UIN SUSKA RIAU
Hak Cipta Dilindungi Undang-Undang
1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:
a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.
b. Pengutipan tidak merugikan kepentingan yang wajar UIN Suska Riau.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin UIN Suska Riau.
© Hak Cipta milik UIN Suska Riau
State Islamic University of Sultan Syarif Kasim Riau

BAB II
LANDASAN TEORI

2.1 *Information Retrieval System*

Sistem temu kembali informasi (*information retrieval system*) merupakan sistem yang dapat digunakan untuk menemukan informasi yang *relevan* dengan kebutuhan dari penggunanya secara otomatis dari suatu koleksi informasi (Mandala, 2002). Sistem ini berfungsi untuk menemukan informasi yang *relevan* dengan kebutuhan pemakai. Dalam konteks ini, temu kembali informasi berkaitan dengan representasi, penyimpanan, dan akses terhadap dokumen representasi dokumen. (Harjanto, 2012).

Suatu *query* merupakan *keyword* (kata kunci) yang kemudian dilakukan pencocokan dengan *keyword* yang dimiliki oleh suatu dokumen sehingga akhinya dapat menarik kembali informasi dari suatu dokumen tersebut.

Proses yang berjalan dalam sistem temu kembali informasi adalah proses *indexing subsystem*, yang merupakan proses persiapan ulang dilakukan terhadap dokumen sehingga dokumen siap diproses, dan *searching subsystem* (*macthing system*) yang merupakan proses menemukan kembali informasi (dokumen) yang *relevan* terhadap *query* yang diberikan. (Harjanto, 2012).

Adapun tahap-tahap yang terjadi pada proses *indexing*, yaitu (Harjanto, 2012):

1. *Tokenizing* dokumen, yaitu 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.
2. *Stopword removal* dokumen, yaitu kata-kata yang sering muncul dalam dokumen namun artinya tidak deskriptif dan tidak memiliki keterkaitan dengan tema tertentu. Pada bahasa Indonesia, *stopword* disebut juga sebagai kata yang tidak penting, misalnya “di”, “oleh”, “pada”, “sebuah”, “karena” dan lain sebagainya. Untuk lebih lanjut, daftar *stopword* bisa dilihat pada lampiran A.

Hak Cipta Dilindungi Undang-Undang
1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:
a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.
b. Pengutipan tidak merugikan kepentingan yang wajar UIN Suska Riau.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin UIN Suska Riau.
© Hak cipta milik UIN Suska Riau
State Islamic University of Sultan Syarif Kasim Riau
UIN SUSKA RIAU

3. *Stemming* dokumen, yaitu tahap penghilangan imbuhan sehingga didapatkan kata dasar dari *term-term* dokumen inputan.
4. *Term Weighting*, yaitu proses pembobotan pada setiap term (kata) yang ada didalam dokumen.

2.2 *Text Preprocessing*

Text Preprocessing adalah mempersiapkan teks menjadi data yang akan mengalami proses pengolahan pada tahapan berikutnya (Mustaqhfiri, 2011).

Tujuan dilakukan *pre-processing* adalah memilih setiap kata dari dokumen dan merubahnya menjadi kata dasar yang memiliki arti sempit dan proses teks mining akan memberikan hasil yang lebih memuaskan (Septiawan, 2010).

Adapun tahapan pada teks *preprocessing* adalah pemecahan kalimat, proses *case folding*, *filtering* kalimat, proses *tokenizing* kata dan proses *stemming* (Mustaqhfiri, 2011). Berikut adalah diagram alir dari *text preprocessing*:

Gambar 2.1. Diagram Alir Text Preprocessing

2.2.1 Pemecahan Kalimat

Pemecahan kalimat teks menjadi kalimat-kalimat. Adapun yang menjadi pemisah kumpulan kalimat adalah tanda tanya “?”, tanda titik ”.”, dan tanda seru “!” (Mustaqhfiri, 2011).

2.2.2 *Case Folding*

Case Folding adalah proses mengubah semua huruf yang ada pada dokumen teks menjadi huruf kecil semua. Dalam bahasa pemrograman *PHP* digunakan fungsi *strtolower* untuk mengubah huruf menjadi huruf kecil semua.

