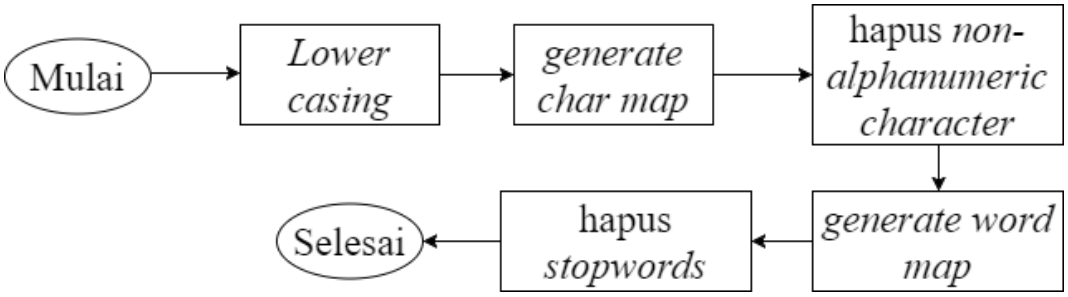
0.1. Perancangan Sistem

0.1.1 *Preprocessing*

Gambar 1 menunjukan rincian tahapan untuk proses *preprocessing* yang merupakan bagian dari tahapan *merging context seeds* pada Gambar **??**.



Gambar 1: Alur *Preprocessing*

Diketahui dokumen X(*suspicious-document*) dan dokumen Y (*source-document*) yang dijadikan *input* untuk sistem. *Preprocessing* dilakukan terhadap kedua dokumen. Sebagai contoh, proses yang ditunjukan pada Sub-Bab ini merupakan proses untuk dokumen *suspicious-document00044.txt*

Lower casing* dan *Generate Character Map

Preprocessing dimulai dengan *Lower-casing*, atau mengubah seluruh alfabet yang ada pada dokumen menjadi huruf kecil. Kemudian memetakan seluruh karakter yang ada pada dokumen ke dalam *char map* untuk mengetahui letak kemunculan tiap karakter pada dokumen.

```
A GED is accepted by all public and most private colleges
and universities, as well as most employers.  GMAT is the
most effective test available for admission to business scho-
ols.emand for qualified professionals in education and re-
search industry is increasing.  Prepare for the GRE and get
going.  Getting admission in a law school is now easy wi-
th the LSAT Preparation course which thoroughly teaches you
the techniques and helps to build skills to take the LSAT
test.  The cost of the course is $125 and is open to the
first 40 individuals who meet the following qualifications:
.......
```

Akan membangun *char map* seperti pada Tabel 1.

Tabel 1: Contoh *Char Map*

index	karakter
0	a
1	
2	g
3	e
4	d
5	
6	i
7	s
8	
....	....
12160	
12161	g
12162	e
12163	t

Hapus karakter *non-alphanumeric*

Langkah berikutnya menghapus seluruh baris yang memiliki karakter yang tidak termasuk ke dalam *alphanumeric* karakter. Karakter yang termasuk ke dalam karakter *non-alphanumeric* ditunjukan pada Tabel 2.

Tabel 2: Daftar Karakter *Non-alphanumeric*

‘	∼	!	@	#	$
%	ˆ	&	*	(	)
-	_	+	=	{	[
}	]	—	\	;	:
”	’	<	,	>	.
?	/				

Generate Word Map

Setelah menghilangkan seluruh karakter *non-alphanumeric* maka dibangun *word map* dari karakter yang tersisa yang ada pada *char map* dengan menyimpan informasi kemunculan awal dan akhir karakter pada kata tersebut. Tiap kata akan dibangun dengan pemisah berupa karakter spasi. Contoh dari tabel *word map* ditunjukan pada Tabel 3. Dengan banyak jumlah data : **129**

Tabel 3: Contoh *Word Map*

Indeks awal	Indeks akhir	Kata
0	0	a
2	4	ged
6	7	is
9	16	accepted
18	19	by
...	...	...

