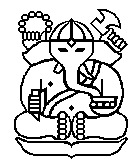
STEMMER BAHASA INDONESIA

IF 5282 – PEMROSESAN BAHASA ALAMI

DITARI SALSABILA 23514004

HASAN ASYARI 23514025

IRFAN AFIF 23514027



SEKOLAH PASCA SARJANA TEKNIK INFORMATIKA

SEKOLAH TEKNIK ELEKTRO & INFORMATIKA

INSTITUT TEKNOLOGI BANDUNG

2015

Contents

[PEMOTONG KALIMAT 3](#_Toc411721447)

[TOKENISASI 4](#_Toc411721448)

[1. Algoritma pemisah tanda baca 4](#_Toc411721449)

[2. Algoritma tokenisasi : 4](#_Toc411721450)

[Stemmer Bahasa Indonesia 5](#_Toc411721451)

[1. Pemeriksaan kata-kata plural 7](#_Toc411721452)

[2. Remove Prefix 8](#_Toc411721453)

[3. Remove Suffix 9](#_Toc411721454)

[4. Pemeriksaan Akhir 10](#_Toc411721455)

STEMMING BAHASA INDONESIA

# PEMOTONG KALIMAT

Proses pemotongan kalimat yang dilakukan adalah memecah teks masukan menjadi kalimat-kalimat. Permasalahan utama dalam proses ini adalah simbol akhir kalimat seperti tanda titik, tidak hanya digunakan sebagai tanda akhir kalimat, tetapi dapat juga digunakan di tengah kalimat. Tanda titik terutama banyak digunakan di singkatan-singkatan. Oleh karena itu dibutuhkan strategi pemisah kalimat yang dapat menentukan akhir dari kalimat dalam teks masukan.

Proses pertama yang dilakukan adalah penggantian baris (end line : \n) diganti dengan spasi. Selanjutnya masukan dipecah kata perkata berdasarkan karakter spasi. Setelah itu proses dilakukan dengan pemeriksaan kata per kata.

Kondisi-kondisi yang digunakan dalam pemeriksaan kata adalah :

1. Jika bertemu dengan simbol-simbol : (."), (!), (?), (?", (!"). Tanda seru dan tanda Tanya sudah ditentukan dengan pasti sebagai akhir kalimat. Begitu juga tanda seru dan tanda tanya yang diikuti oleh kutip, digunakan sebagai akhir kalimat. Hal ini terjadi pada kasus dimana terdapat kalimat langsung. Tanda titik yang diikuti tanda kutip juga digunakan sebagai penanda akhir kalimat.
2. Jika kata diakhiri tanda titik, maka ada beberapa kondisi yang diperiksa :
   1. Jika kata lebih dari 5 huruf, maka dianggap sebagai akhir kalimat.
   2. Jika kata terdapat pada kamus seperti Rp dan diperiksa sebagai huruf kecil, maka kata tidak dianggap sebagai akhir kalimat.
   3. Jika kata diawali dengan huruf besar, maka tidak dianggap sebagai akhir kalimat.
   4. Selain itu dianggap sebagai akhir kalimat.

# TOKENISASI

## Algoritma pemisah tanda baca

Input berupa array list kata. Dilakukan pengecekan untuk tiap kata, jika kata tersebut mengandung karakter non alfa numerik maka tiap karakter pada kata akan dicek satu-persatu. Dilakukan pemisahan kata menjadi beberapa bagian berdasarkan karakter non alfa numerik tersebut. Output berupa array list di mana kata dan karakter alfanumerik sudah terpisah.

## Algoritma tokenisasi :

Input berupa array list yang dapat berupa kata atau karakter non alfanumerik. Tiap elemen pada array list diubah ke huruf kecil terlebih dahulu. Dilakukan pengecekan untuk setiap elemen dan elemen berikutnya pada array list. Jika elemen saat ini dan elemen berikutnya berada di daftar kata majemuk, maka kedua kata tersebut dianggap sebagai satu token dan digabungkan. Daftar kata majemuk disimpan di file teks frasa, yang dipanggil berdasarkan huruf awal pada elemen. Output berupa array list yang tiap itemnya dapat berupa kata dasar atau kata majemuk/frase.

# Stemmer Bahasa Indonesia

Stemming adalah proses untuk mengubah kata berimbuhan, maupun kata hasil dari proses tokeninasi ke bentuk kata dasarnya.

Algortima stemming yang kami gunakan sebagian besar bersumber dari Library PHP Stemming bahasa Indonesia yang disediakan di <https://github.com/sastrawi/sastrawi>. Struktur Code Program yang digunakan dalam library tersebut banyak yang **kami terjemahkan langsung** ke Program Stemming yang kami buat ini.

Ada beberapa proses yang kami ubah secara dasar, antara lain Pembacaan Rule Ambiguitas yang kami terjemahkan langsung dari Algoritma Nazief dan Adriani, Implementasi hashmap untuk dictionary lookup yang digunakan serta perubahan struktur rule reader agar memudahkan membaca rule dari file.

Berikut adalah FlowChart Stemming secara umum dalam program yang kami buat. Berikut adalah bagaimana proses stemming kami bekerja:



Ada beberapa tahapan dasar dari program stemming yang ada dalam tugas ini:

1. Pemeriksaan untuk kata-kata plural.
2. Pemeriksaan kata diluar rule umum yang ada seperti pelajar dan belajar.
3. Pemeriksaan jika kata tersebut hanya terdiri kurang dari 4 huruf.
4. Perubahan struktur kata melalui penghapusan Prefix maupun Suffix dari kata yang diperiksa
5. Hingga pada tahapan pemeriksaan akhir, yaitu perubahan prefix dan memeriksa kata yang telah dihapus suffixnya terlebih dahulu.

Seluruh tahapan diatas dilengkapi dengan dictionary lookup untuk memastikan jika kata awal maupun kata perubahan telah tersedia dalam kamus kata dasar.

## Pemeriksaan kata-kata plural



Tahapan ini memeriksa apabila kata awal yang dimasukkan adalah kata plural/jamak. Jika kata tersebut kata plural maka kata tersebut akan dipisahkan terlebih dahulu menjadi kata individual, melalui aturan dalam flowchart diatas.

## Remove Prefix



Tahapan ini menghapus prefix dari kata berdasarkan rule prefix yang sudah disediakan. Tahapan ini di loop tiga kali, karena diasumsikan bahwa setiap kata berimbuhan maksimal memiliki 3 prefix. Setiap perubahan dari penghapusan prefix, dilakukan pengecekan dikamus kata dasar untuk memastikan penghapusan prefix tidak merubah dan menyebabkan kata menjadi tidak beraturan.

## Remove Suffix



Tahapan ini untuk menghapus dan memeriksa akhiran dari kata yang dihapus. Ada tiga jenis Suffix yang dihapus melalui tahapan ini, Possesive Prounoun, Particle dan Derivitional Suffix. Ketiga jenis suffix tersebut dihapus secara bertahap sesuai flowchart diatas.

## Pemeriksaan Akhir



Tahapan ini dilakukan setelah penghapusan prefix dan suffix ditahap sebelumnya. Tahapan ini melalui **reversed removal checker.** Dimana kata yang sudah dihapus suffix-nya akan diperiksa lagi dengan melalui penghapusan prefix seperti pada flowchart diatas. Tahapan ini untuk memastikan bahwa ketika penghapusan suffix dan prefix tidak ada kata yang terlewatkan akibat perubahan struktur ditahapan sebelumnya.