2.2.3 *Filtering* Kalimat

Filtering merupakan proses penghilangan *stopword*. *Stopword* yaitu kata-kata yang sering muncul dalam dokumen namun artinya tidak deskriptif dan tidak memiliki keterkaitan dengan tema tertentu. Misal, “di”,”oleh”,”karena”, dan lain sebagainya.

Proses ini dilakukan pada judul dokumen, abstrak dokumen dan masukan *query* secara terpisah. Proses ini lebih mudah dan lebih cepat diproses setelah kata diekstrak dari teks dokumennya. Kata yang diperoleh dari tahap *filtering* diperiksa dengan daftar *stopword*, apabila sebuah kata masuk di dalam daftar *stopword* maka kata tersebut tidak akan diproses lebih lanjut (Utomo, 2011).

2.2.4 Tokenisasi *Kata*

Tokenizing adalah proses pemotongan *string input* berdasarkan tiap kata yang menyusunnya. Pemecahan kalimat menjadi kata-kata tunggal dilakukan dengan men-*scan* kalimat dengan pemisah *white space* (spasi, tab, dan *newline*) (Mustaqhfiri, 2011).

Proses ini dilakukan pada judul dokumen, abstrak dokumen dan masukan *query* secara terpisah (Utomo, 2011).

2.2.5 *Stemming* Bahasa Indonesia

Stemming merupakan suatu proses yang terdapat dalam sistem IR yang mentransformasikan kata-kata yang terdapat dalam suatu dokumen ke kata-kata akarnya (*root word*) dengan menggunakan aturan-aturan tertentu. Sebagai contoh, kata bersama, kebersamaan, menyamai, akan distem ke *root word*nya yaitu "sama" (Septiawan, 2010).

Terdapat 5 aturan tahapan pada proses *stemming* dalam bahasa indonesia menggunakan *Porter Stemmer* (Tala, 2003), yaitu sebagai berikut:

1. Penanganan terhadap partikel infleksional, yaitu : lah, tah, kah. Contoh: pergilah, berlarilah.
2. Penanganan terhadap kata ganti infleksional, yaitu : ku, mu, dan nya. Contoh: punyaku, miliknya.
3. Penanganan terhadap prefiks derivasional pertama, yaitu meng dan semua variasinya, peng, dan semua variasinya, di, ter, dan ke. Contoh: mengubur, pengukur, terlambat.
4. Penanganan terhadap prefiks derivasional kedua, yaitu: ber dan semua variasinya, per dan semua variasinya. Contoh: berlatih, belajar, perkata
5. Penanganan terhadap surfiks derivasional yaitu kan, an, dan i. Contoh: ambilkan, janjian, dan dekati.

UIN SUSKA RIAU
Hak Cipta Dilindungi Undang-Undang
1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:
a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.
b. Pengutipan tidak merugikan kepentingan yang wajar UIN Suska Riau.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin UIN Suska Riau.
© Hak cipta milik UIN Suska Riau
State Islamic University of Sultan Syarif Kasim Riau

Kelima proses *stemming* diatas dapat dirumuskan kedalam beberapa tabel sebagai berikut :

Tabel 2.1. Aturan Pembentukan Partikel *Infleksional*

Surfiks	Pengganti	Kondisi Tambahan	Contoh
Kah	Null	Null	Bukukah→buku
Lah	Null	Null	Terimalah→ terima
Tah	Null	Null	Apatah → apa
Pun	Null	Null	Bukupun→ buku

Tabel 2.2. Aturan Pembentukan Kata Ganti Milik *Infleksional*

Surfiks	Pengganti	Kondisi Tambahan	Contoh
Ku	Null	Null	Bukuku → buku
Mu	Null	Null	Bukumu → buku
Nya	Null	Null	Bukunya→ buku

Tabel 2.3. Aturan Pembentukan *Prefiks Devirasional* Pertama

Surfiks	Pengganti	Kondisi Tambahan	Contoh
Meng	Null	Null	Mengukur→ ukur
Meny	S	V....	Menyapu → sapu
Men	Null	Null	Menduga→ duga
Men	T	V...	Menuduh→ tuduh
Mem	P	V...	Memilah→ pilah
Mem	Null	Null	Membaca→ baca
Me	Null	Null	Merusak → rusak
Peng	Null	Null	Pengukur → ukur
Peny	S	V...	Penyelam→ selam
Pen	Null	Null	Pendaki → daki
Pen	T	V...	Penari → tari
Pem	Null	Null	Pembaca→ baca
Di	Null	Null	Diukur →ukur
Ter	Null	Null	Tersipu → sipu
Ke	Null	Null	Kekasih → kasih

