SISTEM IDENTIFIKASI BAHASA JAWA DAN BAHASA INDONESIA DOKUMEN TEKS BERBASIS KARAKTER N-GRAM

Skripsi

oleh
FIDELIA VERA SENTOSA
71140005

PROGRAM STUDI INFORMATIKA FAKULTAS TEKNOLOGI INFORMASI
UNIVERSITAS KRISTEN DUTA WACANA
2018

SISTEM IDENTIFIKASI BAHASA JAWA DAN BAHASA INDONESIA DOKUMEN TEKS BERBASIS KARAKTER N-GRAM

Skripsi

Diajukan kepada Program Studi Informatika Fakultas Teknologi Informasi
Universitas Kristen Duta Wacana
Sebagai Salah Satu Syarat dalam Memperoleh Gelar
Sarjana Komputer

Disusun oleh

FIDELIA VERA SENTOSA
71140005

PROGRAM STUDI INFORMATIKA FAKULTAS TEKNOLOGI INFORMASI
UNIVERSITAS KRISTEN DUTA WACANA
2018

PERNYATAAN KEASLIAN SKRIPSI

Saya menyatakan dengan sesungguhnya bahwa skripsi dengan judul:

SISTEM IDENTIFIKASI BAHASA JAWA DAN BAHASA INDONESIA DOKUMEN TEKS BERBASIS KARAKTER N-GRAM

yang saya kerjakan untuk melengkapi sebagian persyaratan menjadi Sarjana Komputer pada pendidikan Sarjana Program Studi Informatika Fakultas Teknologi Informasi Universitas Kristen Duta Wacana, bukan merupakan tiruan atau duplikasi dari skripsi kesarjanaan di lingkungan Universitas Kristen Duta Wacana maupun di Perguruan Tinggi atau instansi manapun, kecuali bagian yang sumber informasinya dicantumkan sebagaimana mestinya.

Jika dikemudian hari didapati bahwa hasil skripsi ini adalah hasil plagiasi atau tiruan dari skripsi lain, saya bersedia dikenai sanksi yakni pencabutan gelar kesarjanaan saya.

Yogyakarta, 7 Januari 2019

FIDELIA VERA SENTOSA
71140005

HALAMAN PERSETUJUAN

Judul Skripsi	:	SISTEM IDENTIFIKASI BAHASA JAWA DAN BAHASA INDONESIA DOKUMEN TEKS BERBASIS KARAKTER N-GRAM
Nama Mahasiswa	:	FIDELIA VERA SENTOSA
N I M	:	71140005
Matakuliah	:	Skripsi (Tugas Akhir)
Kode	:	TIW276
Semester	:	Gasal
Tahun Akademik	:	2018/2019

2018/2019

Telah diperiksa dan disetujui di
Yogyakarta,
Pada tanggal 30 November 2018

Dosen Pembimbing I	Dosen Pembimbing II
Lucia Dwi Krisnawati, Dr. Phil.	Aditya Wikan Mahastama, S.Kom., M.Cs.

HALAMAN PENGESAHAN

SISTEM IDENTIFIKASI BAHASA JAWA DAN BAHASA INDONESIA DOKUMEN TEKS BERBASIS KARAKTER N-GRAM

Oleh: FIDELIA VERA SENTOSA / 71140005

Dipertahankan di depan Dewan Penguji Skripsi
Program Studi Informatika Fakultas Teknologi Informasi
Universitas Kristen Duta Wacana - Yogyakarta
Dan dinyatakan diterima untuk memenuhi salah satu syarat memperoleh gelar
Sarjana Komputer
pada tanggal 12 Desember 2018

Yogyakarta, 7 Januari 2019
Mengesahkan,

Dewan Penguji:

1. Lucia Dwi Krisnawati, Dr. Phil.
2. Aditya Wikan Mahastama, S.Kom., M.Cs.
3. R. Gunawan Santosa, Drs. M.Si.
4. Maria Nila Anggia Rini, S.T, M.T.I

Dekan

(Budi Susanto, S.Kom., M.T.)