Hapus *stopwords*

Tahap akhir *preprocessing* adalah menghilangkan *stopwords*. *Stopwords* merupakan kata yang umum muncul, sehingga dianggap memberikan *noise* pada data. Daftar *stopwords* dapat dilihat pada Tabel 4

Tabel 4: Daftar *Stopwords*

i	herself	was	because	from	any	t
me	it	were	as	up	both	can
my	its	be	until	down	each	will
myself	itself	been	while	in	few	just
we	they	being	of	out	more	don
our	them	have	at	on	most	should
ours	their	has	by	off	other	now
ourselves	theirs	had	for	over	some	
you	themselves	having	with	under	such	
your	what	do	about	again	no	
yours	which	does	against	further	nor	
yourself	who	did	between	then	not	
yourselves	whom	doing	into	once	only	
he	this	a	through	here	own	
him	that	an	during	there	same	
his	these	the	before	when	so	
himself	those	and	after	where	than	
she	am	but	above	why	too	
her	is	if	below	how	very	
hers	are	or	to	all	s	

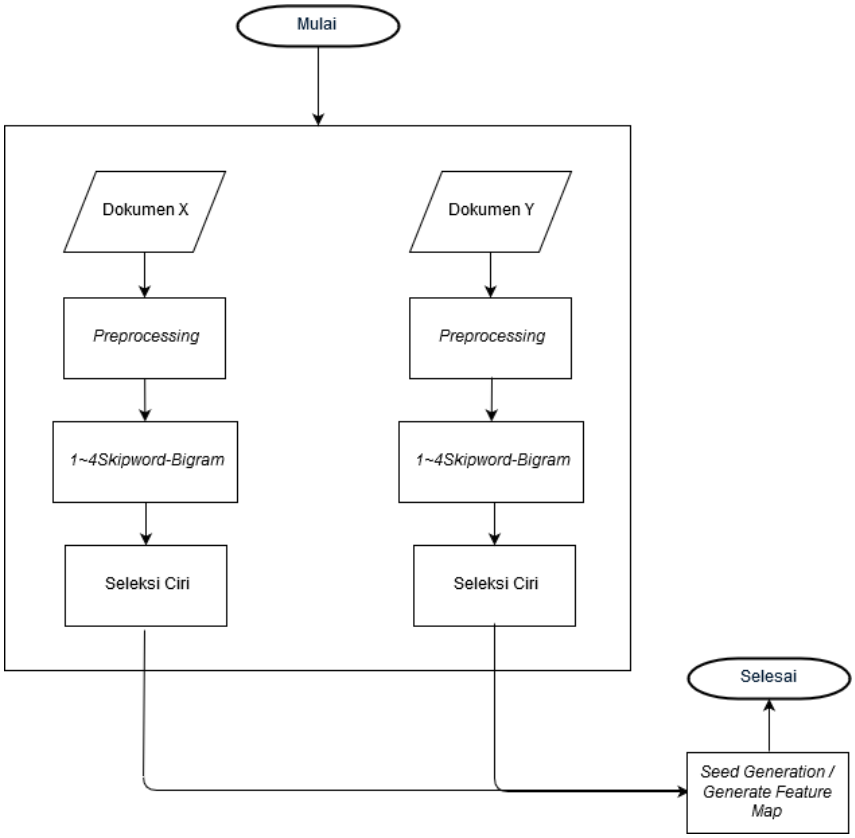
Sehingga pada akhir tahap *preprocessing*, *word map* yang dihasilkan ditunjukan pada Tabel 5. Dengan jumlah data : **84**

Tabel 5: Contoh Word Map yang Telah Dihilangkan *Stopwords*

Indeks awal	Indeks akhir	Kata
2	4	ged
9	16	accepted
25	30	public
41	47	private
49	56	colleges
...	...	...
12149	12152	what

0.1.2 *Seed Generation*

Gambar 2 menunjukan rincian tahapan untuk proses *seed generation* yang merupakan bagian dari tahapan *merging context seeds* pada Gambar **??**. Seluruh proses yang ditampilkan merupakan proses yang dilakukan terhadap dokumen X *(suspicious-document00017.txt)* dan dokumen Y *(source-document00135.txt)*.



