IMPLEMENTASI SUFFIX TREE CLUSTERING UNTUK PENGELOMPOKAN DOKUMEN HASIL PENCARIAN ONLINE PADA MESIN PENCARI DAN JEJARING SOSIAL

JAJANG MACHPUDIN SURYANA
10107277

Pencarian yang dilakukan tidak selamanya optimal

Solusi :

Clustering terhadap hasil pencarian



LATAR BELAKANG



Mengimplementasikan metode STC untuk pengelompokan dokumen hasil pencarian online

MAKSUD

Menganalisis metode *Suffix Tree Clustering* pada pengelompokan dokumen secara umum.

Mengimplementasikan metode *Suffix Tree Clustering* pada pengelompokan dokumen hasil pencarian online.

Untuk mendapatkan persamaan dan perbedaan pola dokumen, sehingga dapat diperoleh ringkasan yang berguna untuk pengelompokan.



TUJUAN

BATASAN MASALAH

❶Studi kasus dilakukan di mesin pencari Google dan jejaring sosial Twitter.

❷Sumber data berasal dari hasil pencarian dokumen berdasarkan suatu kata kunci.

❸Metode yang digunakan adalah metode *Suffix Tree Clustering*.



BATASAN MASALAH

❹*Clustering* data dilakukan pada halaman hasil yang telah ditelusuri oleh mesin pencari Google dan Twitter. Hasil *clustering* akan berada di samping isi halaman hasil pencarian.

❺Metode analisis dan pembangunan perangkat lunak yang digunakan adalah metode analisis terstruktur.