UIN SUSKA RIAU
Hak Cipta Dilindungi Undang-Undang
1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:
a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.
b. Pengutipan tidak merugikan kepentingan yang wajar UIN Suska Riau.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin UIN Suska Riau.
© Hak cipta milik UIN Suska Riau
State Islamic University of Sultan Syarif Kasim Riau
UIN SUSKA RIAU

Tabel 2.4. Aturan Pembentukan *Prefiks Devirasional* Kedua

Surfiks	Pengganti	Kondisi Tambahan	Contoh
Ber	Null	Null	Berlari → lari
Bel	Null	Ajar	Belajar → ajar
Be	Null	Kerja	Bekerja → kerja
Per	Null	Null	Perjelas → jelas
Pel	Null	Ajar	Pelajar → ajar
Pe	Null	Null	Pekerja → kerja

Tabel 2.5. Aturan Pembentukan *Surfiks Devirasional*

Surfiks	Pengganti	Kondisi Tambahan	Contoh
Kan	Null	Prefiks (ke, peng)	tarikkan→ tarik (meng)ambilkan→ ambil
An	Null	Prefiks (di, meng, ter)	makanan → makan (per)janjian → janji
I	Null	Prefiks (ber, ke, peng)	tandai →tanda (men)dapati→ dapat

Contoh **:**

1. Sekolah → sekolah (kata dasar, tidak dilakukan *stemming*)
2. Duduklah → duduk (dilakukan proses *stemming*)

Berdasarkan urutan tahapan pada penanganan kata berimbuhan, maka terdapat beberapa kemungkinan dalam kesulitan membedakan suatu suku kata merupakan imbuhan atau bagian kata dasar:

1. Kata dasar mempunyai suku kata terakhir (partikel *infleksional*) serta kata tersebut tidak mendapat imbuhan apapun. Contoh: istilah.
2. Kata dasar mempunyai suku kata terakhir (partikel *infleksional*) dan mempunyai *prefiks*. Contoh: bersalah.
3. Kata dasar mempunyai suku kata terakhir (kata ganti *infleksional*) serta kata tersebut tidak mendapat imbuhan apapun. Contoh: bangku.

UIN SUSKA RIAU
Hak Cipta Dilindungi Undang-Undang
1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:
a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.
b. Pengutipan tidak merugikan kepentingan yang wajar UIN Suska Riau.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin UIN Suska Riau.
© Hak Cipta Milik UIN Suska Riau
State Islamic University of Sultan Syarif Kasim Riau
UIN SUSKA RIAU

4. Kata dasar mempunyai suku kata terakhir (kata ganti *infleksional*) dan mengandung *prefiks*. Contoh: bertanya.
5. Kata dasar mempunyai suku kata pertama (*prefiks devasional*) serta kata tersebut tidak mendapat imbuhan apapun. Contoh: diagram, kenang.
6. Kata dasar mempunyai suku kata kedua (*prefiks devasional* pertama) serta kata tersebut tidak mendapat imbuhan apapun. Contoh: disiplin, pentungan.
7. Kata dasar mempunyai suku kata pertama (*prefiks devasional* kedua) serta kata tersebut tidak mendapat imbuhan apapun. Contoh: pelangi, perban.
8. Kata dasar mempunyai suku kata pertama (*prefiks devasional*) dan mempunyai *surfiks devirasional*. Contoh: belakangan, pejamkan.
9. Kata dasar mempunyai suku kata terakhir (*sufiks devasional*) serta kata tersebut tidak mendapat imbuhan apapun. Contoh: koleksi, dominan.

Berdasarkan dari permasalahan tersebut, maka dibuat kamus-kamus kecil untuk melengkapi *stemming* ini. Terdapat 9 kamus kecil yaitu :

1. Kamus partikel. Seperti: masalah.
2. Kamus partikel ber*prefiks*. Seperti: menikah.
3. Kamus milik. Seperti : bangku.
4. Kamus milik ber*prefiks*. Seperti: bersuku.
5. Kamus prefiks 1. Seperti: median.
6. Kamus prefiks 1 ber*sufiks*. Seperti: terapan.
7. Kamus *prefiks* 2. Seperti: percaya.
8. Kamus *prefiks* 2 ber*sufiks*. Seperti: perasaan.
9. Kamus *sufiks*. Seperti: pantai.