Ketua Program Studi

(Gloria Virginia, Ph.D.)

UCAPAN TERIMA KASIH

Pertama-tama Penulis mengucapkan puji syukur dan terima kasih kepada Tuhan yang Maha Kuasa atas berkat dan rahmat-Nya, penulis mampu menyelesaikan tugas akhir yang berjudul "Sistem Identifikasi Bahasa Jawa dan Bahasa Indonesia Dokumen Teks Berbasis Karakter N-gram".

Meskipun masih mendapati beberapa halangan dan hambatan selama mengerjakan tugas akhir ini, Penulis mendapatkan bantuan, dukungan dan kerjasama dari berbagai pihak yang membantu sehingga Penulis mampu menyelesaikan tugas akhir ini. Maka dari itu, Penulis ingin mengucapkan terima kasih terkhusus pada:

1. Ibunda Lani Kristanti, Nenek Ang Kas Nio, Kakak Stefanie Mayasari, Kakak Jessica Fina Susanti, Kakak Andretti Devin Kurniawan, Kakak Budiarto Suprayogo, dan semua keluarga di Solo yang selalu memberikan dukungan doa serta bantuan baik berupa material maupun moral bagi penulis untuk bisa menyelesaikan tugas akhir ini.
2. Bapak Budi Susanto, S.Kom., M.T. selaku Dekan Fakultas Teknologi Informasi Universitas Kristen Duta Wacana.
3. Ibu Gloria Virginia, Ph. D. selaku ketua program studi Informatika Universitas Kristen Duta Wacana sekaligus dosen wali penulis yang dengan sabar memberikan dukungan, arahan, dan nasihat kepada penulis.
4. Ibu Lucia Dwi Krisnawati, Dr. Phill. selaku dosen pembimbing I yang telah memberikan waktunya untuk penulis supaya dapat melakukan konsultasi mengenai tugas akhir, memberikan arahan, penjelasan, dan masukan untuk sistem identifikasi bahasa ini.
5. Bapak Aditya Wikan Mahastama, S.Kom., M.Cs. selaku dosen pembimbing II yang senantiasa memberikan arahan bagi penulis dalam menyelesaikan tugas akhir ini.

6. Erwin Winata yang telah membantu dan menemani penulis selama pengerjaan program dan penulisan tugas akhir.
7. Terimakasih kepada teman-teman: Ofri Cantika, Steishy Mega, Karlina, Wahyuni, Emylia Intan, Cyndy Alisia, Stephani Nugroho, Olivia Citra, teman-teman SWAG dan teman-teman lainnya yang telah memberikan bantuan, dukungan dan semangat kepada penulis dalam pengerjaan tugas akhir ini.
8. Terimakasih kepada McDonalds dan Kedai IQ yang telah menyediakan tempat yang nyaman untuk penulis dalam proses pengerjaan tugas akhir ini.
9. Terimakasih juga kepada seluruh pihak yang tidak dapat dituliskan satu-persatu yang selalu memberikan dukungan baik secara langsung maupun tidak langsung selama pengerjaan tugas akhir. Kiranya Tuhan senantiasa melimpahkan berkat dan rahmat-Nya bagi kita semua.

Penulis menyadari bahwa masih terdapat banyak kekurangan dalam penelitian ini, baik dalam penulisan ataupun pembahasan. Akhir kata penulis mengucapkan terima kasih kepada semua pihak yang telah berkontribusi dalam penelitian tugas akhir ini. Penulis juga berharap semoga tugas akhir ini dapat bermanfaat bagi para pembaca.