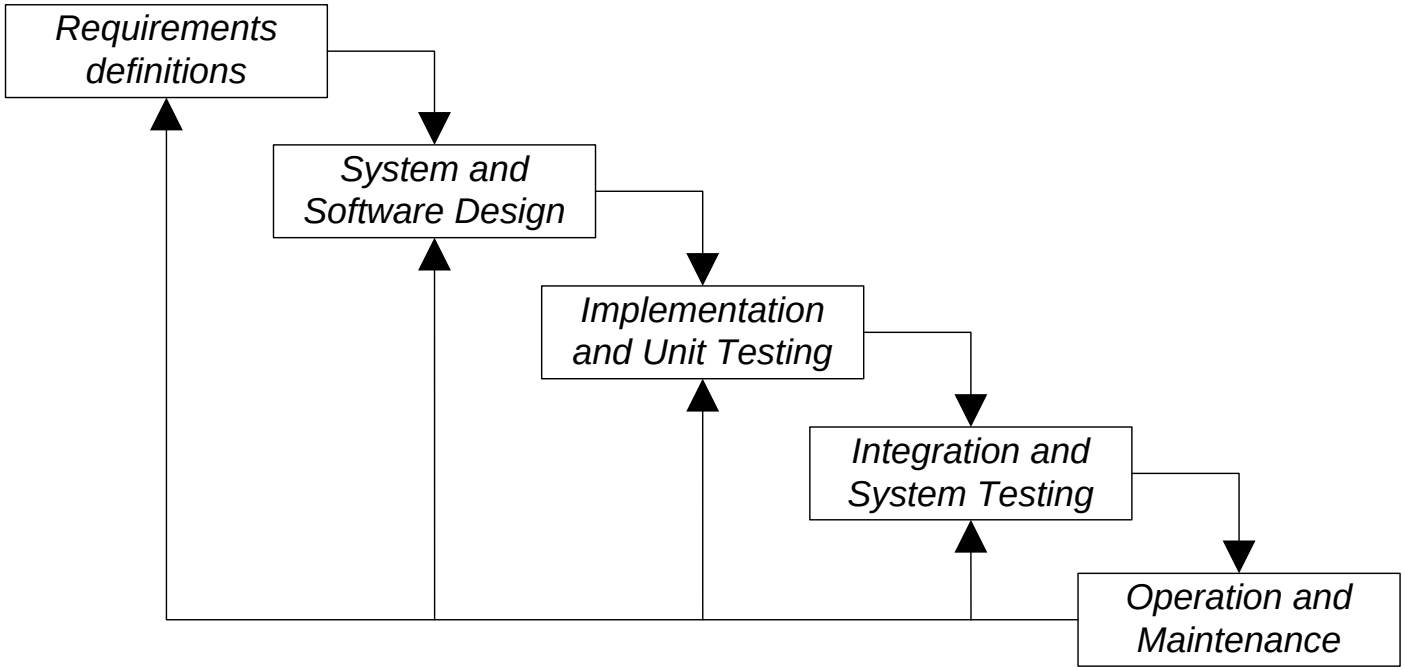
METODOLOGI PENGUMPULAN DATA

① Studi Literatur

② Observasi



METODE PEMBUATAN PERANGKAT LUNAK



Document preprocessing

- Penghapusan Stop-word
  - penghapusan tanda baca
  - penghapusan elemen html

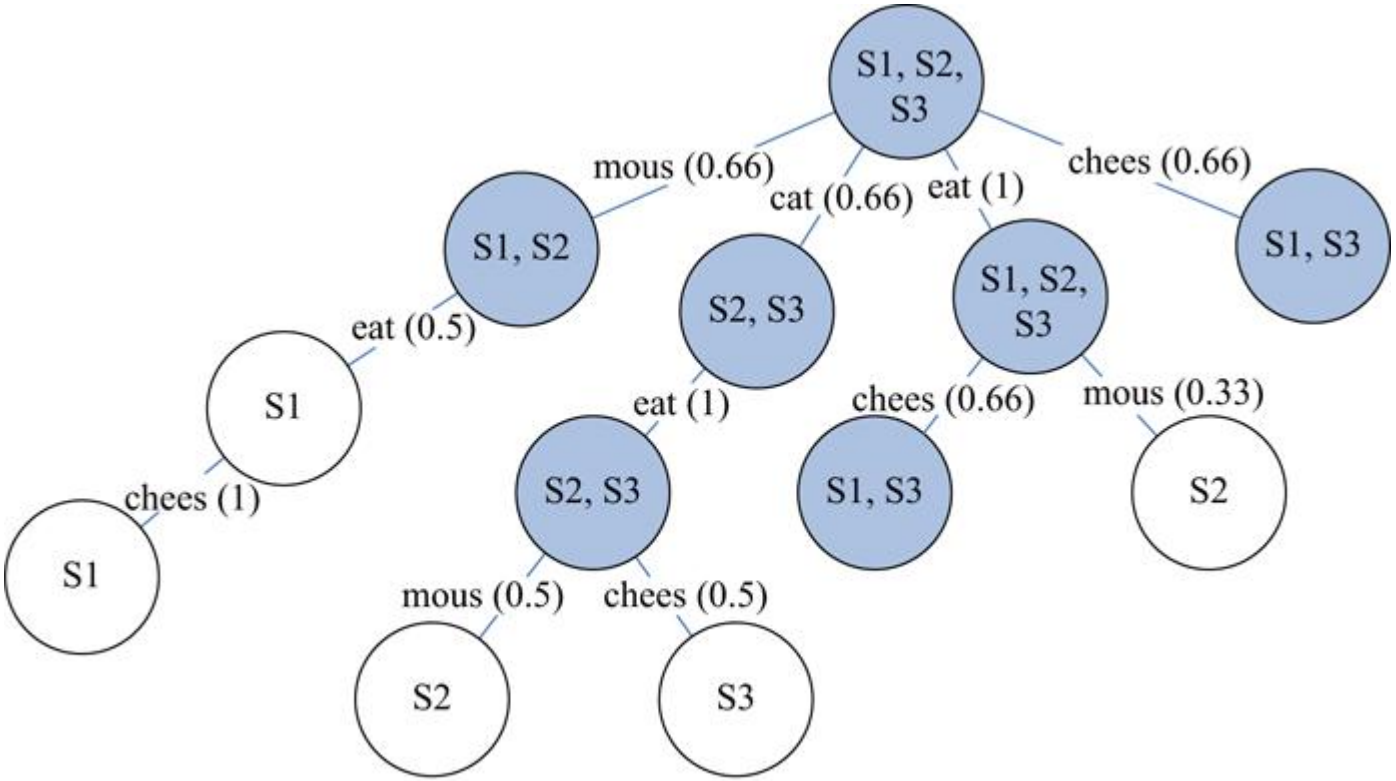
- Stemming kata
  - Information, informing → inform