Kondisi ukuran adalah jumlah minimum suku kata dalam sebuah kata. Karena dalam bahasa Indonesia, kata dasar setidaknya mempunyai dua suku kata. Maka kondisi ukuran dalam proses *stemming* bahasa Indonesia adalah dua suku kata didefinisikan memiliki satu vokal.

UIN SUSKA RIAU
Hak Cipta Dilindungi Undang-Undang
1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:
a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.
b. Pengutipan tidak merugikan kepentingan yang wajar UIN Suska Riau.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin UIN Suska Riau.
© Hak cipta milik UIN Suska Riau
State Islamic University of Sultan Syarif Kasim Riau

Hak Cipta Dilindungi Undang-Undang
1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:
a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.
b. Pengutipan tidak merugikan kepentingan yang wajar UIN Suska Riau.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin UIN Suska Riau.

2.3 Pembobotan

Pembobotan kemunculan kata dalam suatu dokumen digunakan untuk perhitungan tingkat kemiripan antar dokumen dengan *query* (Abror, 2011). Salah satu metode pembobotan yang sering digunakan adalah *TF-IDF*.

Terms Frequency – Inverse Document Frequency (TF-IDF)

Metode ini merupakan metode untuk menghitung nilai atau bobot suatu kata *(term)* pada dokumen. Metode ini akan mengabaikan setiap kata-kata yang tergolong tidak penting. Oleh sebab itu, sebelum melalukan metode ini, proses *stemming* dan *stopword removal* harus dilakukan terlebih dahulu oleh sistem (Pradnyana, 2012).

Faktor lain yang diperhatikan dalam pemberian bobot adalah kalimat yang muncul pada sedikit dokumen harus dipandang sebagai kata yang lebih penting daripada kalimat yang muncul pada banyak dokumen. Pembobotan akan memperhitungkan faktor kebalikan frekuensi dokumen yang mengandung suatu kalimat (*inverse document frequency*) (Pradnyana, 2012).

Nilai IDF sebuah *term* (kata) dapat dihitung menggunakan persamaan berikut:

$$\mathrm{idf}_{\mathrm{t}} = \log(\frac{N}{df_t}) \qquad (2.1)$$

N adalah jumlah dokumen yang berisi *term* (t) dan n adalah jumlah kemunculan frekuensi *term* terhadap D.

Adapun algoritma yang digunakan untuk menghitung bobot (W) masing-masing dokumen terhadap kata kunci (*query*) yaitu:

$$\mathrm{W}_{\mathrm{dt}} = tf_{d,t} \bullet idf_t \qquad (2.2)$$

Keterangan :

d= dokumen ke-d

t= *term* ke-t dari kata kunci

tf= *term* frekuensi/ frekuensi kata

W= bobot dokumen ke-d terhadap *term* ke-t

Setelah bobot (W) masing-masing dokumen diketahui, maka dilakukan proses pengurutan (*sorting*) dimana semakin besar nilai W, maka semakin besar

pula tingkat kesamaan (*similarity*) dokumen tersebut terhadap kata yang dicari, demikian pula sebaliknya.

Fungsi metode ini adalah untuk mencari representasi nilai dari tiap-tiap dokumen dari suatu kumpulan data training (*training set*). Dari sini akan dibentuk suatu vektor antara dokumen dengan kata (*documents with terms*) yang kemudian untuk kesamaan antar dokumen dengan *cluster* akan ditentukan oleh sebuah *prototype* vektor yang disebut juga dengan *cluster centroid* (Arifin).

2.4 *Cosine Similarity*

Metode *Cosine Similarity* merupakan metode yang digunakan untuk menghitung *similarity* (tingkat kesamaan) antar dua buah objek (Pradnyana, 2012).

Metode *cosine similarity* ini menghitung *similarity* antara dua buah objek (misalkan D1 dan D2) yang dinyatakan dalam dua buah vektor dengan menggunakan *keywords* (kata kunci) dari sebuah dokumen sebagai ukuran.

