Makalah

Sistem Temu Kembali Informasi

(Tokenisasi, Stopword Removal, dan Stemming)

Disusun oleh :

Arif Budi Setiawan (15.01.53.0064)

Saiful Budi Yanto (15.01.53.0081)

Dosen Pengampu :

Dr. Drs. Eri Zuliarso,M.Kom

Fakultas Teknologi Informasi

Program Studi S1 - Teknik Informatika

Universitas Stikubank Semarang

2017

DAFTAR ISI

KATA PENGANTAR

Puji dan syukur penulis panjatkan kepada Tuhan Yang Maha Esa atas selesainya makalah yang telah kami buat dengan judul "Tokenisasi, Stopword Removal, Stemming, TF/IDF". Atas dukungan yang diberikan dalam penyusunan makalah ini, maka kami mengucapkan terima kasih kepada :

1. Dr.Drs. Eri Zuliarso,M.Kom. selaku Dosen Mata Kuliah Sistem Temu Kembali Informasi.
2. Semua pihak yang turut serta memberikan motivasi.

Semoga makalah ini dapat memberikan manfaat bagi para pembaca, Penulis menyadari masih banyak kekurangan dalam makalah ini. Oleh karena itu kritik dan saran yang membangun dari rekan-rekan sangat dibutuhkan untuk penyempurnaan makalah ini.

Semarang, 14 September 2017

Penulis

BAB 1

PENDAHULUAN

1.1.Latar Belakang

Sistem Temu Kembali Informasi (Information Retrieval) digunakan untuk menemukan kembali informasi-informasi yang relevan terhadap kebutuhan pengguna dari suatu kumpulan informasi secara otomatis. Salah satu aplikasi umum dari sistem temu kembali informasi adalah search-engine atau mesin pencarian yang terdapat pada jaringan internet. Pengguna dapat mencari halaman-halaman Web yang dibutuhkannya melalui mesin tersebut.

Oleh sebab itu, metode-metode untuk menemukan kembali teks terus ditingkatkan. Salah satunya yaitu Information Retrieval System (IRS). IRS merupakan pencarian informasi dalam satu atau lebih dokumen, atau mencari informasi dari database. IR menggunakan perhitungan untuk menentukan apakah informasi tersebut relevan bagi penggunanya. Di dalam IR akan melalui beberapa tahapan, yaitu Text Preprocessing, Pembobotan, dan Indexing.

1.2.Rumusan Masalah

1. Apa pengertian Sistem Temu Kembali Informasi dan tujuannya ?
2. Apa pengertian dari Tokenisasi, Stopword Removal, dan Stemming?
3. Bagaimana proses dari Tokenisasi, Stopword Removal, dan Stemming?

1.3.Tujuan

Dapat memahami pengertian, proses, dan manfaat dari Tokenisasi, Stopword Removal, dan Stemming.