Gambar 2: Alur *Seeds Generation*

Ekstraksi Ciri

Pada tahap ini dokumen X(*suspicious*) dan dokumen Y(*source*)yang telah melalui tahap *preprocessing* akan melalui proses ekstraksi ciri dengan *skipword-bigram* 1-4 dengan tetap menyimpan informasi letak kemunculan karakter pada kata/*token*. Dari proses ekstraksi ini akan dibangun *feature map* yang ditunjukan oleh Tabel 6 seperti yang dijelaskan pada Persamaan **??**.

Tabel 6: *Feature Map* Dokumen X

i awal	i akhir	Kata	f1	f2	f3	f4
2	4	ged	*_ged	*_ged	*_ged	*_ged
9	16	accepted	ged_accepted	*_accepted	*_accepted	*_accepted
25	30	public	accepted_public	ged_public	*_public	*_public
41	47	private	all_private	public_private	ged_private	*_private
49	56	colleges	private_colleges	public_colleges	accepted_colleges	ged_colleges

Seleksi Fitur/Ciri

Pada metode *Merging Context Seeds*, fitur yang mempunyai jumlah kemunculan yang tinggi dianggap tidak relevan. Sehingga apabila ada fitur yang kemunculannya melebihi **4**, fitur akan dihapus. Nilai 4 merupakan parameter yang sudah diuji[**?**] untuk mendapatkan fitur yang optimal.

Sebagai contoh : ***jika*** *fitur* ***patch_alabama*** *pada dokumen X muncul sebanyak* ***6****. Maka fitur tersebut akan dihapus seluruhnya dari feature map, baik fitur tersebut muncul di f1,f2,f3, atau f4.*

Seed Generation

Pada tahap ini fitur yang ada pada dokumen X dan dokumen Y. Akan dicari irisan antara kedua dokumen tersebut dari fitur yang ada. Apabila terdapat fitur yang sama, maka akan dibangun *seed set* yang merupakan titik awal pendeteksian *passage* yang diplagiat pada kedua dokumen, atau disebut dengan *passage reference*. Tabel 7 menunjukan hasil *seed generation* antara dokumen X *(suspicious-document00017.txt)* dan dokumen Y *(source-document00135.txt)*.

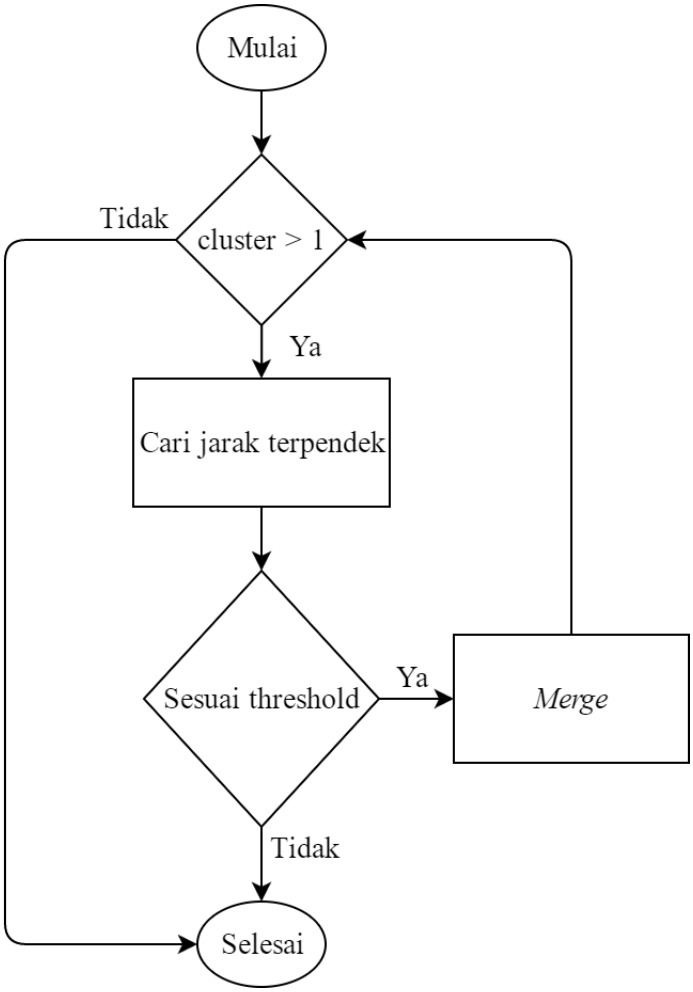
Tabel 7: *Passage Reference* Dokumen X dan Dokumen Y

