan metode yang digunakan dalam tugas akhir ini serta hasil dari implementasi aplikasi perangkat lunak yang telah dibuat. Sistematika penulisan buku tugas akhir secara garis besar antara lain:

1. Pendahuluan
   1. Latar Belakang
   2. Rumusan Masalah
   3. Batasan Tugas Akhir
   4. Tujuan
   5. Metodologi
   6. Sistematika Penulisan
2. Tinjauan Pustaka
3. Desain dan Implementasi
4. Uji Coba dan Evaluasi
5. Kesimpulan dan Saran
6. Daftar Pustaka

# JADWAL KEGIATAN

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Tahapan | 2012 | | | | | | | | | | | | | | | | | | | | |
| Februari | | | | Maret | | | | April | | | | Mei | | | | | Juni | | | |
| Penyusunan Proposal |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Studi Literatur |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Perancangan sistem |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Implementasi |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Uji coba dan evaluasi |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Penyusunan buku |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |

# DAFTAR PUSTAKA

1. **Feldman, R., and Sanger, J.** *The Text Mining Handbook*. New York: Cambridge University Press, 2006. 9780521836579
2. **Sommerville, Ian.** *Software Engineering.* Boston : Addison-Wesley, 2007. 9780321313799.
3. **B. Frakes, William dan Baeza-Yates, Richardo.** *Information Retrieval Data Structures and Algorithm.* Prentice Hall International Edition, 2007.
4. **Radicati, Sara.** *Email Statistics Report,2010.* The Radicati Group, Inc., April 2004. [Dikutip: 24 Januari 2012.] http://www.radicati.com.
5. **Zainal Arifin, Agus dan Novan Setiono, Ari.** *Klasifikasi Dokumen Berita Kejadian Berbahasa Indonesia dengan Algoritma Single Pass Clustering.* Jurusan Teknik Informatika, Institut Teknologi Sepuluh Nopember.
6. **Z. Tala, Fadillah.** *A Study of Stemming Effects on Information Retrieval in Bahasa Indonesia.* Institute for Logic, Languange and Computation, Universiteit van Amsterdam The Netherlands.
7. **Bort, Guillaume.** *Play! Framework.* http://www.playframework.org/.