IMPLEMENTASI ALGORITMA *SORTED NEIGHBORHOOD METHOD* DAN *N-GRAM* DALAM APLIKASI DETEKSI DUPLIKASI DATA (STUDI KASUS: RUMAH SAKIT SWASTA DI JAKARTA)

TUGAS AKHIR

AMELIA FAHMI

1132001008

PROGRAM STUDI INFORMATIKA

FAKULTAS TEKNIK DAN ILMU KOMPUTER

UNIVERSITAS BAKRIE

JAKARTA

2019

IMPLEMENTASI ALGORITMA *SORTED NEIGHBORHOOD METHOD* DAN *N-GRAM* DALAM APLIKASI DETEKSI DUPLIKASI DATA (STUDI KASUS: RUMAH SAKIT SWASTA DI JAKARTA)

TUGAS AKHIR

Diajukan sebagai salah satu syarat untuk memperoleh gelar Sarjana Komputer

AMELIA FAHMI

1132001008

PROGRAM STUDI INFORMATIKA

FAKULTAS TEKNIK DAN ILMU KOMPUTER

UNIVERSITAS BAKRIE

JAKARTA

2019

HALAMAN PERNYATAAN ORISINALITAS

Tugas akhir ini adalah hasil karya saya sendiri dan semua sumber baik yang dikutip maupun dirujuk telah saya nyatakan dengan benar.

Nama : **Amelia Fahmi**

Nim : **1132001008**

Tanda Tangan:

Tanggal : **15 April 2019**

HALAMAN PENGESAHAN

Tugas akhir ini diajukan oleh:

Nama	: Amelia Fahmi
Nim	: 1132001008
Program studi	: Informatika
Fakultas	: Teknik dan Ilmu Komputer
Judul Tugas Akhir	: Implementasi Algoritma *Sorted Neighborhood Method* dan *N-Gram* dalam aplikasi deteksi duplikasi data (Studi Kasus: Rumah Sakit Swasta di Jakarta)

Telah berhasil dipersidangkan dihadapan Dosen Penguji dan diterima sebagai bagian persyaratan yang diperlukan untuk memperoleh gelar Sarjana Komputer pada Program Studi Informatika Fakultas Teknik dan Ilmu Komputer, Universitas Bakrie.

Dewan Penguji

Pembimbing	: Yusuf Lestanto, S.T., M., Sc.	(.....)
Penguji	: Guson P. Kuntarto, S.T., M., Sc.	(.....)
Penguji	: Prof. Dr. Hoga Saragih, S.T., M.T.	(.....)

Ditetapkan di : Jakarta

Tanggal : 15 April 2019

UCAPAN TERIMAKASIH

Alhamdulillahirabil'alamin puji dan syukur ke hadirat Allah SWT, karena atas rahmat-Nya dan karunia-Nya lah Tugas Akhir ini dapat diselesaikan. Tugas Akhir ini berjudul "Implementasi Algoritma *Sorted Neighborhood Method* dan *N-Gram* dalam aplikasi deteksi duplikasi data (Studi Kasus: Rumah Sakit Swasta di Jakarta)".

Terselesaikannya Tugas Akhir ini tidak luput dari bantuan banyak pihak yang telah membantu dalam penelitian dan penulisan. Oleh karena itu, dengan segala kerendahan hati ucapan terimakasih ini disampaikan kepada:

1. Yusuf Lestanto, S.T., M,Sc., selaku dosen pembimbing, yang telah meluangkan waktunya serta memberikan, motivasi, bimbingan, saran serta perbaikan dalam menyelesaikan Tugas Akhir ini.
2. Kedua Orang tua tercinta, Bapak Fahmi Rizal dan Ibu Nurlaila, kedua saudara/saudari kandung, Adelia dan Ibnu Haykal. Terimakasih telah memberikan dukungan, semangat dan doa yang sangat berarti.
3. Prof. Dr. Hoga Saragih, S.T., M.T., selaku Kepala Program Studi Informatika dan dosen penguji Tugas Akhir, yang senantiasa memberikan motivasi, saran serta perbaikan Tugas Akhir.
4. Guson P. Kuntarto, S.T., M., Sc., selaku dosen penguji Tugas Akhir yang senantiasa memberikan saran serta perbaikan Tugas Akhir.
5. Seluruh Bapak/Ibu dosen Program Studi Informatika Universita Bakrie, yang telah memberikan banyak ilmu, pengetahuan, wawasan selama perkuliahan.
6. Sahabat-sahabat, Denok, Sika, Junita, Intan, Bariqi, Noddie yang senantiasa memberikan motivasi, dukungan, sukacita dan kebersamaan selama ini.
7. Sahabat-sahabat mahasiswa Universitas Bakrie, Dina, Nida, Fu, Riri, Kiki, Fildzah, Hani, Rica dan Ghifari yang senantiasa memberikan

motivasi, dukungan selama pengerjaan Tugas Akhir dan kebersamaan selama masa perkuliahan.

8. Teman-teman seperjuangan Informatika 2013. Terimakasih telah memberikan semangat dan bekerja sama selama empat tahun masa perkuliahan.
9. Adik-adik Informatika 2014, 2015 dan 2016. Terimakasih telah memberikan dukungan dan semangat.
10. Seluruh pihak yang terlibat dalam penyusunan Tugas Akhir ini yang tidak dapat disebutkan satu persatu.