Ukuran ini memungkinkan perangkingan dokumen sesuai dengan kemiripannya (relevansi) terhadap *query*. Setelah semua dokumen dirangking, sejumlah tetap dokumen *top-scoring* dikembalikan kepada pengguna. Ukuran pada *cosine similarity* ini menghitung nilai *cosinus* sudut antara dua vektor. Jika terdapat dua vektor dokumen dj dan *query* q, serta *term* diekstrak dari lokasi dokumen maka nilai *cosinus* antara dj dan q didefenisikan sebagai berikut :

$$sim\ D_1, D_2 = \frac{\sum_i t_{1i} t_{2i}}{\sqrt{\sum_i t_{1i}^2} \times \sqrt{\sum_i t_{2i}^2}} \qquad (2.3)$$

Berikut adalah contoh untuk menghitung nilai *similarity*, pada sebuah dokumen dengan judul "**Udara**"

Udara
Udara adalah suatu campuran gas yang terdapat pada lapisan yang mengelilingi bumi dan komponen campuran gas tersebut tidak slalu konstan. Udara juga merupakan atmosfer yang berada di sekeliling bumi yang memiliki fungsi sangat penting. Karena di dalamnya terkandung oksigen dan karbon dioksida.

UIN SUSKA RIAU
Hak Cipta Dilindungi Undang-Undang
1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:
a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.
b. Pengutipan tidak merugikan kepentingan yang wajar UIN Suska Riau.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin UIN Suska Riau.
© Hak Cipta milik UIN Suska Riau
State Islamic University of Sultan Syarif Kasim Riau
UIN SUSKA RIAU

UIN SUSKA RIAU
Hak Cipta Dilindungi Undang-Undang
1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:
a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.
b. Pengutipan tidak merugikan kepentingan yang wajar UIN Suska Riau.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin UIN Suska Riau.
© Hak cipta milik UIN Suska Riau
State Islamic University of Sultan Syarif Kasim Riau

Langkah-langkah menetukan nilai *cosine similarity* adalah sebagai berikut:

1. Langkah pertama adalah menghapus tanda baca yang ada pada dokumen. Melakukan *tokenisasi, stopword removal, dan stemming*. Hasilnya dapat dilihat sebagai berikut :

   Q = udara

   D1= udara campuran gas lapisan keliling bumi komponen gas konstan.

   D2= udara atmosfer keliling bumi fungsi penting

   D3= kandung oksigen karbondioksida

2. Menentukan bobot setiap term dari ke 3 dokumen dengan persamaan (2.1)

	A	B	C	D	E	F	G
1		tf				df	idf
2		Q	D1	D2	D3		log (N/n)
3	udara	1	1	1		2	0.1761
4	campur		1			1	0.4771
5	gas		1			1	0.4771
6	keliling		1	1		2	0.1761
7	bumi		1	1		2	0.1761
8	atmosfer			1		1	0.4771
9	fungsi			1		1	0.4771
10	penting			1		1	0.4771
11	kandung				1	1	0.4771
12	oksigen				1	1	0.4771
13	karbondioksida				1	1	0.4771

Gambar 2.2 Mengitung Nilai *IDF*

3. Kemudian menentukan bobot W masing-masing dokumen terhadap kata kunci (*query*) yaitu dengan persamaan (2.2)

	A	B	C	D	E
1		Wdt = tf.idf			
2		Q	D1	D2	D3
3	udara	0.1761	0.1761	0.1761	0.0000
4	campur	0.0000	0.4771	0.0000	0.0000
5	gas	0.0000	0.4771	0.0000	0.0000
6	keliling	0.0000	0.1761	0.1761	0.0000
7	bumi	0.0000	0.1761	0.1761	0.0000
8	atmosfer	0.0000	0.0000	0.4771	0.0000
9	fungsi	0.0000	0.0000	0.4771	0.0000
10	penting	0.0000	0.0000	0.4771	0.0000
11	kandung	0.0000	0.0000	0.0000	0.4771
12	oksigen	0.0000	0.0000	0.0000	0.4771
13	karbondioksida	0.0000	0.0000	0.0000	0.4771

Gambar 2.3 Mengitung Bobot W

4. Selanjutnya adalah menghitung jumlah perkalian bobot Wij.Wiq

	A	B	C	D
1		Wij.Wiq		
2		D1	D2	D3
3	udara	0.0310	0.0310	0.0000
4	campur	0.0000	0.0000	0.0000
5	gas	0.0000	0.0000	0.0000
6	keliling	0.0000	0.0000	0.0000
7	bumi	0.0000	0.0000	0.0000
8	atmosfer	0.0000	0.0000	0.0000
9	fungsi	0.0000	0.0000	0.0000
10	penting	0.0000	0.0000	0.0000
11	kandung	0.0000	0.0000	0.0000
12	oksigen	0.0000	0.0000	0.0000
13	karbondioksida	0.0000	0.0000	0.0000