INTISARI

SISTEM IDENTIFIKASI BAHASA JAWA DAN BAHASA INDONESIA DOKUMEN TEKS BERBASIS KARAKTER N-GRAM

Dalam beberapa tahun terakhir, jumlah akan ketersediaan dokumen semakin bertambah dan beragam seiring dengan berkembangnya internet. Namun, informasi maupun data yang ada bersifat heterogen dan tidak terstruktur sehingga sulit untuk dikumpulkan secara manual. Maka, dibutuhkan suatu sistem yang dapat melakukan pengidentifikasian bahasa secara otomatis menggunakan komputer, supaya lebih efisien jika dibandingkan dengan cara manual manusia.

Klasifikasi dokumen teks merupakan permasalahan mendasar dan penting. Mengingat bahwa bahasa Indonesia merupakan *under resource langauge* sama halnya dengan bahasa Jawa, maka identifikasi bahasa sangat diperlukan. Oleh karena itu, permasalahan ini merupakan masalah yang bisa dikatakan cukup kompleks dikarenakan penggunaan kata yang tergolong tidak sedikit. Salah satu metode yang dapat digunakan untuk mengklasifikasikan naskah dokumen tersebut adalah menggunakan n-gram.

Sistem identifikasi bahasa Jawa dan bahasa Indonesia dengan karakter n-gram yang telah dikembangkan membuktikan bahwa berhasil mengidentifikasi bahasa dari sebuah naskah dokumen dengan nilai akurasi 85,07463%. Hal ini menunjukkan bahwa n-gram dapat diterapkan untuk mengidentifikasikan suatu naskah dokumen.

Kata Kunci: Identifikasi Bahasa, N-gram, Bahasa Jawa, Bahasa Indonesia.

DAFTAR ISI

DAFTAR GAMBAR

DAFTAR TABEL

DAFTAR LAMPIRAN

INTISARI

SISTEM IDENTIFIKASI BAHASA JAWA DAN BAHASA INDONESIA DOKUMEN TEKS BERBASIS KARAKTER N-GRAM

Dalam beberapa tahun terakhir, jumlah akan ketersediaan dokumen semakin bertambah dan beragam seiring dengan berkembangnya internet. Namun, informasi maupun data yang ada bersifat heterogen dan tidak terstruktur sehingga sulit untuk dikumpulkan secara manual. Maka, dibutuhkan suatu sistem yang dapat melakukan pengidentifikasian bahasa secara otomatis menggunakan komputer, supaya lebih efisien jika dibandingkan dengan cara manual manusia.

Klasifikasi dokumen teks merupakan permasalahan mendasar dan penting. Mengingat bahwa bahasa Indonesia merupakan *under resource langauge* sama halnya dengan bahasa Jawa, maka identifikasi bahasa sangat diperlukan. Oleh karena itu, permasalahan ini merupakan masalah yang bisa dikatakan cukup kompleks dikarenakan penggunaan kata yang tergolong tidak sedikit. Salah satu metode yang dapat digunakan untuk mengklasifikasikan naskah dokumen tersebut adalah menggunakan n-gram.

Sistem identifikasi bahasa Jawa dan bahasa Indonesia dengan karakter n-gram yang telah dikembangkan membuktikan bahwa berhasil mengidentifikasi bahasa dari sebuah naskah dokumen dengan nilai akurasi 85,07463%. Hal ini menunjukkan bahwa n-gram dapat diterapkan untuk mengidentifikasikan suatu naskah dokumen.

Kata Kunci: Identifikasi Bahasa, N-gram, Bahasa Jawa, Bahasa Indonesia.

