SKRIPSI

KLASIFIKASI DOKUMEN MENURUT BAHASA BERBASIS N-GRAM

Ricky Slamat Putra

NPM: 2013730011

PROGRAM STUDI TEKNIK INFORMATIKA
FAKULTAS TEKNOLOGI INFORMASI DAN SAINS
UNIVERSITAS KATOLIK PARAHYANGAN
2018

UNDERGRADUATE THESIS

N-GRAM BASED DOCUMENTS CLASSIFICATION BY LANGUAGE

Ricky Slamat Putra

NPM: 2013730011

DEPARTMENT OF INFORMATICS
FACULTY OF INFORMATION TECHNOLOGY AND SCIENCES
PARAHYANGAN CATHOLIC UNIVERSITY
2018

LEMBAR PENGESAHAN

KLASIFIKASI DOKUMEN MENURUT BAHASA BERBASIS N-GRAM

Ricky Slamat Putra

NPM: 2013730011

Bandung, 16 Mei 2018

Menyetujui,

Pembimbing Utama	Pembimbing Pendamping
Dott. Thomas Anung Basuki	-

Ketua Tim Penguji	Anggota Tim Penguji
Dr.rer.nat. Cecilia Esti Nugraheni	Rosa De Lima, M.Kom.

Mengetahui,

Ketua Program Studi

Mariskha Tri Adithia, P.D.Eng

PERNYATAAN

Dengan ini saya yang bertandatangan di bawah ini menyatakan bahwa skripsi dengan judul:

KLASIFIKASI DOKUMEN MENURUT BAHASA BERBASIS N-GRAM

adalah benar-benar karya saya sendiri, dan saya tidak melakukan penjiplakan atau pengutipan dengan cara-cara yang tidak sesuai dengan etika keilmuan yang berlaku dalam masyarakat keilmuan.

Atas pernyataan ini, saya siap menanggung segala risiko dan sanksi yang dijatuhkan kepada saya, apabila di kemudian hari ditemukan adanya pelanggaran terhadap etika keilmuan dalam karya saya, atau jika ada tuntutan formal atau non-formal dari pihak lain berkaitan dengan keaslian karya saya ini.

Dinyatakan di Bandung,
Tanggal 16 Mei 2018

Ricky Slamat Putra
NPM: 2013730011

ABSTRAK

Dalam era sekarang, dokumen semakin banyak seiring berkembangnya internet. Dokumen yang tersedia juga terdiri dari berbagai bahasa sehingga membuat internet sebagai gudang dokumen dari berbagai macam bahasa. Semakin banyak ketersediaan dokumen akan meningkatkan kompleksitas pencarian dokumen. Dalam pencarian informasi tingkat relevansi suatu dokumen sangat penting agar informasi yang didapatkan bersesuaian.

Di dunia terdapat beragam bahasa begitu juga dokumen yang ada. Oleh karena itu diperlukan klasifikasi dokumen. Klasifikasi adalah suatu proses pengelompokan berdasarkan ciri-ciri kemiripan. Klasifikasi yang akan dilakukan berdasarkan bahasa dari dokumen. Masing - masing bahasa memiliki karakteristik dalam frekuensi penggunaan huruf. Oleh karena itu digunakan metode n-gram dalam proses klasifikasi setiap dokumen.

N-gram adalah suatu metode pengolahan dokumen yang biasanya digunakan dalam *spelling correction*, *word prediction* dan pengolahan lainnya. Dalam skripsi ini n-gram akan digunakan sebagai metode yang mencari karakteristik dari masing-masing dokumen yang akan menghasilkan suatu *language model*. *Language model* ini berisi data frekuensi penggunaan huruf pada suatu dokumen. Data tersebut terdiri dari n-gram per karakter seperti unigram (satu karakter), bigram (dua karakter) dan trigram (tiga karakter).

Proses klasifikasi ini akan melakukan proses pelatihan sebelum dapat mengklasifikasi dokumen. Proses pelatihan ini adalah proses mencari karakteristik n-gram dari suatu bahasa dengan membuat hasil *language model* dari seluruh dokumen bahasa tersebut. Hasil *language model* tersebut merepresentasikan titik pusat *cluster* suatu bahasa.

K-Means adalah algoritma yang mencari jarak antar suatu data ke *cluster*. K-means akan digunakan untuk mencari jarak antara dokumen yang diklasifikasi dengan *cluster* bahasa yang ada. Jarak terkecil dari k-means ini merepresentasikan kesamaan karakteristik dokumen terhadap *cluster*.