ALGORITMA STC

Pembentukan cluster dasar

berdasarkan scoring dokumen

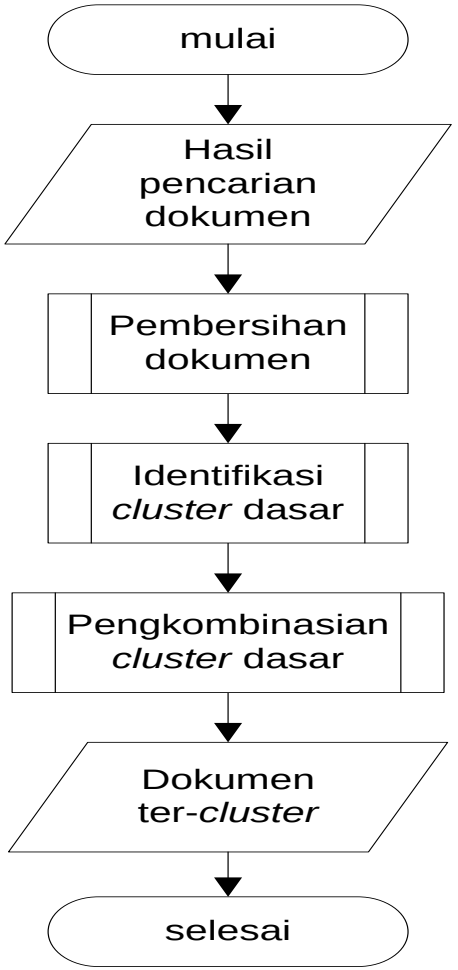
Pengkombinasian Cluster

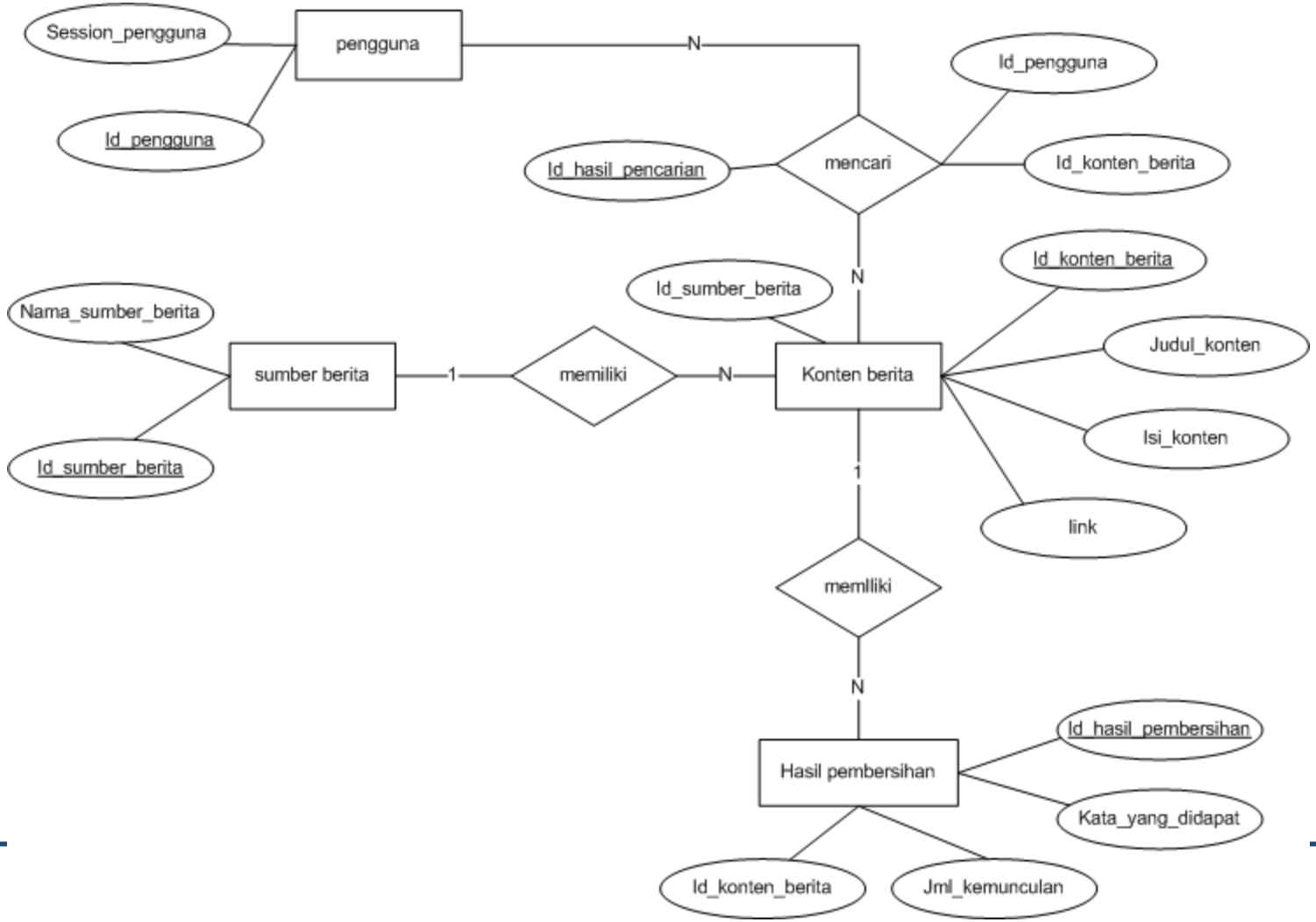


Note: Threshold = 0.6

ALGORITMA STC

ANALISIS ALGORITMA




Session_pengguna
pengguna
N
Id_pengguna
Id_pengguna
Id_hasil_pencarian
mencari
Id_konten_berita
N
Id_konten_berita
Id_sumber_berita
Nama_sumber_berita
sumber berita
1
memiliki
N
Konten berita
Judul_konten
Isi_konten
Id_sumber_berita
1
link
memIliki
N
Hasil pembersihan
Id_hasil_pembersihan
Kata_yang_didapat
Id_konten_berita
Jml_kemunculan