BAB 1
PENDAHULUAN

1.1 Latar Belakang Masalah

Kebudayaan Indonesia sangatlah beragam, salah satu ragam kebudayaan tersebut adalah bahasa. Bahasa Jawa merupakan salah satu bahasa daerah di Indonesia dari sekian banyak bahasa daerah yang ada. Dalam beberapa tahun terakhir, jumlah akan ketersediaan dokumen semakin bertambah dan beragam seiring dengan berkembangnya internet. Internet adalah sumber informasi dan data yang paling luas yang pernah dibangun oleh umat manusia. Namun, informasi maupun data yang ada bersifat heterogen dan tidak terstruktur sehingga sulit untuk dikumpulkan secara manual dan rumit untuk digunakan dalam proses otomatis (Castrillo, 2015). Jika jumlah dokumen semakin kompleks, maka proses pencarian suatu dokumen tertentu juga semakin sulit didapat kerelevanan nya (Alifian, et al., n.d.). Mengingat perkembangan teknologi yang semakin mempermudah manusia untuk bertukar informasi yang salah satu jenis informasi tersebut adalah berupa teks, maka dibuat berbagai macam kategori untuk menyaring informasi yang relevan terhadap kebutuhan setiap orang. Salah satu dari kategori tersebut adalah berdasarkan bahasa. Maka, dibutuhkan suatu sistem aplikasi yang dapat melakukan pengidentifikasian bahasa secara otomatis menggunakan komputer, supaya lebih cepat jika dibandingkan dengan cara manual manusia.

Telah banyak upaya untuk melestarikan kebudayaan bahasa di Indonesia. Salah satu contoh nyata mengenai pelestarian kebudayaan tersebut adalah beberapa Universitas di Indonesia yang salah satunya adalah Universitas Kristen Duta Wacana telah memuat mata kuliah *Digital Humanities. Digital Humanities* adalah bidang akademik yang terkait dengan penerapan alat dan metode komputasi untuk disiplin ilmu tradisional seperti sastra, sejarah, dan filsafat. *Digital Humanities* juga dapat diartikan sebagai tindakan mendigitalisasi karya

seni, sastra, dan kemanusiaan. Dengan adanya mata kuliah *Digital Humanities* mendorong setiap individu untuk ikut aktif dalam pelestarian budaya, khususnya budaya Indonesia. Banyak teks dokumen bersejarah yang harus di lestarikan dengan digital, walaupun selama ini telah dilestarikan, namun karena pelestarian tersebut tidak menggunakan digital, maka tidak sedikit teks-teks bersejarah yang sobek atau berjamur. Hal ini penulis ketahui saat penulis berkunjung ke salah satu museum di Yogyakarta. Supaya bisa tetap terbaca maka pelestarian tersebut harus secara digital.

Klasifikasi dokumen teks merupakan permasalahan mendasar dan penting. Mengingat bahwa bahasa Indonesia merupakan *under resource langauge* juga bahasa Jawa, maka identifikasi bahasa sangat diperlukan. Oleh karena itu, permasalahan ini merupakan masalah yang bisa dikatakan cukup kompleks dikarenakan penggunaan kata yang tergolong tidak sedikit. Salah satu metode yang dapat digunakan untuk mengklasifikasikan naskah dokumen tersebut adalah menggunakan n-gram.

1.2 Rumusan Masalah

Berdasarkan uraian diatas maka masalah yang akan diteliti adalah sebagai berikut:

1. Bagaimana mengekstraksi n-gram menjadi fitur klasifikator bahasa?
2. Bagaimana menerapkan ukuran *out of space* untuk mengidentifikasi bahasa?

1.3 Batasan Masalah

Adapun batasan masalah dalam penelitian ini adalah sebagai berikut :

1. Format dokumen uji merupakan .pdf atau .docx atau .txt.
2. Bahasa yang digunakan dalam dokumen uji yaitu Bahasa Indonesia atau Bahasa Jawa (latin).
3. Teks sumber bersifat *monolingual* (Bahasa Jawa atau Bahasa Indonesia) atau *bilingual* (Bahasa Jawa dan Bahasa Indonesia).

1.4 Tujuan Penelitian

Tujuan dari diadakannya penelitian ini yaitu untuk menyediakan sistem identifikasi Bahasa Jawa dan Bahasa Indonesia pada sebuah dokumen dengan aplikasi dalam lingkup prapemrosesan untuk aplikasi yang lebih besar.