Kata-kata kunci: N-gram, Klasifikasi, Bahasa, K-Means

ABSTRACT

In the present era, the number of documents is getting bigger along the development of the internet. Documents are available in many languages so as to make the Internet as a repository of documents from various languages. The more document is available will increase the complexity of document search. In search of information the relevance level of a document is very important so that the information obtained is relevant.

In the world there are various languages as well as existing documents. Therefore a document classification system is needed. Classification is a grouping process based on the characteristics of similarities. The classification will be based on the language of the document. Each language has characteristics in the frequency of the use of letters, therefore the n-gram method is used in the process for each document.

N-gram is a word processing method commonly used in spelling correction, word prediction and other word processing. In this undergraduate thesis n-gram will be used as a method that looks for the characteristics of each document that will produce a language model. This language model contains the frequency data of the use of letters in a document. The data consists of n-gram per character such as unigram (one character), bigram (two characters) and trigram (three characters).

This classification process will do the training process before it can classify the document. The process of this training is the process of finding the n-gram characteristics of a language by making the language model results from all the language documents. The language result of the model represents the cluster's central point of a language.

K-Means is an algorithm that searches the distance between data to the cluster. K-means will be used to find the distance between documents classified with existing language clusters. The smallest distance of k-means represents the similarity of document characteristics to the cluster

Keywords: N-gram, Classification, Language, K-Means

Dipersembahkan untuk kedua orang tua, diri sendiri, dan semua orang yang berperan pada pembuatan skripsi ini

KATA PENGANTAR

Puji syukur kepada Tuhan Yang Maha Esa. Oleh karena kasih-Nya yang begitu besar, penulis dapat menyelesaikan tugas akhir yang berjudul "Klasifikasi Dokumen Berbasis N-gram" dengan lancar. Penulis akan senantiasa berdoa agar penulisan ilmiah ini dapat memberi informasi yang bermanfaat dan menjadi inspirasi untuk penelitian penelitian berikutnya.

Selama menempuh kuliah, terutama pada saat penulisan ilmiah ini, penulis mendapat banyak dukungan dari berbagai pihak. Oleh karena itu, penulis ingin berterima kasih kepada pihak-pihak yang telah mendukung penulis:

1. Keluarga yang selalu memberikan dukungan kepada penulis berupa doa maupun nasihat.
2. Pak Thomas Anung Basuki sebagai dosen pembimbing yang telah sabar membimbing penulis dalam penyusunan penulisan ilmiah ini.
3. Bu Heni dan Bu Rosa selaku penguji yang telah memberikan kritik dan saran yang membangun sehingga tugas akhir ini menjadi lebih baik.
4. Special thanks untuk Ratna Wijaya yang telah membantu dan menyemangati sehingga skripsi ini berjalan dengan lancar
5. Jalaludin Yod, Dede, Pepew, Anton, Box, dan Jacinta yang telah membantu ketika saya menghadapi masalah dalam skripsi.
6. Member sumur Fready, Nando, Ratna yang telah rutin memberi asupan susu murni pada tengah malam.
7. Seluruh teman-teman IT yang membuat kenangan selama kuliah menjadi lebih indah dan seru.
8. Pihak lain yang tidak dapat disebutkan satu-persatu, yang telah memberikan kontribusi terhadap pembuatan tugas akhir ini.

Penulis menyadari bahwa penulisan ilmiah ini belum sempurna. Oleh karena itu penulis memohon maaf jika terdapat kesalahan. Penulis juga mengharapkan kritik dan saran yang membangun demi perbaikan dan kemajuan penulis.

Bandung, Mei 2018

Penulis

DAFTAR ISI

DAFTAR GAMBAR

DAFTAR TABEL

BAB 1

PENDAHULUAN

1.1 Latar Belakang

Dalam era sekarang, ketersediaan dokumen semakin banyak dan beragam seiring dengan berkembangnya internet. Sebelum adanya internet masyarakat harus mengunjungi tempat informasi itu berada dan memakan waktu yang banyak. Perkembangan internet khususnya dalam bidang pendidikan membuat masyarakat lebih mudah untuk mendapatkan informasi yang mereka inginkan, oleh karena itu dalam era ini masyarakat bisa bertukar dan mendapatkan informasi dengan mudah. Masyarakat bisa mengunduh dan membagikan informasi berupa dokumen kepada orang lain hanya dalam hitungan detik. Informasi yang ada pun tidak hanya dalam satu bahasa namun ada berbagai bahasa. Sehingga membuat internet menjadi gudang dokumen dengan berbagai macam bahasa.

Bagi sebagian masyarakat, kebutuhan informasi juga sangat penting. Jika jumlah dokumen semakin banyak, maka proses pencarian suatu dokumen tertentu juga semakin sulit didapatkan relevansinya. Dalam pencarian informasi, tingkat relevansi suatu dokumen sangat penting agar informasi yang didapatkan bersesuaian. Hal ini akan lebih mudah jika dokumen tersebut sudah tersedia sesuai dengan kategorinya masing-masing.