ERD

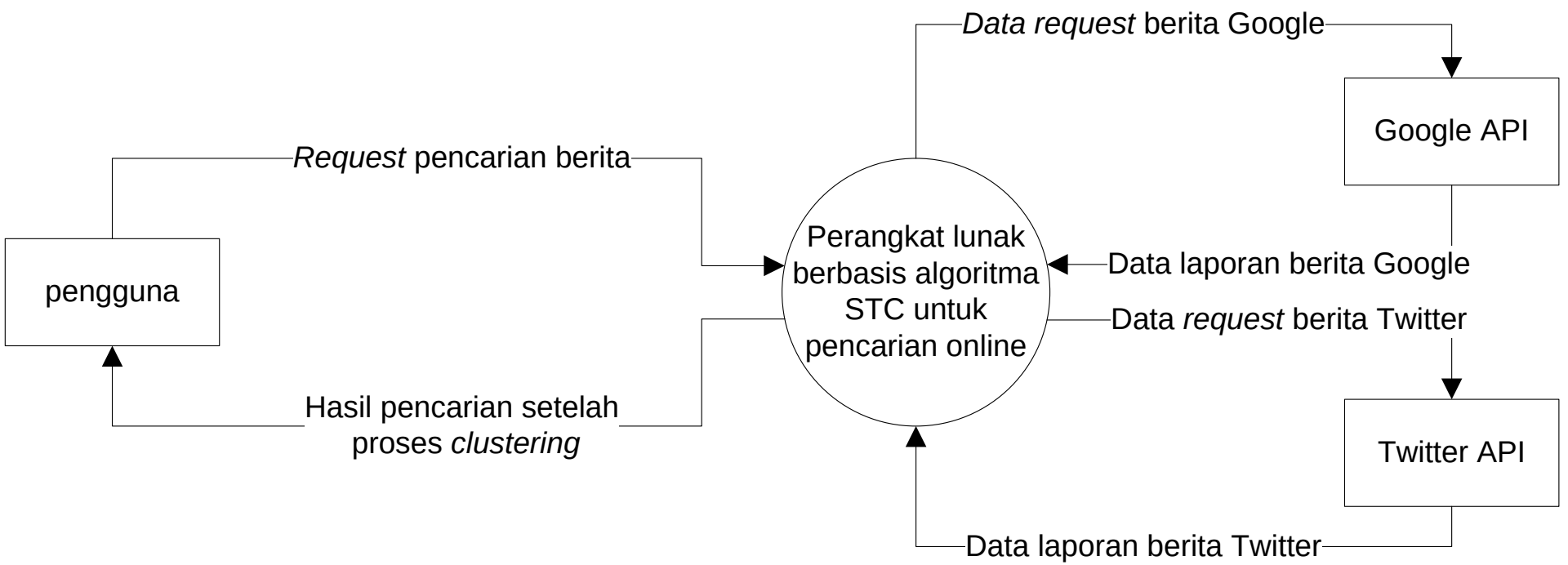
