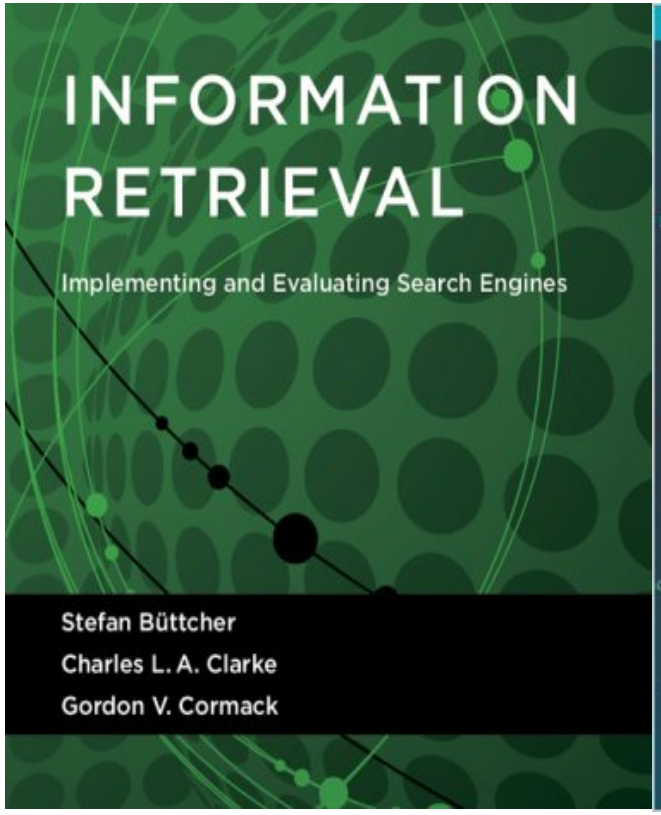
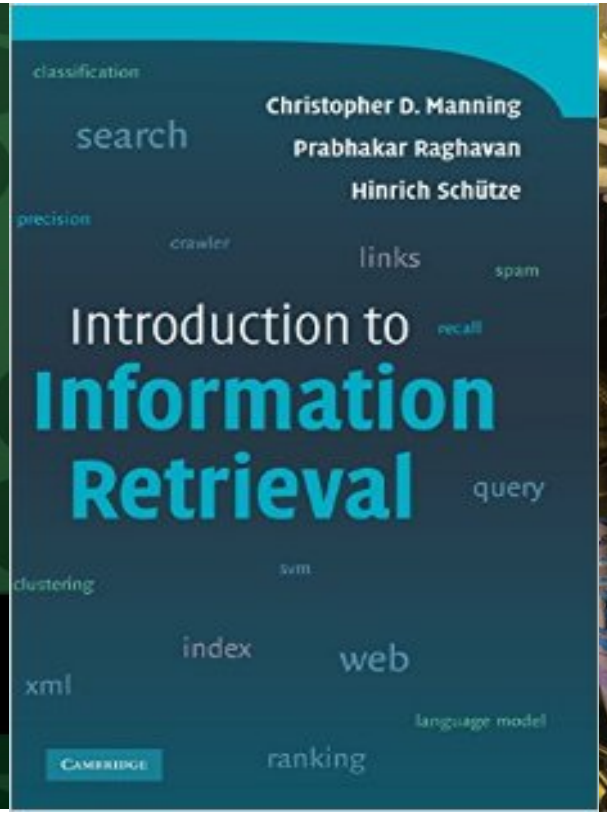
Sistem Temu Kembali Informasi