Di dunia terdapat beribu-ribu macam bahasa begitu juga dengan dokumen yang ada. Dengan adanya dokumen dalam berbagai bahasa tersebut maka diperlukan klasifikasi dokumen. Klasifikasi adalah suatu proses pengelompokan berdasarkan ciri-ciri persamaan dan perbedaan. Persamaan dan perbedaan utama dalam dokumen terletak pada bahasa yang digunakannya. Sehingga klasifikasi yang mudah digunakan pada dokumen adalah klasifikasi bahasa.

Pada skripsi ini, akan dibuat sebuah perangkat lunak yang dapat mengklasifikasi dokumen berdasarkan bahasa. Dengan menggunakan perangkat lunak tersebut, pengguna dapat mengelompokkan dokumen berdasarkan bahasa yang digunakan. Perangkat lunak ini menggunakan metode n-gram untuk klasifikasi.

1.2 Rumusan Masalah

Berdasarkan latar belakang yang sudah disebutkan sebelumnya, maka rumusan masalah yang dibangun adalah sebagai berikut:

1. Bagaimana konsep n-gram dan implementasinya pada perangkat lunak ?
2. Bagaimana menggunakan n-gram pada klasifikasi dokumen ?
3. Seberapa baik n-gram dapat digunakan untuk klasifikasi dokumen ?

1.3 Tujuan

Adapun tujuan penelitian yang dilakukan adalah untuk:

1. Memahami konsep n-gram dan implementasinya pada klasifikasi dokumen.

2. Membuat perangkat lunak yang dapat mengklasifikasi dokumen dengan menggunakan n-gram.
3. Mengukur seberapa baik n-gram dapat digunakan untuk mengklasifikasi dokumen.

1.4 Batasan Masalah

Agar pembahasan masalah tidak terlalu luas, masalah yang dikaji di dalam penelitian ini memiliki batasan, yaitu:

1. Dokumen yang diklasifikasi adalah dokumen yang memiliki bahasa yang menggunakan karakter ASCII
2. Dokumen yang diklasifikasi memiliki bahasa yang cukup populer dalam pengertian bahasa tersebut memiliki dokumen yang cukup banyak dalam internet. Hal ini untuk memudahkan proses pelatihan dokumen pada suatu profil bahasa. Jika bahasa yang dokumennya di internet tersedia sedikit maka tingkat keakuratan klasifikasi akan rendah.
3. Penggunaan n-gram pada klasifikasi dokumen terbatas hanya sampai pada trigram.
4. Bahasa yang diuji hanya 5 bahasa yaitu Indonesia, Inggris, Melayu, Jerman dan Italia.

1.5 Metodologi

Tahap-tahap yang dilakukan dalam penelitian ini adalah sebagai berikut:

1. Studi pustaka tentang n-gram, pembelajaran mesin dan klasifikasi dokumen berdasarkan bahasa.
2. Mengumpulkan dokumen sebanyak 500 file untuk masing-masing bahasa dari wikipedia.
3. Melakukan analisis n-gram terhadap klasifikasi dokumen.
4. Melakukan perancangan
5. Mengimplementasikan n-gram untuk klasifikasi dokumen.
6. Melakukan pengujian aplikasi yang dibuat.

1.6 Sistematika Pembahasan

Sistematika pembahasan dari penelitian ini, adalah:

1. Bab 1 Pendahuluan

   Bab ini berisi latar belakang, rumusan masalah, tujuan, batasan masalah, metodologi penelitian, dan sistematika pembahasan yang merupakan ringkasan dari setiap bab secara umum.

2. Bab 2 Dasar Teori

   Bab ini berisi teori-teori dasar mengenai n-gram dan klasifikasi dokumen berdasarkan bahasa.

3. Bab 3 Analisis

   Bab ini berisi tentang analisis n-gram terhadap klasifikasi dokumen.

4. Bab 4 Perancangan Perangkat Lunak

   Bab ini berisi perancangan perangkat lunak untuk aplikasi klasifikasi dokumen dan implementasi n-gram pada klasifikasi dokumen.

5. Bab 5 Implementasi dan Pengujian

   Bab ini berisi implementasi n-gram pada klasifikasi dokumen dan pengujian pada bab ini mencakup pengujian fungsional dan pengujian eksperimen.

6. Bab 6 Kesimpulan dan Saran

   Bab ini berisi kesimpulan yang diperoleh selama pengembangan perangkat lunak dan saran untuk pengembangan lebih lanjut dari perangkat lunak agar dapat mencapai hasil yang lebih baik.