1.5 Manfaat Penelitian

Manfaat yang diperoleh setelah penelitian selesai yaitu sistem dapat digunakan untuk mengenali dokumen uji tersebut merupakan Bahasa Jawa atau Bahasa Indonesia untuk pembangunan korpus, dan menyediakan modul pra proses dalam aplikasi-aplikasi *Natural Language Processing* yang berhubungan dengan Bahasa Indonesia dan Bahasa Jawa.

1.6 Metodologi Penelitian

Dalam penyelesaian penelitian ini, terdapat beberapa metodologi penelitian, antara lain

1.6.1 *Pre-processing*

Tahapan *pre-processing* ini merupakan tahap dimana semua angka dan tanda baca yang ada pada dokumen uji ataupun dokumen latih (untuk pembentukan profil Bahasa Jawa atau Bahasa Indonesia) dihilangkan. Semua spasi yang ada pada dokumen dihilangkan dan diganti dengan karakter '_' sebagai penanda berakhirnya suatu kata.

1.6.2 *Learning*

1.6.2.1 Pengumpulan Data

Tahap pengumpulan data pada sistem ini yaitu tahap dalam pembuatan profil Bahasa Indonesia dan Bahasa Jawa. Dimana pembuatan profil Bahasa Jawa diambil dari Trawaca dan dari halaman sastra.org. Dimana trawaca merupakan korpus citra dokumen beraksara Jawa dan dokumen beraksara latin berbahasa Jawa.

1.6.2.2 Pembentukan n-gram

Pembentukan n-gram merupakan sebuah proses transformasi dari dokumen (teks) ke dalam bentuk n-gram (bi-gram, tri-gram, quad-gram, dan penta-gram).

1.6.2.3 Input ke dalam Database

Tahapan ini adalah tahap lanjutan setelah dokumen ditransformasi kedalam bentuk n-gram, tahap ini hanya diproses untuk dokumen latih (pembuatan profil) guna untuk menyimpan ranking berdasarkan frekuensi n-gram.

1.6.3 Testing

1.6.3.1 Penghitungan Jarak

Penghitungan jarak adalah tahap penghitungan selisih ranking antara dokumen uji dengan profil dari masing-masing kategori dalam dokumen latih menggunakan mekanisme *out of space.*

1.6.3.2 Perbandingan Jarak

Tahap setelah mendapatkan jarak pada penghitungan jarak sebelumnya adalah perbandingan jarak antara profil Bahasa Jawa dan profil Bahasa Indonesia, dimana jarak terkecil atau jarak minimum adalah bahasa yang teridentifikasi.

1.7 Sistematika Penulisan

Sistematika penulisan laporan tugas akhir ini dibagi ke dalam 5 bab. Kelima bab tersebut antara lain adalah Bab 1 Pendahuluan, Bab 2 Tinjauan Pustaka dan Landasan Teori, Bab 3 Perancangan Sistem, Bab 4 Implementasi dana Analisis Sistem, dan Bab 5 Kesimpulan dan Saran:

Bab 1. Pendahuluan, pada bab ini akan menguraikan tentang latar belakang masalah, rumusan masalah, batasan masalah, tujuan penelitian, manfaat penelitian, dan metode penelitian yang akan dilakukan.

Bab 2. Tinjauan Pustaka dan Landasan Teori, yang memuat beberapa tinjauan pustaka yang diperlukan serta mendukung penelitian dan pembuatan sistem, beserta uraian dari konsep- konsep atau teori-teori yang digunakan.

Bab 3. Perancangan Sistem, merupakan bab dengan penjelasan proses secara detail. Dimana bab ini akan membahas perancangan sistem yang berisi tahapan dalam perancangan, pembangunan dan pengembangan sistem, termasuk aliran data dan rancangan antarmuka masukan dan keluaran.