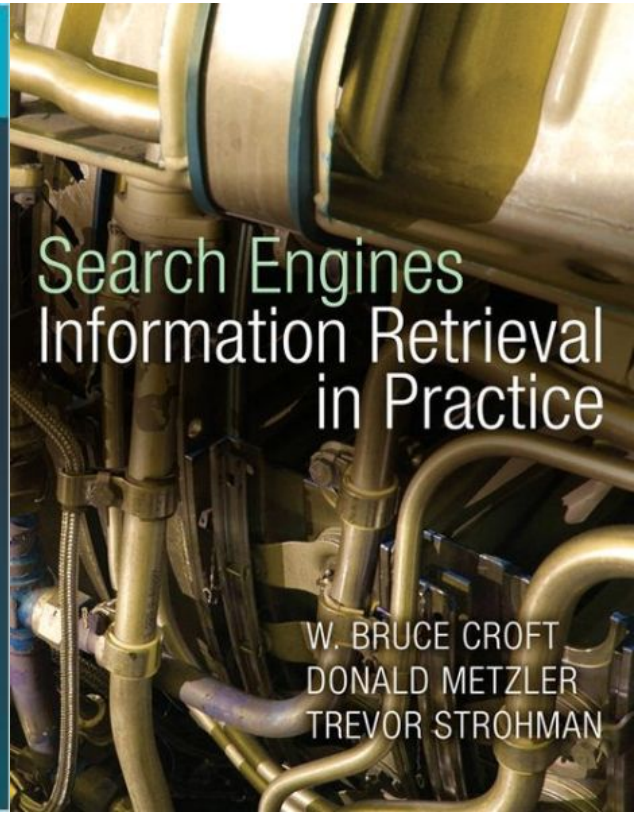
"Document Preprocessing"