BAB 2

PEMBAHASAN

2.1.Pengertian Sistem Temu Kembali Informasi

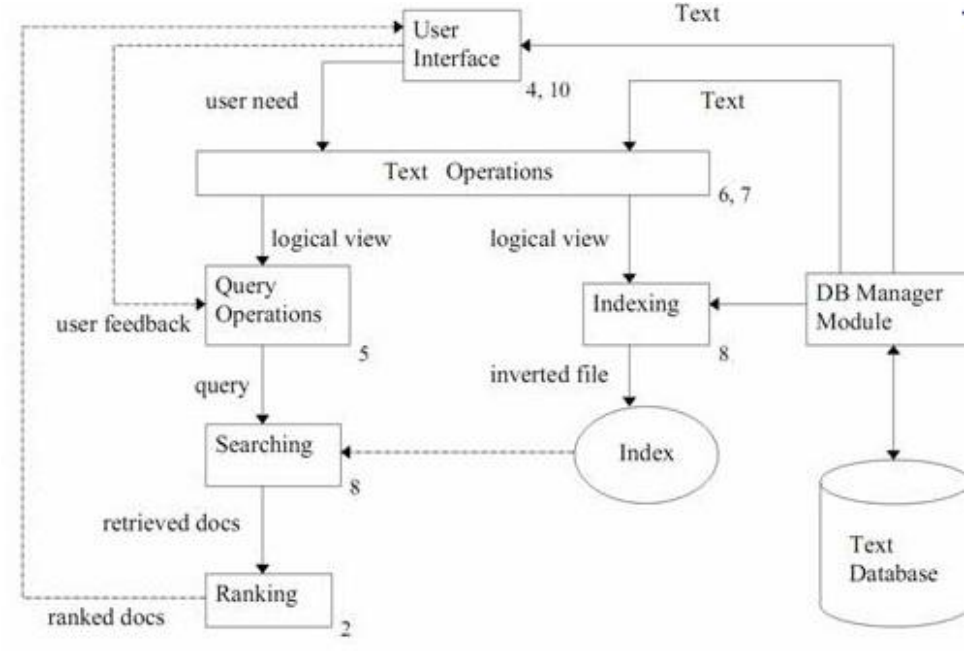
Sistem temu kembali informasi berasal dari kata Information Retrieval System (IRS). Temu kembali informasi adalah sebuah media layanan bagi pengguna untuk memperoleh informasi atau sumber informasi yang dibutuhkan oleh pengguna. Sistem temu kembali informasi merupakan sistem informasi yang berfungsi untuk menemukan informasi yang relevan dengan kebutuhan pemakai. Sistem temu kembali informasi berfungsi sebagai perantara kebutuhan informasi pengguna dengan sumber informasi yang tersedia. Pengertian yang sama mengenai sistem temu kembali informasi menurut Sulistyo-Basuki sistem temu kembali informasi adalah kegiatan yang bertujuan untuk menyediakan dan memasok informasi bagi pemakai sebagai jawaban atas permintaan atau berdasarkan kebutuhan pemakai. Dapat dinyatakan bahwa sistem temu kembali informasi memiliki fungsi dalam menyediakan kebutuhan informasi sesuai dengan kebutuhan dan permintaan penggunanya.

Definisi lain yang mengemukakan bahwa: “Sistem temu kembali informasi adalah suatu proses yang dilakukan untuk menemukan dokumen yang dapat memberikan kepuasan bagi pengguna dalam memenuhi kebutuhan informasinya’. Tujuan utama sistem temu kembali informasi adalah untuk menemukan dokumen yang sesuai dengan kebutuhan informasi pengguna secara efektif dan efisien, sehingga dapat memberikan kepuasan baginya, dan sasaran akhir dari sistem temu kembali informasi adalah kepuasan pemakai.

Maka dapat disimpulkan bahwa sistem temu kembali informasi merupakan sebuah sistem yang berguna dalam memanggil dan menempatkan dokumen dari/dalam basis data sesuai dengan permintaan pengguna. Sistem temu kembali informasi memiliki tujuan akhir, yaitu memberikan kepuasan informasi bagi pengguna sistem. Jadi, temu kembali informasi merujuk pada keseluruhan kegiatan yang meliputi pembuatan wakil informasi (representation), penyimpanan (storage), pengaturan (organization) sampai kepada pengambilan (access).

2.2.Konsep Dasar Information Retrieval System

Konsep dasar dalam Information Retrieval System terdiri dari Indexing, Searching dan perengkingan relevansi keyword query. Dimana proses indexing dilakukan untuk membentuk database index terhadap koleksi dokumen yang dimasukkan, atau dengan kata lain, indexing merupakan proses persiapan yang dilakukan terhadap dokumen sehingga dokumen siap untuk retrive. Proses indexing sendiri meliputi 2 proses, yaitu dokukmen indexing dan term indexing. Dari term indexing akan dihasilkan koleksi kata yang akan digunakan untuk meningkatkan performansi pencarian pada tahap selanjutnya.