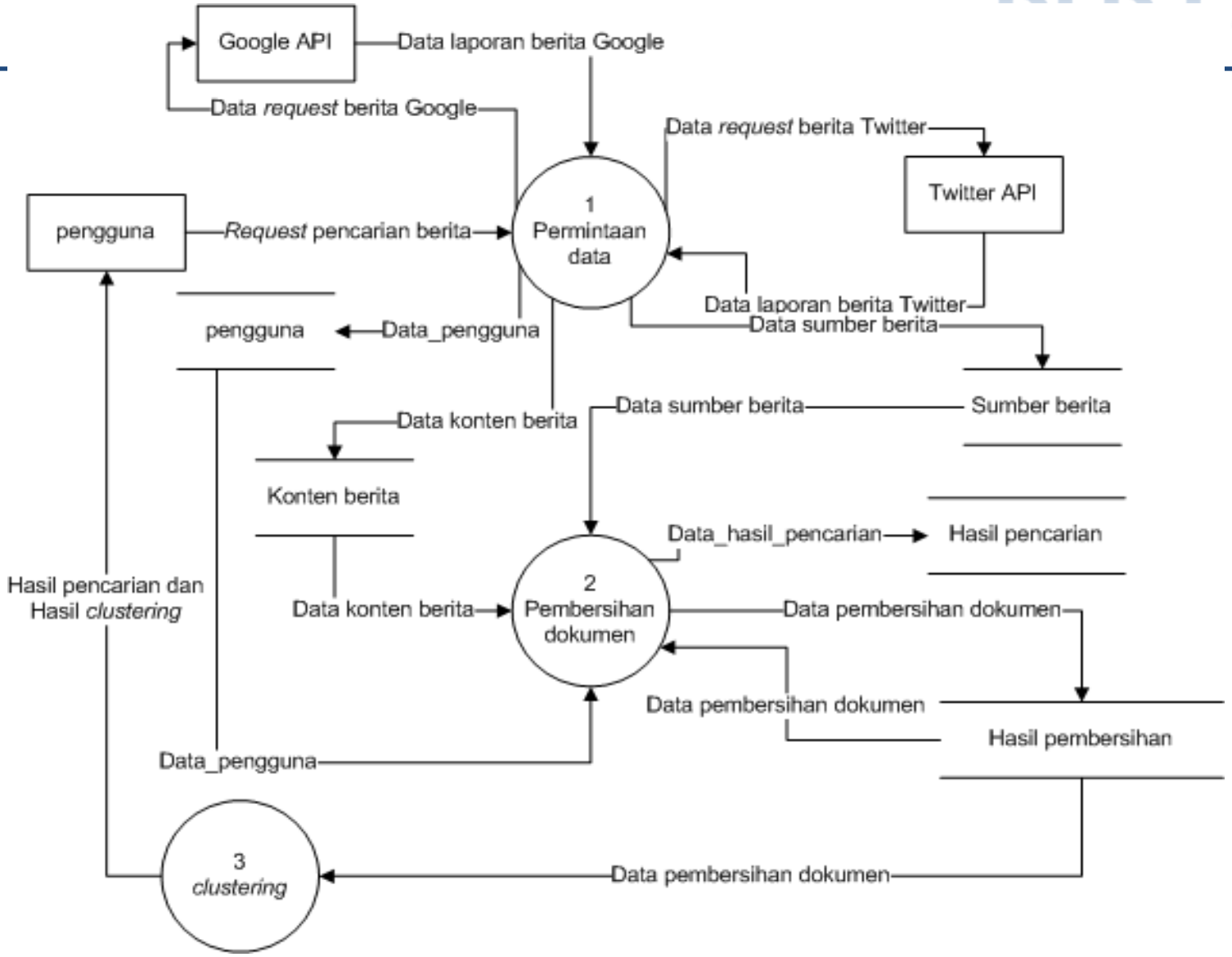
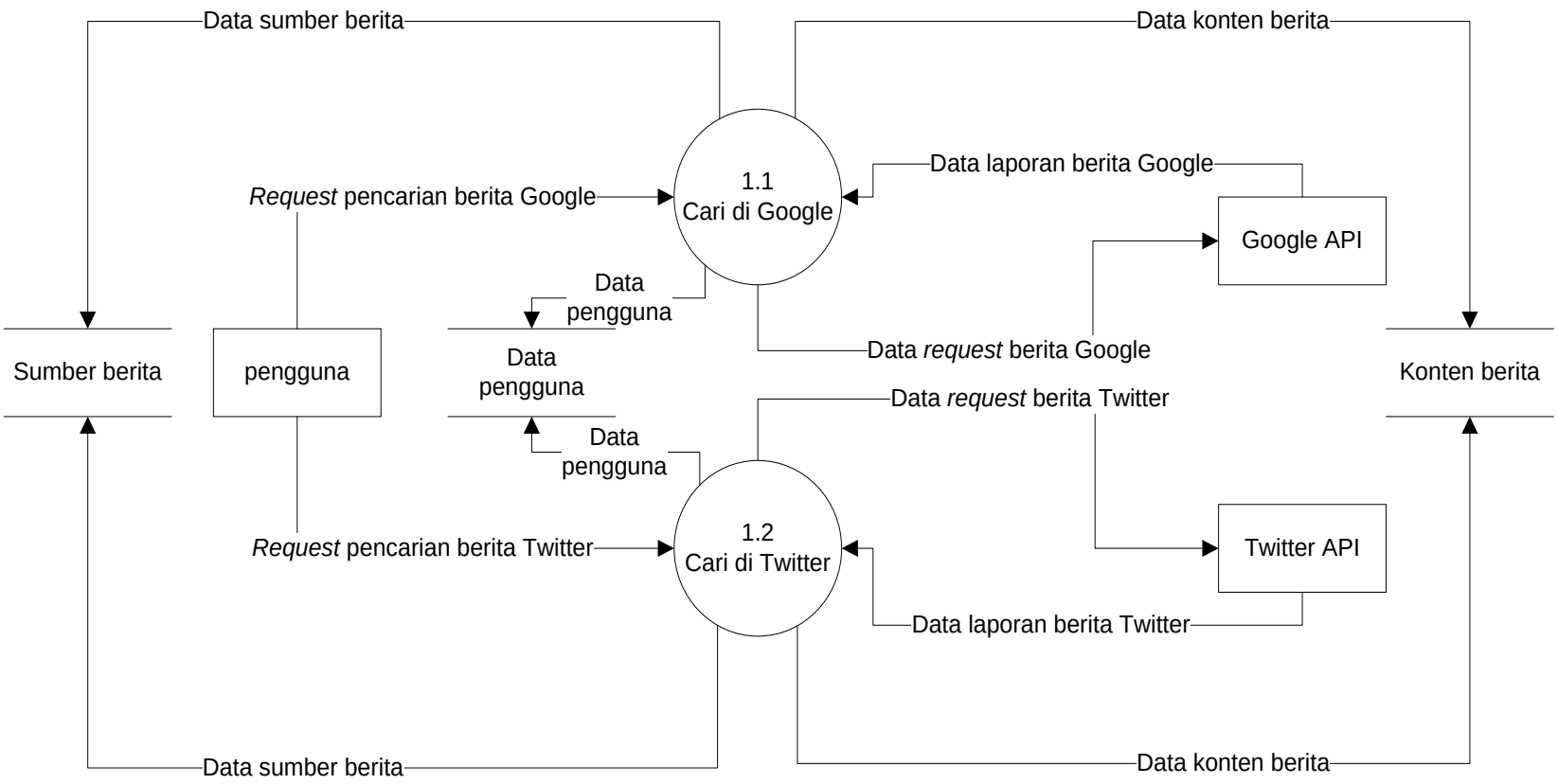