Tim Dosen STKI

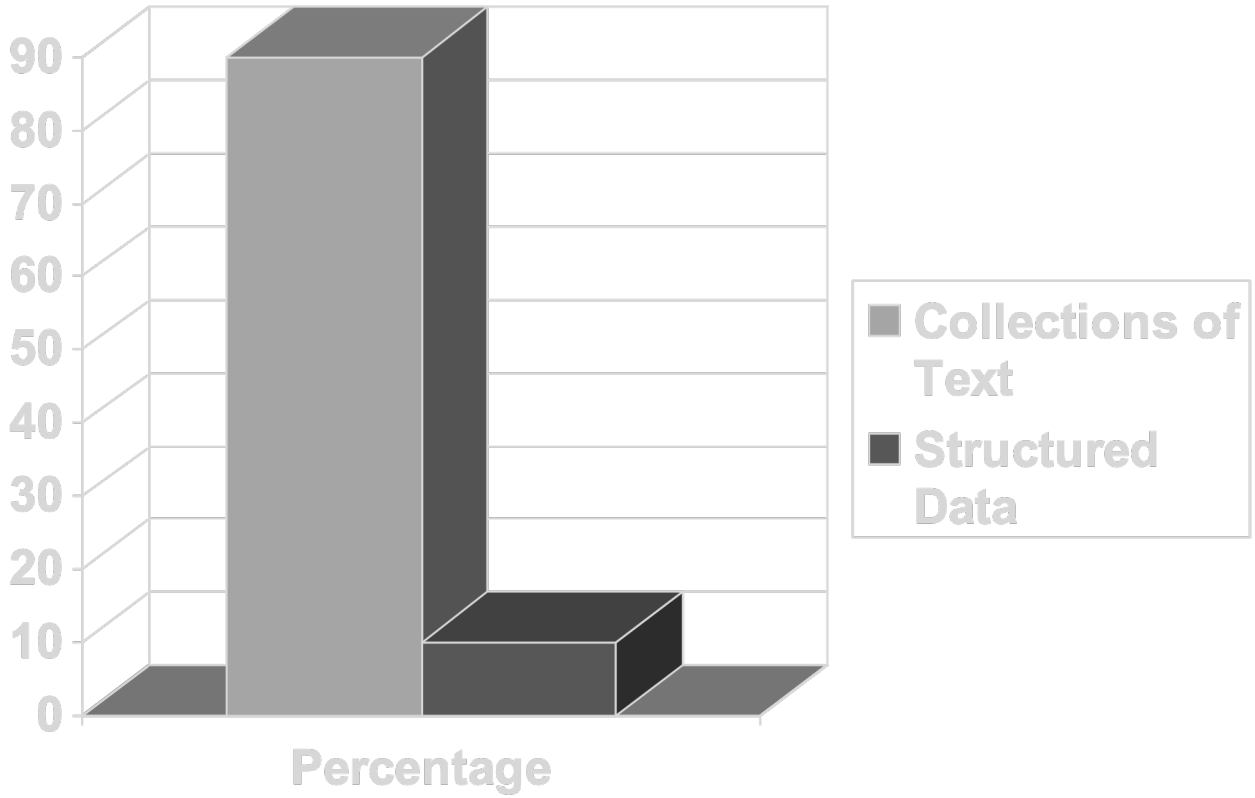
Buku Penunjang & Literatur







Latar Belakang



Latar Belakang

- Dokumen-dokumen yang ada kebanyakan **tidak memiliki struktur yang pasti** sehingga informasi di dalamnya tidak bisa diekstrak secara langsung.
- Tidak semua kata **mencerminkan** makna/isi yang terkandung dalam sebuah dokumen.

Latar Belakang

- Preprocessing diperlukan untuk memilih kata yang akan digunakan sebagai **indeks**
- **Indeks** ini adalah kata-kata yang **mewakili dokumen** yang nantinya digunakan untuk membuat pemodelan untuk Information Retrieval maupun aplikasi teks mining lain.

Definisi

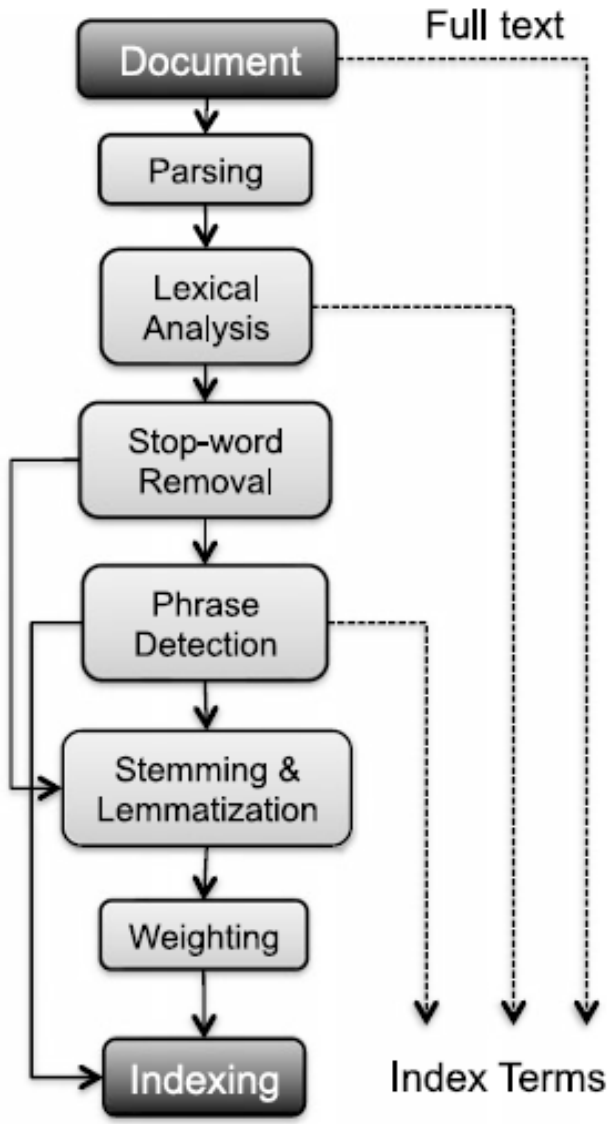
Definisi Pemrosesan Teks (Text Preprocessing) adalah suatu proses **pengubahan** bentuk data yang **belum terstruktur** menjadi data yang **terstruktur** sesuai dengan kebutuhan, untuk proses mining yang lebih lanjut (sentiment analysis, peringkasan, clustering dokumen, etc.).

Singkatnya...