Gambar 1 Information Retrieval System

2.3.Tokenisasi/Word Token/Parsing

Tahap tokenizing disebut juga sebagai parsing Yaitu pengambilan kata-kata (term) dari kumpulan dokumen menjadi kumpulan term dengan cara menghapus karakter tanda baca yang terdapat pada dokumen dan mengubah kumpulan term menjadi lowercase

Didalam sistem temu kembali terdapat proses text mining yang memiliki definisi menambang data yang berupa teks dimana sumber data biasanya didapat dari dokumen, dan tujuannya adalah mencari kata-kata yang dapat mewakili isi dari dokumen sehingga dapat dilakukan analisa keterhubungan antar dokumen. Tahapan yang dilakukansecara umum dalam text mining adalah : tokenizing, filtering, stemming, tagging dan analyzing.

Pada proses tersebut masing-masing melakukan fungsinya masing-masing. Proses tokenizing adalah tahap pemotongan string input berdasarkan tiap kata yang menyusunnya. Proses ini menghasilkan kata –kata yang berdiri sendiri. Sedangkan proses filtering adalah tahap mengambil kata – kata penting dari hasil token. Bisa menggunakan algoritma stop list (membuang kata-kata yang kurang penting atau word list. Proses ini akan dihasilkan kata yang penting saja dan membuang kata kata yang kurang penting.

Tahap stemming adalah tahap mencari root dari tiap kata hasil filtering. Tahap ini akan menghilangkan imbuhan pada suatu kalimat. Sedangkan tahap tagging adalah tahap mencari bentuk awal/ root dari tiap kata lampau atau kata hasil stemming.

Apakah proses tokenizing penting untuk dilakukan ?

Sangat penting, karena didalam proses ini merupakan tahap pemotongan string input berdasarkan tiap kata yang menyusunnya. Proses ini menghasilkan kata –kata yang berdiri sendiri. Dan kemudian dilakukan proses filtering. Tahap filtering mengambil kata-kata yang penting dari hasil proses token. Dan setelah itu baru dilakukan proses stemming , tagging dan analyzing. Sehingga antara tahap – tahap ini saling terkait dan berhubungan.

Seperti yang terlihat pada gambar pada proses preprosesing untuk tokenisasi, semua term dalam dokumen yang dibaca diganti dengan huruf kecil. Setelah itu tiap term akan dicek apakah tanda baca atau tidak. Jika tanda baca maka akan dihapus/dibuang. Prosesakan dilanjutkan untuk membuat term menjadi token-token yang terpisah.

2.4.Stopword Removal / Filtering

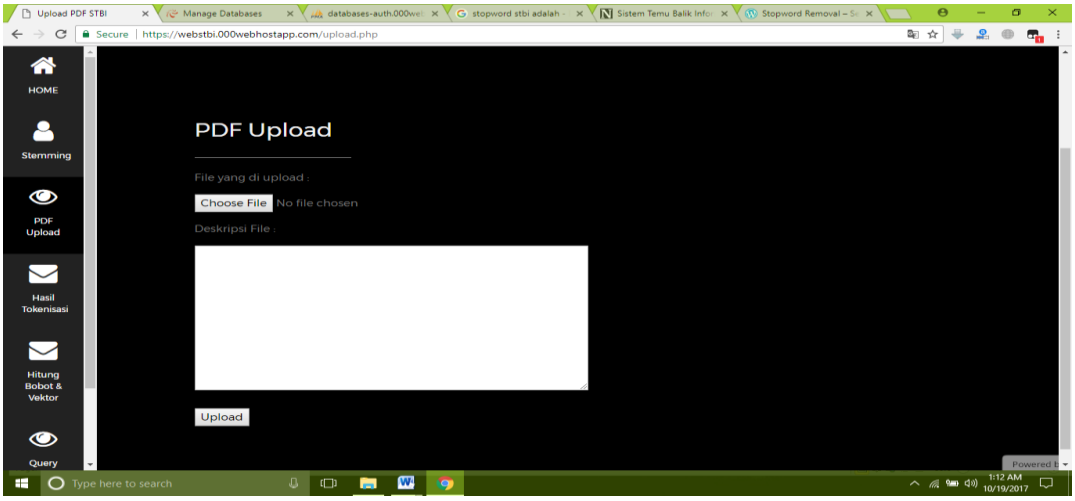
Tahap Stopword Removal atau Filtering adalah Proses penghapusan atau pembuangan kata-kata yang sering ditampilkan dalam dokumen atau kata-kata yang tidak deskriptif (tidak penting) yang dapat dibuang dengan pendekatan bag-of-words seperti: and, or, not, tetapi, yang, sedangkan dan sebagainya.