Gambar 2.4 Mengitung Nilai Perkalian Bobot Terhadap *Query*

5. Yang terakhir adalah menetapkan rumus *cosine similarity* dengan menghitung kemiripan *query* dengan dokumen 1, dokumen 2, dan dokumen 3. Maka didapat nilai *similarity* sebagai berikut :

	A	B	C	D	E	F	G	H
1		w kuadrat				Wij.Wiq		
2		Q1	D1	D2	D3	D1	D2	D3
3	udara	0.0310	0.0310	0.0310	0.0000	0.0310	0.0310	0.0000
4	campur	0.0000	0.2276	0.0000	0.0000	0.0000	0.0000	0.0000
5	gas	0.0000	0.2276	0.0000	0.0000	0.0000	0.0000	0.0000
6	keliling	0.0000	0.0310	0.0310	0.0000	0.0000	0.0000	0.0000
7	bumi	0.0000	0.0310	0.0310	0.0000	0.0000	0.0000	0.0000
8	atmosfer	0.0000	0.0000	0.2276	0.0000	0.0000	0.0000	0.0000
9	fungsi	0.0000	0.0000	0.2276	0.0000	0.0000	0.0000	0.0000
10	penting	0.0000	0.0000	0.2276	0.0000	0.0000	0.0000	0.0000
11	kandung	0.0000	0.0000	0.0000	0.2276	0.0000	0.0000	0.0000
12	oksigen	0.0000	0.0000	0.0000	0.2276	0.0000	0.0000	0.0000
13	karbondioksida	0.0000	0.0000	0.0000	0.2276	0.0000	0.0000	0.0000
14		panjang vektor				jumlah perkalian bobot		
15		0.1761	0.7405	0.8809	0.8264	0.0310	0.0310	0.0000
16								
17						cosine similarity		
18						0.2378	0.1999	0.0000

Gambar 2.5 Mengitung Nilai *Cosine Similarity*

UIN SUSKA RIAU
Hak Cipta Dilindungi Undang-Undang
1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:
a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.
b. Pengutipan tidak merugikan kepentingan yang wajar UIN Suska Riau.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin UIN Suska Riau.
© Hak cipta milik UIN Suska Riau
State Islamic University of Sultan Syarif Kasim Riau

Tabel 2.6 Nilai *Cosine Similarity*

Dokumen 1	Dokumen 2	Dokumen 3
0.2378	0.1999	0.0000

2.5 *Maximum Marginal Relevance* (*MMR*)

Algoritma *MMR* merupakan salah satu metode *extractive summary* yang digunakan dalam peringkasan *single* dokumen maupun multidokumen, mempunyai karakter yang efektif dan sederhana (Mustaqhfiri, 2011). Algoritma MMR digunakan untuk menghitung tingkat kesamaan (*similarity*) antara satu dokumen dengan dokumen yang lain dalam sebuah postingan. Bobot tertinggi dokumen nantinya yang akan menjadi ringkasan terpenting (inti) dari dokumen tersebut.

Pada peringkasan dokumen dengan metode MMR, dilakukan proses segementasi dokumen manjadi kalimat dan dilakukan pengelompokan sesuai dengan gender kalimat tersebut (Mustaqhfiri, 2011). Algoritma MMR nantinya digunakan dengan mengkombinasikan matriks *cosine similarity* untuk merangking kalimat-kalimat yang terlebih dahulu dibobot dengan menggunakan algoritma *TF-IDF*.

Kebanyakan mesin pencarian *information retrieval* (IR) modern menghasilkan daftar perangkingan dari dokumen yang diukur dari penurunan relevansi terhadap permintaan user (*user query*). Penaksiran pertama untuk mengukur hasil peringkasan yang *relevan* adalah dengan mengukur hubungan antar informasi yang ada dalam dokumen dengan *query* yang diberikan oleh *user* dan menambah kombinasi linier sebagai sebuah matrik (Mustaqhfiri, 2011).