DIAGRAM KONTEKS

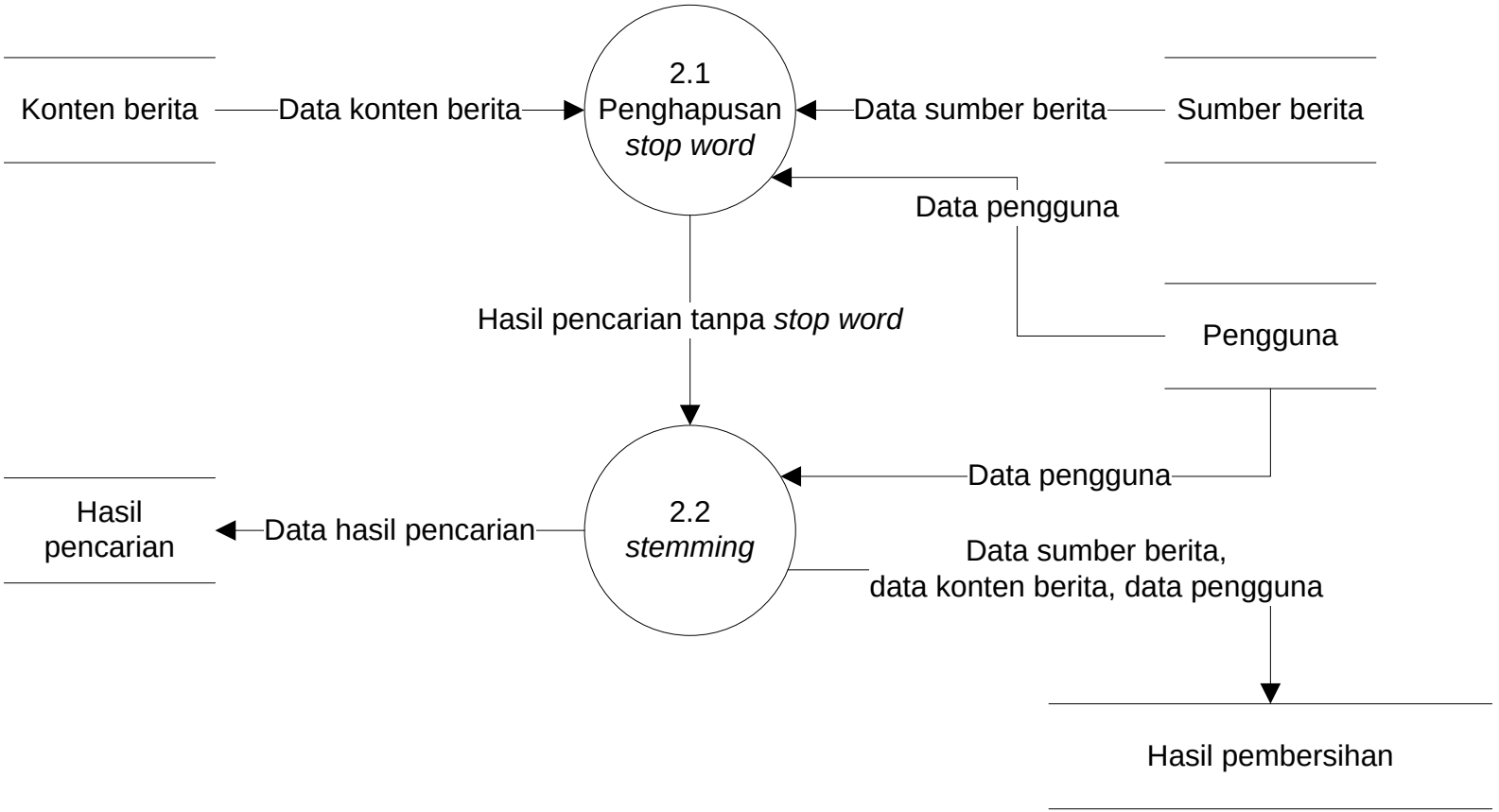
DFD LV.1


Google API
Data laporan berita Google
Data *request* berita Google
Data *request* berita Twitter
Twitter API
pengguna
Request pencarian berita
1
Permintaan
data
Data laporan berita Twitter
Data sumber berita
pengguna
Data_pengguna
Data konten berita
Data sumber berita
Sumber berita
Konten berita
Data_hasil_pencarian
Hasil pencarian
Hasil pencarian dan
Hasil *clustering*
Data konten berita
2
Pembersihan
dokumen
Data pembersihan dokumen
Data pembersihan dokumen
Hasil pembersihan
Data_pengguna
3
clustering
Data pembersihan dokumen