- **Preprocessing** adalah **merubah** teks menjadi term index
- Tujuan: menghasilkan sebuah set **term index** yang bisa **mewakili** dokumen

Langkah-langkah Text Pre-processing

- Langkah-langkah umum dalam Text Pre-processing


Full text
Document
Parsing
Lexical Analysis
Stop-word Removal
Phrase Detection
Stemming & Lemmatization
Weighting
Indexing
Index Terms

Langkah 1 : Parsing

- Tulisan dalam sebuah dokumen bisa jadi terdiri dari **berbagai macam** bahasa, character sets, dan format;
- Sering juga, dalam satu dokumen yang sama berisi tulisan dari **beberapa bahasa**. Misal, sebuah email berbahasa Indonesia dengan lampiran PDF berbahasa Inggris.

Langkah 1 : Parsing

Parsing Dokumen berurusan dengan **pengenalan dan "pemecahan"** struktur dokumen menjadi komponen-komponen terpisah. Pada langkah preprocessing ini, kita menentukan mana yang dijadikan **satu unit dokumen**;

Langkah 1 : Parsing

Contoh, **email dengan 4 lampiran** bisa dipisah menjadi **5 dokumen** : 1 dokumen yang merepresentasikan isi (body) dari email dan 4 dokumen dari masing-masing lampiran

Langkah 1 : Parsing

- Contoh lain, buku dengan **100 halaman** bisa dipisah menjadi **100 dokumen**; masing-masing halaman menjadi 1 dokumen
- **Satu *tweet*** bisa dijadikan sebagai **1 dokumen**. Begitu juga dengan sebuah komentar pada forum atau review produk.

Langkah 2 : Lexical Analysis

Lebih populer disebut Lexing atau **Tokenization / Tokenisasi**

Langkah 2 : Lexical Analysis

- Tokenisasi adalah proses **pemotongan** string input berdasarkan tiap kata penyusunnya.
- Pada prinsipnya proses ini adalah **memisahkan** setiap kata yang menyusun suatu dokumen.

Langkah 2 : Lexical Analysis

Pada proses ini dilakukan **penghilangan** angka, tanda baca dan karakter selain huruf alfabet, karena karakter-karakter tersebut dianggap sebagai pemisah kata (delimiter) dan tidak memiliki pengaruh terhadap pemrosesan teks.

Langkah 2 : Lexical Analysis

Pada tahapan ini juga dilakukan proses ***case folding***, dimana semua huruf diubah menjadi huruf kecil.

Langkah 2 : Lexical Analysis

- Pada tahapan ini juga **Cleaning**
- Cleaning adalah proses **membersihkan** dokumen dari komponen-komponen yang tidak memiliki hubungan dengan informasi yang ada pada dokumen, seperti tag html, link, dan script, dsb.

Tokens, Types, and Terms

- Text: “apakah culo dan boyo bermain bola di depan rumah boyo?”
- **Token** adalah kata-kata yang dipisah-pisah dari teks aslinya tanpa mempertimbangkan adanya duplikasi
- **Token**: “culo”, “dan”, “boyo”, “bermain”, “bola”, “di”, “depan”, “rumah”, “boyo”

Tokens, Types, and Terms

- Text: “apakah culo dan boyo bermain bola di depan rumah boyo?”
- **Type** adalah token yang memperhatikan adanya duplikasi kata. Ketika ada duplikasi hanya dituliskan sekali saja.
- **Type**: “culo”, “dan”, “boyo”, “bermain”, “bola”, “di”, “depan”, “rumah”

Tokens, Types, and Terms

- Text: “apakah culo dan boyo bermain bola di depan rumah boyo?”
- **Token**: “culo”, “dan”, “boyo”, “bermain”, “bola”, “di”, “depan”, “rumah”, “boyo”
- **Type**: “culo”, “dan”, “boyo”, “bermain”, “bola”, “di”, “depan”, “rumah”

Tokens, Types, and Terms

- Text: “apakah culo dan boyo bermain bola di depan rumah boyo?”
- **Term** adalah type yang sudah dinormalisasi (dilakukan stemming, filtering, dsb)
- **Term** : “culo”, “boyo”, “main”, “bola”, “depan”, “rumah”