2.5.Stemming

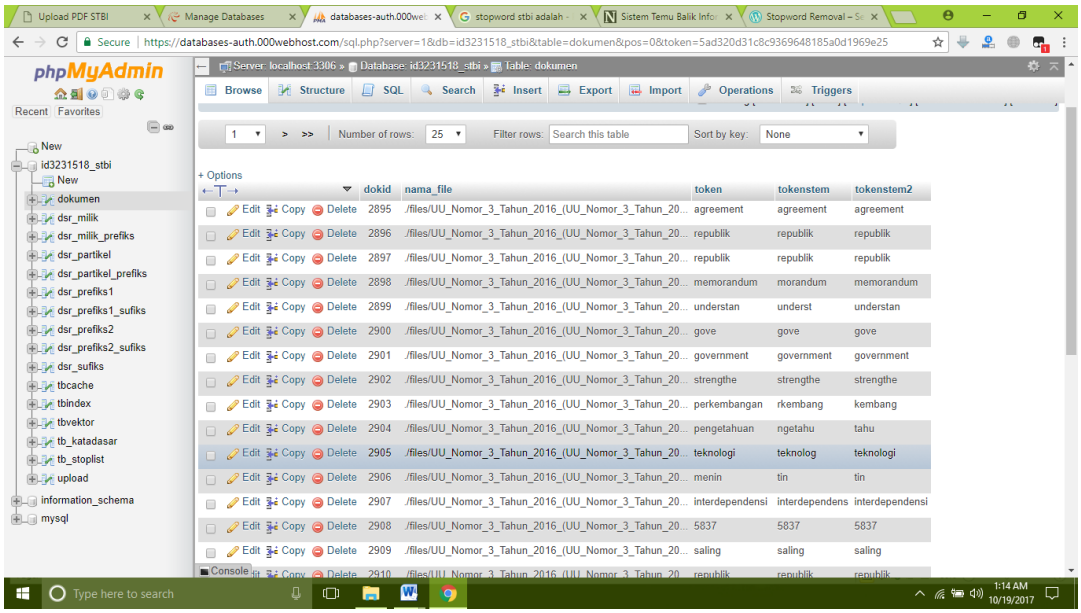
Proses stemming adalah proses pembuangan prefix dan sufix suatu kata bentukan menjadi kata dasar. Proses stemming dilakukan untuk mendapatkan hasil peringkat halaman informasi yang relevan. Proses stemming dilakukan dengan cara menghilangkan semua imbuhan (affixes) baik yang terdiri dari awalan (prefixes), sisipan (infixes), akhiran (suffixes) dan confixes (kombinasi dari awalan dan akhiran) pada kata turunan. Stemming digunakan untuk mengganti bentuk dari suatu kata menjadi kata dasar dari kata tersebut yang sesuai dengan struktur morfologi bahasa Indonesia yang benar (Tala, 2003).

Stemming digunakan untuk mereduksi bentuk term untuk menghindari ketidakcocokan yang dapat mengurangi recall, di mana term-term yang berbeda namun memiliki makna dasar yang sama direduksi menjadi satu bentuk.