Dalam penyusunan Tugas Akhir ini tentunya masih terdapat banyak kekurangan. Untuk itu, saran dan kritik yang membangun dari pembaca selalu diterima. Akhirnya, semoga Tugas Akhir ini dapat bermanfaat bagi kita semua.

Jakarta, 15 April 2019

Amelia Fahmi

HALAMAN PERNYATAAN PERSETUJUAN PUBLIKASI

Sebagai *civitas* akademik Universitas Bakrie, saya yang bertanda tangan di bawah ini:

Nama	: Amelia Fahmi
Nim	: 1132001008
Program studi	: Informatika
Fakultas	: Teknik dan Ilmu Komputer
Jenis Tugas Akhir	: Implementasi

Demi pengembangan ilmu pengetahuan, menyetujui untuk memberikan kepada Universitas Bakrie **Hak Bebas Royalti Noneksklusif (*Non-exclusive Royalty-Free Right*)** atas tugas akhir saya yang berjudul:

Implementasi Algoritma *Sorted Neighborhood Method* dan *N-Gram* dalam aplikasi deteksi duplikasi data (Studi Kasus: Rumah Sakit Swasta di Jakarta)

Beserta perangkat yang ada (jika diperlukan). Dengan Hak Bebas Royalti Noneksklusif ini, Universitas Bakrie berhak menyimpan, mengalihmedia/formatkan, mengelola dalam bentuk pangkalan data (*database*), merawat dan mempublikasikan tugas akhir saya selama tetap menyantumkan nama saya sebagai penulis/pencipta dan sebagai pemilik Hak Cipta untuk kepentingan akademis.

Demikian pernyataan ini saya buat dengan sebenarnya.

Dibuat di : Jakarta

Tanggal : 15 April 2019

Yang menyatakan,

Amelia Fahmi

IMPLEMENTASI ALGORITMA *SORTED NEIGHBORHOOD METHOD* DAN *N-GRAM* DALAM APLIKASI DETEKSI DUPLIKASI DATA (STUDI KASUS: RUMAH SAKIT SWASTA DI JAKARTA)

Amelia Fahmi

ABSTRAK

Penelitian ini bertujuan untuk mengimplementasikan algoritma *Sorted Neighborhood Method* dan *N-Gram* dalam mendeteksi duplikasi pada data pasien Rumah Sakit Swasta di Jakarta. Dalam beberapa aplikasi untuk menjamin kualitas data dan kinerja aplikasi, data harus dibersihkan. Salah satu hal terpenting dalam pembersihan data adalah deteksi duplikasi data. Deteksi duplikasi data merupakan proses identifikasi pada *record* yang berbeda tetapi memiliki kemiripan yang tinggi. Oleh karena itu aplikasi yang dibangun dalam penelitian ini adalah aplikasi deteksi duplikasi data berbasis *web* dan dirancang menggunakan PHP. Aplikasi ini bertujuan untuk menghindari kesalahpahaman yang disebabkan kesamaan data pasien Rumah Sakit Swasta di Jakarta pada tingkat operasional, serta memberikan kemudahan bagi tenaga medis ataupun pihak terkait dalam mendapatkan informasi mengenai riwayat penyakit pasien. Dalam penelitian ini untuk mengetahui nilai efektivitas algoritma *Sorted Neighborhood Method* dan *N-Gram* terhadap aplikasi yang dibangun, dilakukan pengujian menggunakan hasil perhitungan sampel data pasien untuk menghitung *precision, recall* dan *f-measure*. Hasil pengujian metode tersebut dapat dilihat pada grafik hasil pengujian yang menunjukan bahwa nilai *f-measure* yang lebih *optimal* pada sampel data pasien terdapat pada perpaduan nilai token 4 gram 4 dan threshold 0.5 dengan nilai *f-measure* sebesar 0.76.

Kata kunci: Deteksi Duplikasi, *Sorted Neighborhood Method, N-Gram*

IMPLEMENTASI ALGORITMA *SORTED NEIGHBORHOOD METHOD* DAN *N-GRAM* DALAM APLIKASI DETEKSI DUPLIKASI DATA (STUDI KASUS: RUMAH SAKIT SWASTA DI JAKARTA)

Amelia Fahmi

ABSTRACT

This research aims to implement Sorted Neighborhood Method and N-Gram Algorithm in detecting duplication in patient data of Hospital in Jakarta. In some applications to ensure data quality and application performance, data must be cleaned. One of the most important things in data cleaning is data duplication detection. Data duplication detection is a process of identifying different records but having a high similarity. Therefore the application that builds in this research is a web-based data duplication detection application and designed using PHP. This application aims to avoid misunderstanding caused by the similarity of patients data Hospital in Jakarta at the oprational level, as well as provide convenience for medical personnel or related parties in obtaining information about the patient disease history. In this study to know the value of the effectiveness of the Sorted Neighborhood Method and N-Gram algorithm towards the applications that is built, testing is done using the results of the calculation of patient data samples to calculate precision, recall dan f-measure. The result of the evaluation of the method can be seen in the graph of test result that show the value of the f-measure that is more optimal of the sample patient data obtained in a combination of value of token 4 gram 4 and threshold 0.5 with the f-measure value of 0.76.

Keywords: *Duplicate Detection, Sorted Neighborhood Method, N-Gram*

DAFTAR ISI

DAFTAR GAMBAR

DAFTAR TABEL

DAFTAR RUMUS

DAFTAR LAMPIRAN

DAFTAR SINGKATAN

SNM : *Sorted Neighborhood Method*

SRS : *Software Requirement Specification*