Tokens, Types, and Terms

- Text: “apakah culo dan boyo bermain bola di depan rumah boyo?”
- **Token**: “culo”, “dan”, “boyo”, “bermain”, “bola”, “di”, “depan”, “rumah”, “boyo”
- **Type**: “culo”, “dan”, “boyo”, “bermain”, “bola”, “di”, “depan”, “rumah”
- **Term**: “culo”, “boyo”, “main”, “bola”, “depan”, “rumah”

Contoh Tokenisasi

Teks English	They are applied to the words in the texts.
Tokens	they
	are
	applied
	to
	the
	words
	in
	the
	texts

Contoh Tokenisasi

Teks Bahasa	Namanya adalah Santiago. Santiago sudah memutuskan untuk mencari sang alkemis.
Tokens	namanya
	adalah
	santiago
	santiago
	sudah
	memutuskan
	untuk
	mencari
	sang
	alkemis

Langkah 3 : Stopword Removal

- Disebut juga **Filtering**
- **Filtering** adalah tahap **pemilihan** kata-kata penting dari hasil token, yaitu kata-kata apa saja yang akan digunakan untuk mewakili dokumen.

Stopword Removal : Metode

- Algoritma **stoplist**
- **Stoplist** atau **stopword** adalah **kata-kata yang tidak deskriptif (tidak *penting*)** yang dapat dibuang dengan pendekatan *bag-of-words*.

Stopword Removal : Metode

- Algoritma **stoplist**
- Kita memiliki database kumpulan **kata-kata yang tidak deskriptif (tidak *penting*),** kemudian kalau hasil tokenisasi itu ada yang merupakan kata tidak penting dalam database tersebut, maka hasil tokenisasi itu dibuang.

Stopword Removal : Metode

- Algoritma **stoplist**
- Contoh stopwords adalah i’m, you, one, two, they, are, to, the, in, dst.

Stopword Removal : Metode

Hasil Token	Hasil Filtering
they	-
are	-
applied	applied
to	-
the	-
words	words
in	-
the	-
texts	texts

Stopword Removal : Metode

- Algoritma **stoplist**
- Contoh stopwords adalah untuk, sang, sudah, adalah, dst.

Stopword Removal : Metode

Hasil Token	Hasil Filtering
namanya	namanya
adalah	-
santiago	santiago
santiago	santiago
sudah	-
memutuskan	memutuskan
untuk	-
mencari	mencari
sang	-
alkemis	alkemis

Stopword Removal : Metode

- Algoritma **wordlist**
- **Wordlist** adalah **kata-kata yang deskriptif (*penting*)** yang harus disimpan dan tidak dibuang dengan pendekatan *bag-of-words*.

Stopword Removal : Metode

- Algoritma **wordlist**
- Kita memiliki database kumpulan **kata-kata yang deskriptif (*penting*),** kemudian kalau hasil tokenisasi itu ada yang merupakan kata penting dalam database tersebut, maka hasil tokenisasi itu disimpan.

Stopword Removal : Metode

- Algoritma **wordlist**
- Contoh wordlist adalah applied, words, texts, dst.

Stopword Removal : Metode

Hasil Token	Hasil Filtering
they	-
are	-
applied	applied
to	-
the	-
words	words
in	-
the	-
texts	texts

Stopword Removal : Metode

- Algoritma **wordlist**
- Contoh wordlist adalah santiago, namanya, mencari, memutuskan, alkemis, dst.

Stopword Removal : Metode

Hasil Token	Hasil Filtering
namanya	namanya
adalah	-
santiago	santiago
santiago	santiago
sudah	-
memutuskan	memutuskan
untuk	-
mencari	mencari
sang	-
alkemis	alkemis

Using Stop Words or Not?

- Kebanyakan aplikasi text mining ataupun IR **bisa ditingkatkan** performanya dengan penghilangan stopword.
- Akan tetapi, secara umum Web search engines seperti ***google*** sebenarnya **tidak menghilangkan** stop word, karena algoritma yang mereka gunakan berhasil memanfaatkan stopword dengan baik.