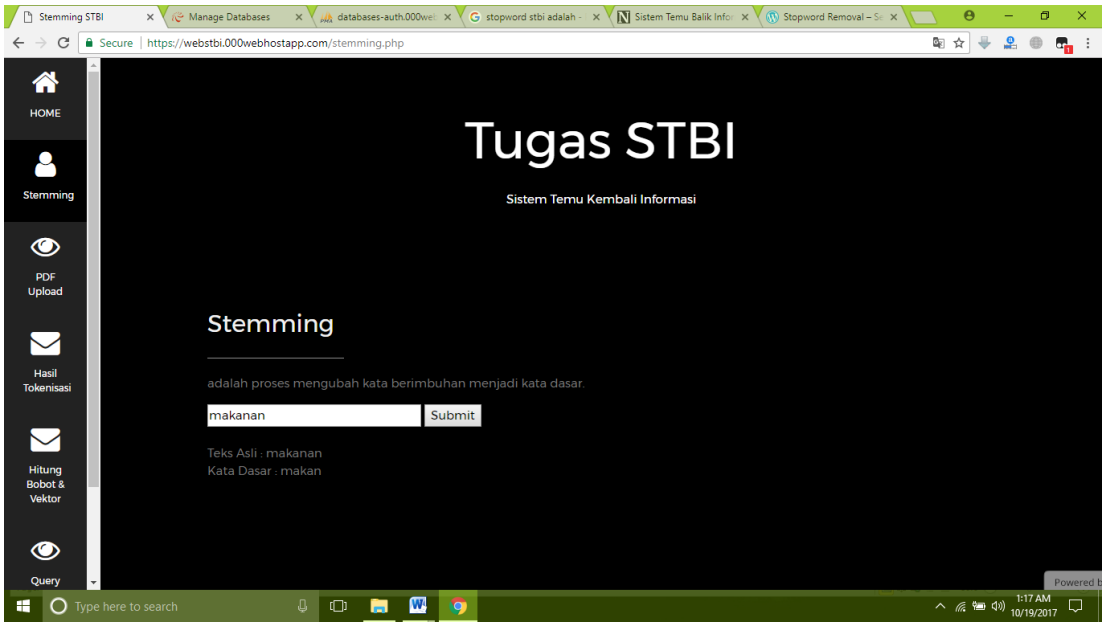
2.6.Preview Aplikasi Tokenisasi, Stopword Removal, dan Stemming



Gambar 2 Form upload dokumen untuk dilakukan tokenisasi



Gambar 3 Hasil dari proses tokenisasi yang sudah dilakukan stopword removal



Gambar 4 Proses Stemming dengan penghilangan imbuhan menjadi kata dasar

BAB III

PENUTUP

3.1.Kesimpulan

1. Token adalah kata-kata yang dipisah-pisah dari teks aslinya tanpa mempertimbangkan adanya duplikasi.
2. Tokenisasi adalah proses mengubah dokumen menjadi kumpulan term dengan cara menghapus semua karakter tanda baca yang terdapat pada token. Hingga pada akhirnya yang diperoleh hanya kumpulan kata-kata dari suatu teks/dokumen.
3. Stoplist atau stopword adalah kata-kata yang tidak deskriptif (tidak penting) yang dapat dibuang dengan pendekatan bag-of-words.
4. Stopword removal disebut juga filtering, adalah tahap pemilihan kata-kata penting dari hasil token, yaitu kata-kata apa saja yang akan digunakan untuk mewakili dokumen.
5. Proses Stemming adalah proses pembentukan kata dasar.
6. Stemming digunakan untuk mereduksi bentuk term untuk menghindari ketidakcocokan yang dapat mengurangi recall, di mana term-term yang berbeda namun memiliki makna dasar yang sama direduksi menjadi satu bentuk.

DAFTAR PUSTAKA

https://1104505056unud.wordpress.com/2015/06/12/temu-kembali-informasi-pada-opac-di-unit-perpustakaan-perguruan-tinggi/

http://informationretrievalsystem.blogspot.co.id/2012/08/definisi-information-retrieval.html

http://informationretrievalsystem.blogspot.co.id/2012/08/tujuan-indexing.html

http://informationretrievalsystem.blogspot.co.id/2013/02/cara-kerja-informasi-retrieval.html

http://sistemtemukembaliinformasi.blogspot.co.id/2012/07/tokenisasi.html