Peringkasan dokumen dengan tipe ekstraktif, nilai akhir diberikan pada kalimat Di dalam *MMR* dihitung dengan persamaan :

$$MMR = argmax\,[\lambda * Sim_1\ D, Q\ -\ 1 - \lambda\ * \max Sim_2(D_i, D')] \qquad (2.4)$$

Dimana Di merupakan kalimat dokumen, sedangkan D' adalah kalimat yang sudah dipilih atau diekstrak sebelumnya. Koefisien λ digunakan mengatur kombinasi nilai untuk memberi penekanan bahwa kalimat tersebut relevan dan untuk mengurangi redudansi. Sim_1 dan Sim_2 merupakan dua fungsi *similarity* yang mempresetasikan kesamaan kalimat pada sebuah dokumen dan memilih

Hak Cipta Dilindungi Undang-Undang
1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:
a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.
b. Pengutipan tidak merugikan kepentingan yang wajar UIN Suska Riau.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin UIN Suska Riau.
© Hak cipta milik UIN Suska Riau
State Islamic University Sultan Syarif Kasim Riau
UIN SUSKA RIAU

masing-masing kalimat untuk dijadikan ringkasan. Sim_1 adalah matrik *similarity* kalimat Di terhadap *query* yang diberikan oleh user, sedangkan Sim_2 adalah matrik *similarity* kallimat Di terhadap kalimat yang telah diekstrak sebelumnya (Mustaqhfiri, 2011).

Nilai parameter λ adalah 0 sampai 1 atau range 0,1. Pada saat parameter λ = 1 maka nilai MMR yang diperoleh akan cenderung relevan terhadap dokumen asal. Ketika nilai parameter λ = 0 maka nilai diperoleh cenderung relevan terhadap kalimat yang telah diekstrak sebelumnya yang akan dibandingkan. Untuk peringkasan yang baik pada dokumen pendek digunakan nilai parameter λ = 0.7 atau λ = 0.8 (Mustaqhfiri, 2011).

2.6 *Spam*

Spam adalah penyalahgunaan perangkat elektronik untuk mengirimkan pesan secara terus menerus tanpa dikehendaki oleh si penerima (Pradipa, 2013).

Spam bisa juga berarti aktivitas pengiriman berita atau iklan yang diluar topik yang sedang dibicarakan dalam sebuah newsgroup (Elyyani, 2011). Para pengirim *spam* yang disebut dengan *spammer* akan dilaporkan secara otomatis ke *public blacklist database* yaitu semacam direktori khusus yang mem-*blacklist email* yang bermasalah (Elyyani, 2011).

Spam pada forum bisa berupa iklan atau pesan yang tidak diinginkan diforum internet. Kebanyakan *spam* diforum ini berupa iklan yang di*link*kan ke banyak situs, misalnya situs penjualan, perjudian atau bahkan situs porno.

2.7 Diskusi *Online*

Menurut Kamus Besar Bahasa Indonesia diskusi adalah sebuah pertemuan ilmiah untuk bertukar pikiran mengenai suatu masalah yang dilakukan oleh sekelompok orang yang membahas topik yang sedang menjadi perhatian dihadapan khalayak dan khalayak diberi kesempatan untuk bertanya atau memberikan pendapat.

Diskusi online adalah suatu proses berbicara, percakapan ilmiah yang berisikan pertukaran pendapat, pemunculan ide-ide serta pengujian pendapat yang dilakukan oleh beberapa orang yang tergabung dalam suatu kelompok yang di fasilitasi dengan *internet*. Dalam sebuah diskusi dibutuhkan seorang pemimpin

UIN SUSKA RIAU
Hak Cipta Dilindungi Undang-Undang
1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:
a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.
b. Pengutipan tidak merugikan kepentingan yang wajar UIN Suska Riau.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin UIN Suska Riau.
© Hak cipta milik UIN Suska Riau
State Islamic University of Sultan Syarif Kasim Riau
UIN SUSKA RIAU

yang disebut ketua diskusi. Kegiatan diskusi bertujuan untuk memperoleh suatu kesepakatan, pengertian, dan keputusan bersama tentang suatu masalah. Unsur penting diskusi adalah adanya forum tanya jawab selama proses berlangsung.

UIN SUSKA RIAU
© Hak cipta milik UIN Suska Riau
State Islamic University of Sultan Syarif Kasim Riau
Hak Cipta Dilindungi Undang-Undang
1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:
a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah.
b. Pengutipan tidak merugikan kepentingan yang wajar UIN Suska Riau.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin UIN Suska Riau.