Langkah 4 : *Phrase Detection*

Langkah ini bisa menangkap informasi dalam teks **melebihi** kemampuan dari metode tokenisasi / bag-of-word murni.

Langkah 4 : *Phrase Detection*

Pada langkah ini tidak hanya dilakukan tokenisasi per kata, namun juga mendeteksi adanya 2 kata atau lebih yang menjadi **frase**.

Langkah 4 : *Phrase Detection*

- Contoh, dari dokumen ini : *“search engines are the most visible information retrieval applications”*
- Terdapat dua buah **frase**, yaitu *“search engines”* dan *“information retrieval”*.

Langkah 4 : *Phrase Detection*

- *Phrase detection* bisa dilakukan dengan beberapa cara : menggunakan **rule/aturan** (misal dengan menganggap dua kata yang sering muncul berurutan sebagai frase), bisa dengan ***syntactic analysis***, and **kombinasi** keduanya.

Langkah 4 : *Phrase Detection*

- Metode umum yang diguakan adalah **penggunaan thesauri** untuk mendeteksi adanya frase.
- Contoh : Pada thesauri tersebut terdapat **daftar frase-fase** dalam bahasa tertentu, kemudia kita bandingkan kata-kata dalam teks apakah mengandung frase-frase dalam thesauri tersebut atau tidak.

Langkah 4 : *Phrase Detection*

- Kelemahanya, tahap ini butuh komputasi yang cukup lama
- Kebanyakan aplikasi teks mining atau IR tidak menggunakan *Phrase Detection*
- Sudah cukup dengan Token per Kata
- Akan tetapi, sebenarnya pemanfaatan *Phrase* akan meningkatkan akurasi

Langkah 5 : Stemming

Stemming adalah proses pengubahan **bentuk kata** menjadi **kata dasar** atau tahap mencari root kata dari tiap kata hasil filtering.

Langkah 5 : Stemming

Dengan dilakukanya proses stemming setiap kata berimbuhan akan berubah menjadi kata dasar, dengan demikian dapat lebih **mengoptimalkan** proses **teks mining**.

Langkah 5 : Stemming

Hasil Token	Hasil Filtering	Hasil Stemming
they	-	-
are	-	-
applied	applied	apply
to	-	-
the	-	-
words	words	word
in	-	-
the	-	-
texts	texts	text

Langkah 5 : Stemming

Hasil Token	Hasil Filtering	Hasil Stemming
namanya	namanya	nama
adalah	-	-
santiago	santiago	santiago
santiago	santiago	santiago
sudah	-	-
memutuskan	memutuskan	putus
untuk	-	-
mencari	mencari	cari
sang	-	-
alkemis	alkemis	alkemis

Langkah 5 : Stemming

- Implementasi proses **stemming** sangat beragam , tergantung dengan **bahasa** dari dokumen.
- Beberapa metode untuk Stemming :

Porter Stemmer (English & Indonesia)

Stemming Arifin-Setiono (Indonesia)

Stemming Nazief-Adriani (Indonesia)

Khoja (Arabic)

Stemming : Metode

- Algorithmic: Membuat sebuah **algoritma yang mendeteksi imbuhan**. Jika ada awalan atau akhiran yang seperti imbuhan, maka akan dibuang.

Stemming : Metode

- Algorithmic

Porter's algorithm

Rule			Example		
SSES	→	SS	caresses	→	caress
IES	→	I	ponies	→	poni
SS	→	SS	caress	→	caress
S	→		cats	→	cat

Stemming : Metode

- Metode Algorithmic
- Kelebihan : relatif **cepat**
- Kekurangan : beberapa algoritma **terkadang salah mendeteksi imbuhan**, sehingga ada beberapa kata yang bukan imbuhan tapi dihilangkan
- Contoh : makan -> mak; **an** dideteksi sebagai akhiran sehingga dibuang.