DFD LV.2 PROSES 1

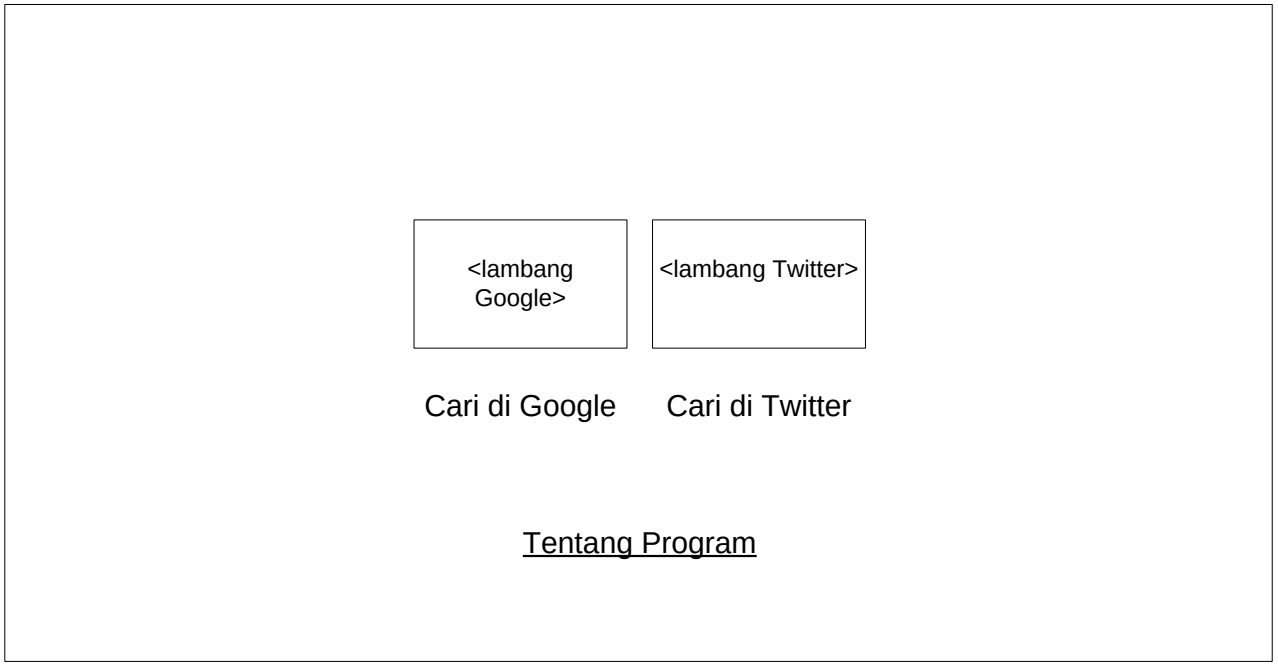


DFD LV.2 PROSES 2



ANTARMUKA HALAMAN UTAMA

A01

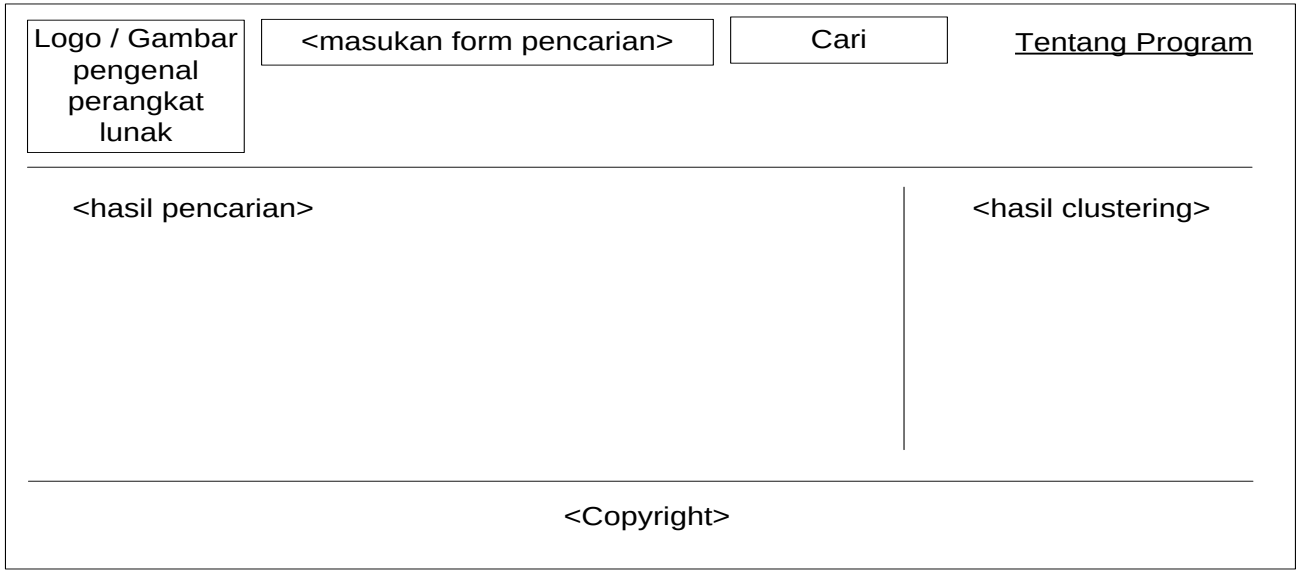


1. Klik “Cari di Google” untuk:
   A. Menuju A02 jika <form pencarian> telah terisi
   B. Menuju M01 jika <form pencarian> belum terisi
2. Klik “Cari di Twitter” untuk:
   A. Menuju A03 jika <form pencarian> telah terisi
   B. Menuju M01 jika <form pencarian> belum terisi
3. Klik “Tentang Program” untuk menuju A04

Ukuran halaman : 1024 x 768 px
Jenis tulisan : Arial
Ukuran tulisan : 12px
Warna latar : putih

ANTARMUKA PENCARIAN GOOGLE

A02



Ukuran halaman : 1024 x 768 px
Jenis tulisan : Arial
Ukuran tulisan : 12px
Warna latar : putih

1. Klik “Cari” untuk:
   A. Menuju A02 jika <form pencarian> telah terisi
   B. Menuju M01 jika <form pencarian> belum terisi
2. Klik “Tentang Program” untuk menuju A04
3. Klik Logo / Gambar pengenal perangkat lunak untuk menuju ke A01

ANTARMUKA PENCARIAN TWITTER

A02

Logo / Gambar pengenal perangkat lunak | <masukan form pencarian> | Cari | Tentang Program

<hasil pencarian>

<hasil clustering>

<Copyright>

Ukuran halaman : 1024 x 768 px
Jenis tulisan : Arial
Ukuran tulisan : 12px
Warna latar : putih

1. Klik “Cari” untuk:
   A. Menuju A02 jika <form pencarian> telah terisi
   B. Menuju M01 jika <form pencarian> belum terisi
2. Klik “Tentang Program” untuk menuju A04
3. Klik Logo / Gambar pengenal perangkat lunak untuk menuju ke A01

ANTARMUKA HASIL CLUSTERING

A05

Logo / Gambar pengenal perangkat lunak

<masukan form pencarian>

Tentang Program

Cari di Google | Cari di Twitter

Hasil clustering dengan kata kunci <kata_kunci>

<hasil pencarian>

1. Klik “Tentang Program” untuk menuju A04
2. Klik Logo / Gambar pengenal perangkat lunak untuk menuju ke A01

Ukuran halaman : 1024 x 768 px
Jenis tulisan : Arial
Ukuran tulisan : 12px
Warna latar : putih

ANTARMUKA RELASI CLUSTER

A06

Logo / Gambar pengenal perangkat lunak

<masukan form pencarian>

Tentang Program

Cari di Google | Cari di Twitter

Hasil clustering dengan kata kunci <kata_kunci>

<cluster 1> dan <cluster 2> berelasi / tidak berelasi kembali

1. Klik Logo / Gambar pengenal perangkat lunak untuk menuju ke A01
2. Klik “Tentang Program” untuk menuju A04
3. Klik “kembali” untuk menuju ke halaman sebelumnya

Ukuran halaman : 1024 x 768 px
Jenis tulisan : Arial
Ukuran tulisan : 12px
Warna latar : putih

ANTARMUKA TENTANG PROGRAM

A04

<informasi tentang program
dan pembuat program>

Kembali ke halaman sebelumnya | Kembali ke halaman utama

1. Klik “Kembali ke halaman sebelumnya” kembali ke halaman tempat tombol “Tentang Program” ditekan
2. Klik “Kembali ke halaman utama” untuk menuju A01

Ukuran halaman : 1024 x 768 px
Jenis tulisan : Arial
Ukuran tulisan : 12px
Warna latar : putih

HASIL PENGUJIAN (WHITE BOX)