Bab 4. Implementasi dan Analisis Sistem, merupakan implementasi dari sistem identifikasi bahasa yang telah dirancang pada bab sebelumnya beserta analisa-analisa yang ditemukan.

Bab 5. Kesimpulan dan Saran, bab terakhir yang berisi kesimpulan atas penelitian yang telah dilakukan oleh penulis dan saran yang dapat diberikan penulis.

BAB 5
KESIMPULAN DAN SARAN

5.1 Kesimpulan

Dari hasil penelitian yang dilakukan, maka dapat dikatakan bahwa karakter n-gram dalam proses identifikasi bahasa memiliki nilai keakurasian 85,07463%. Hal ini membuktikan bahwa profil n-gram (dengan pengambilan 100 *ranking* n-gram teratas) dapat diterapkan dalam proses pengidentifikasian bahasa, maka tujuan dari penelitian telah berhasil dicapai. Untuk memperoleh data yang lebih akurat, program ini memerlukan variasi dan jumlah kata yang lebih banyak lagi untuk dokumen uji dan sampel kedua kategori tersebut. Faktor tersebut dapat mempengaruhi *output* dari sistem identifikasi ini karena jumlah kata tersebut mempengaruhi hasil ranking n-gram dokumen uji yang akan dibandingkan dengan n-gram sampel bahasa.

5.2 Saran

Dari kekurangan yang ada pada sistem yang telah disebutkan pada kesimpulan, maka saran yang dapat diberikan penulis antara lain:

- Memperbanyak dan memperluas variasi profil sampel bahasa Indonesia dan bahasa Jawa .
- Analisa sistem menggunakan metode lebih terukur, supaya dapat dibandingkan dengan penelitian sejenis yang menggunakan metode lain.

DAFTAR PUSTAKA

Baldwin, T., & Lui, M. (2010). Language Identification: The Long and the Short of the Matter. *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL*, (hal. 229-237).

Castrillo, O. (2015). Web Scraping : Applications and Tools.

Cavnar, W. B., & Trenkle, J. M. (2005). N-Gram-Based Text Categorization.

Christianto, D. A. (2018). Rule-Based Lexicon-Based Pos Tagger untuk Teks Berbahasa Jawa. Skripsi, Universitas Kristen Duta Wacana.

Garg, A. (2014, November). A Survey of Language Identification Techniques and Applications. *Journal Of Emerging Technologies In Web Intelligence*, 388-400.

Mahastama, A. W., & Krisnawati, L. D. (2017). Histogram Peak-Based Binarization for Historical Documents. *International Conference on Smart Cities, Automation & Intelligent Computing Systems (ICON-SONICS) 2017)* (hal. 93-98). Yogyakarta: IEEE.

Pratama, N. R. (2011). Pendeteksian Bahasa Pada Teks Menggunakan Metode N-gram. Skripsi, Universitas Kristen Duta Wacana.

Sarma, N., Singh, S. R., & Goswami, D. (2018). Word Level Language Identification in Assamese-Bengali-Hindi-English Code-Mixed Social Media Text. *International Conference on Asian Language Processing (IALP)*, (hal. 272-277). Bandung, Indonesia.

Selamat, A., & Akosu, N. (2014, December). Word-Length Algorithm for Language Identification of Under-Resourced Languages. *Journal of King Saud University – Computer and Information Science*, 457-469.

Sukma, A., Santoso, B. P., Ramadhan, D., Wiraswari, N. M., & Sari, T. R. (t.thn.). Klasifikasi Dokumen Bahasa Jawa Menggunakan Metode N-gram.

Takçı, H., & Ekinci, E. (2011). Minimal Feature Set in Language Identification and Finding Suitable Classification Method with It. 444-448.

Takcı, H., & Soğukpınar, Đ. (2004). Letter Based Text Scoring Method for Language Identification.