Stemming : Metode

- Metode Lemmatization
- Lemmatization : Stemming berdasarkan **kamus**
- Menggunakan *vocabulary* dan *morphological analysis* dari kata untuk menghilangkan imbuhan dan dikembalikan ke bentuk dasar dari kata.

Stemming : Metode

- Metode Lemmatization
- Stemming ini bagus untuk kata-kata yang mengalami **perubahan tidak beraturan** (terutama dalam english)
- Contoh : “see” -> “see”, “saw”, atau “seen”
- Jika ada kata “see”, “saw”, atau “seen”, bisa dikembalikan ke bentuk aslinya yaitu “see”

Stemming : Metode

- Algoritma Porter Stemming merupakan algoritma yang paling populer. Ditemukan oleh Martin Porter pada tahun 1980.
- Mekanisme algoritma tersebut dalam mencari kata dasar suatu kata berimbuhan, yaitu dengan membuang imbuhan–imbuhan (atau lebih tepatnya akhiran pada kata–kata bahasa Inggris karena dalam bahasa Inggris tidak mengenal awalan).

Langkah 5 : Stemming

Hasil Token	Hasil Filtering	Hasil Stemming	Type	Term
they	-	-	-	-
are	-	-	-	-
applied	applied	apply	apply	apply
to	-	-	-	-
the	-	-	-	-
words	words	word	word	word
in	-	-	-	-
the	-	-	-	-
texts	texts	text	text	text

Langkah 5 : Stemming

Hasil Token	Hasil Filtering	Hasil Stemming	Type	Term
namanya	namanya	nama	nama	nama
adalah	-	-	-	-
santiago	santiago	santiago	santiago	santiago
santiago	santiago	santiago	-	-
sudah	-	-	-	-
memutuskan	memutuskan	putus	putus	putus
untuk	-	-	-	-
mencari	mencari	cari	cari	cari
sang	-	-	-	-
alkemis	alkemis	alkemis	alkemis	alkemis

Studi Kasus

Dokumen Ke-i	Isi Dokumen
1	pembukaan daftar wisuda dan pelaksanaan nya lebih baik d umumkan di web tidak hanya di fakultas. sehingga memudahkan mahasiswa yang ada di luar kota. pelaksanaan wisuda sebaiknya terjadwal tidak tergantung pada kuota. sehingga lebih cepat mendapat ijazah.
2	dalam setahun belakangan ini, pengaksesan KRS diganti ke SIAM (sebelumnya menggunakan SINERGI). saat menggunakan sinergi, fitur serta kecepatan akses sangat handal dan nyaman. tapi setelah diganti menggunakan SIAM, keadaan berbalik menjadi buruk (lambat loading dan bahkan sampai logout dengan sendirinya). *KRS tidak hanya berpengaruh bagi mahasiswa semester muda tapi juga keseluruhan mahasiswa
3	Assalamualaikum Wr. Wb. yang menjadi salah satu syarat untuk bisa ujian kompre ada sertifikat TOEIC, sehingga jika belum lulus toeic maka tidak bisa melakukan ujian kompre. saya rasa ini sangat menghambat teman-teman yang memang lemah dibidang bahasa inggris (atau yang kurang beruntung dalam ujian toeic-nya). sehingga mereka tidak bisa fokus untuk ujian kompre-nya. terima kasih..
4	pak/bu dosen saya mau minta keringanan biaya proposional dan spp ,soalnya ibu saya keberatan dengan biaya itu? terima kasih atas perhatiannya.