Penghapusan stop word

Kompleksitas Algoritma : 2

Stemming kata

Kompleksitas Algoritma : 10

Identifikasi cluster

Kompleksitas Algoritma : 2

HASIL PENGUJIAN (BLACK BOX)

Kasus dan Hasil Uji (Data Normal)			
Data Masukan	**Yang Diharapkan**	**Pengamatan**	**Kesimpulan**
Form pencarian diisi dengan : chelsea	Data hasil pencarian berdasarkan kata kunci "chelsea" muncul	Muncul hasil pencarian berdasarkan kata kunci "chelsea".	[✓] Diterima [ ] Ditolak
Kasus dan Hasil Uji (Data Normal)			
Data Masukan	**Yang Diharapkan**	**Pengamatan**	**Kesimpulan**
Form pencarian diisi dengan : 12	Data hasil pencarian berdasarkan kata kunci "12" muncul	Muncul hasil pencarian berdasarkan kata kunci "12".	[✓] Diterima [ ] Ditolak
Kasus dan Hasil Uji (Data Normal)			
Data Masukan	**Yang Diharapkan**	**Pengamatan**	**Kesimpulan**
Form pencarian diisi dengan : ?	Data hasil pencarian berdasarkan kata kunci "?" muncul	Muncul hasil pencarian berdasarkan kata kunci "?".	[✓] Diterima [ ] Ditolak
Kasus dan Hasil Uji (Data Kosong)			
Data Masukan	**Yang Diharapkan**	**Pengamatan**	**Kesimpulan**
Form pencarian kosong	Muncul pesan kesalahan "Form pencarian belum terisi. Silahkan isi dahulu sebelum melakukan pencarian."	Dapat menampilkan pesan kesalahan "Form pencarian belum terisi. Silahkan isi dahulu sebelum melakukan pencarian."	[✓] Diterima [ ] Ditolak

HASIL PENGUJIAN

1. Apakah Anda tidak memiliki kesulitan dalam menggunakan *website* searchjaka.com?

A. Sangat Setuju
B. Setuju
C. Cukup Setuju
D. Kurang Setuju
E. Tidak Setuju

Pertanyaan	No.	Keterangan	Responden	Presentase (%)
1	1	Sangat setuju	3	20
	2	Setuju	9	60
	3	Cukup setuju	3	20
	4	Kurang setuju	0	0
	5	Tidak setuju	0	0
Jumlah			**15**	**100**

HASIL PENGUJIAN

2. Apakah Anda setuju website searchjaka.com memiliki navigasi yang mudah?

A. Sangat Setuju
B. Setuju
C. Cukup Setuju
D. Kurang Setuju
E. Tidak Setuju

Pertanyaan	No.	Keterangan	Responden	Presentase (%)
2	1	Sangat setuju	4	26,67
	2	Setuju	10	66,67
	3	Cukup setuju	1	6,67
	4	Kurang setuju	0	0
	5	Tidak setuju	0	0
Jumlah			**15**	**100**

HASIL PENGUJIAN

3. Apakah Anda setuju bahwa *website* searchjaka.com ini menampilkan informasi hasil pencarian yang sama seperti Anda melakukan pencarian di Google atau Twitter?

A. Sangat Setuju
B. Setuju
C. Cukup Setuju
D. Kurang Setuju
E. Tidak Setuju

Pertanyaan	No.	Keterangan	Responden	Presentase (%)
3	1	Sangat setuju	1	6,67
	2	Setuju	7	46,67
	3	Cukup setuju	7	46,67
	4	Kurang setuju	0	0
	5	Tidak setuju	0	0
Jumlah			**15**	**100**

HASIL PENGUJIAN

4. Apakah Anda setuju dengan adanya *website* searchjaka.com dapat membantu Anda untuk menemukan hasil pencarian dengan lebih cepat?

A. Sangat Setuju
B. Setuju
C. Cukup Setuju
D. Kurang Setuju
E. Tidak Setuju

Pertanyaan	No.	Keterangan	Responden	Presentase (%)
4	1	Sangat setuju	0	0
	2	Setuju	0	0
	3	Cukup setuju	5	33,33
	4	Kurang setuju	10	66,67
	5	Tidak setuju	0	0
Jumlah			**15**	**100**

HASIL PENGUJIAN

5. Apakah Anda setuju bahwa *website* searchjaka.com memiliki poin-poin pada bagian hasil *clustering* yang sudah cukup mewakili hasil pencarian secara keseluruhan?

A. Sangat Setuju
B. Setuju
C. Cukup Setuju
D. Kurang Setuju
E. Tidak Setuju

Pertanyaan	No.	Keterangan	Responden	Presentase (%)
5	1	Sangat setuju	0	0
	2	Setuju	1	6,67
	3	Cukup setuju	11	73,33
	4	Kurang setuju	3	20
	5	Tidak setuju	0	0
Jumlah			**15**	**100**

KESIMPULAN

Algoritma STC dapat digunakan untuk melakukan pengelompokan terhadap dokumen hasil pencarian secara langsung

Performansi yang dihasilkan sangat cepat untuk set data 10 sampai dengan 15 buah dalam 1 kali hasil pencarian

Cluster yang terbentuk sudah dapat mewakili topik dari halaman pencarian

Hasil pencarian dapat diketahui dengan cepat oleh pengguna

SARAN

Algoritma stemming yang lebih baik dapat menghasilkan cluster yang lebih baik

Dapat menangani lebih dari satu bahasa untuk pengelompokan hasil pencarian

DEMO PROGRAM