Studi Kasus

Dokumen Ke-i	Isi Dokumen	Tokenisasi	Filtering	Stemming
1	pembukaan daftar wisuda dan pelaksanaan nya lebih baik d umumkan di web ub tidak hanya di fakultas. sehingga memudahkan mahasiswa yang ada di luar kota. pelaksanaan wisuda sebaiknya terjadwal tidak tergantung pada kuota. sehingga lebih cepat mendapat ijazah.	pembukaan daftar wisuda dan pelaksanaan nya lebih baik d umumkan di web tidak hanya di fakultas sehingga memudahkan mahasiswa yang ada di luar kota pelaksanaan wisuda sebaiknya terjadwal tidak tergantung pada kuota sehingga lebih cepat mendapat ijazah	pembukaan daftar wisuda pelaksanaan umumkan web fakultas memudahkan mahasiswa kota pelaksanaan wisuda sebaiknya terjadwal tergantung kuota cepat ijazah	buka daftar wisuda laksana umum web fakultas mudah mahasiswa kota laksana wisuda baik jadwal gantung kuota cepat ijazah
2	dalam setahun belakangan ini, pengaksesan KRS diganti ke SIAM (sebelumnya menggunakan SINERGI). saat menggunakan sinergi, fitur serta kecepatan akses sangat handal dan nyaman. tapi setelah diganti menggunakan SIAM, keadaan berbalik menjadi buruk (lambat loading dan bahkan sampai logout dengan sendirinya). *KRS tidak hanya berpengaruh bagi mahasiswa semester muda tapi juga keseluruhan mahasiswa	dalam setahun belakangan ini pengaksesan krs diganti ke siam sebelumnya menggunakan sinergi saat menggunakan sinergi fitur serta kecepatan akses sangat handal dan nyaman tapi setelah diganti menggunakan siam keadaan berbalik menjadi buruk lambat loading dan bahkan sampai logout dengan sendirinya krs tidak hanya berpengaruh bagi mahasiswa semester muda tapi juga keseluruhan mahasiswa	setahun belakangan pengaksesan krs diganti siam sinergi sinergi fitur kecepatan akses handal nyaman diganti siam keadaan berbalik buruk lambat loading logout sendirinya krs berpengaruh mahasiswa semester muda keseluruhan mahasiswa	tahun belakang akses krs ganti siam sinergi sinergi fitur cepat akses handal nyaman ganti siam ada balik buruk lambat loading logout sendiri krs pengaruh mahasiswa semester muda luruh mahasiswa
3	Assalamualaikum Wr. Wb. yang menjadi salah satu syarat untuk bisa ujian kompre ada sertifikat TOEIC, sehingga jika belum lulus toeic maka tidak bisa melakukan ujian kompre. saya rasa ini sangat menghambat teman-teman yang memang lemah dibidang bahasa inggris (atau yang kurang beruntung dalam ujian toeic-nya). sehingga mereka tidak bisa fokus untuk ujian kompre-nya. terima kasih..	assalamualaikum wr wb yang menjadi salah satu syarat untuk bisa ujian kompre ada sertifikat toeic sehingga jika belum lulus toeic maka tidak bisa melakukan ujian kompre saya rasa ini sangat menghambat teman teman yang memang lemah dibidang bahasa inggris atau yang kurang beruntung dalam ujian toeic nya sehingga mereka tidak bisa fokus untuk ujian kompre nya terima kasih	assalamualaikum wr wb syarat ujian kompre sertifikat toeic lulus toeic ujian kompre menghambat lemah dibidang bahasa inggris kurang beruntung ujian toeic fokus ujian kompre terima kasih	assalamualaikum wr wb syarat uji kompre sertifikat toeic lulus toeic uji kompre hambat lemah bidang bahasa inggris kurang untung uji toeic fokus uji kompre terima kasih
4	pak/bu dosen saya mau minta keringanan biaya proposional dan spp ,soalnya ibu saya keberatan dengan biaya itu? terima kasih atas perhatiannya.	pak bu dosen saya mau minta keringanan biaya proposional dan spp soalnya ibu saya keberatan dengan biaya itu terima kasih atas perhatiannya	pak bu dosen minta keringanan biaya proposional spp soalnya ibu keberatan biaya terima kasih perhatiannya	pak bu dosen minta ringan biaya proposional spp soal ibu berat biaya terima kasih hati

Kuis (Latihan Soal)

Lakukan pemrosesan awal untuk dokumen dibawah ini:

https://github.com/feryandi/Dataset-Artikel

Presentasikan didepan kelas untuk kelas minggu depan (minimal 3 Mahasiswa)

Thanks!

Any questions ?