pr_x	f	Kata	$Y_{f_{start}}$	$Y_{f_{end}}$	$X_{f_{start}}$	$X_{f_{start}}$
pr_0	gmat_test	test	1181	1199	130	133
pr_1	gmat_questions	questions	1191	1199	1043	1051
pr_2	gmat_questions	questions	1191	1199	1103	1111
pr_3	gmat_questions	questions	1191	1199	1308	1316
...	...	...	...	...	...	...
pr_{461}	takers_right	right	1300	1304	9868	9872

Kolom $Y_{f_{start}}$ menunjukan kemunculan awal kata pada *feature map* pada dokumen Y, sedangkan $Y_{f_{end}}$ menunjukan akhir kemunculan kata pada *feature map* pada dokumen Y. Sedangkan $X_{f_{start}}$ dan $X_{f_{end}}$ menunjukan kemunculan pada dokumen X. Seluruh letak kemunculan pada Tabel 7 didapatkan dari proses ekstraksi ciri.

0.1.3 Merging

Gambar 3 menunjukan rincian tahapan untuk proses *merging* yang merupakan bagian dari tahapan *merging context seeds* pada Gambar **??**.



Gambar 3: Alur *Merging*

Proses *merging* merupakan proses dimana sistem akan meng-*cluster* setiap *passage reference* menggunakan algoritma *single linkage clustering*. Kriteria *merge* yang digunakan adalah jarak terpendek antara 2 *passage reference* dari kumpulan *passage reference* yang ada. Proses *merging* akan terhenti apabila memenuhi salah satu dari dua kondisi terminasi berikut :

1. Jarak terpendek $\leq \tau; \tau = 7$.
   Nilai 7 dianggap nilai yang paling optimal dari pengujian sementara dengan range nilai 5 - 15.

2. Jumlah cluster = 1.
   Yang berarti seluruh *seed* yang didapat sudah masuk kedalam satu *cluster*.

3. Jumlah cluster = 0.
   Tidak terdapat *seed* pada proses *seed generation* sebelumnya.

Tabel 8 menunjukan jarak antar *seed* atau *passage reference* yang didapat dari proses *seed generation*.

Tabel 8: Jarak Antar *Passage Reference*

Dist	pr_0	pr_1	pr_2	pr_3	pr_4	...	pr_{15}	...
pr_0	x	41.6	44.2	53.3	178.2	...	...	...
pr_1	41.6	x	1.8	8.1	90.2	...	...	...
pr_2	44.2	1.8	x	6.3	88.3	...	...	...
pr_3	53.3	8.1	6.3	x	81.6	...	...	...
pr_4	178.2	90.2	88.3	81.6	x	...	...	...
...	...	...	...	...	...	...	...	...
pr_{13}	...	...	...	...	...	...	0.0	...
...	...	...	...	...	...	...	...	...

Dari hasil *loop* pertama diketahui jarak antara *passage reference* terdekat ada di pr_{13} dan pr_{15} dengan jarak 0.0, maka pr_{13} dan pr_{15} akan di *merge*. Data kandidat yang ditunjukan pada Tabel 9. Nilai kemunculan pada dokumen X dan dokumen Y didapatkan dari proses ekstraksi ciri pada tahap sebelumnya.

Tabel 9: Kandidat *Passage Reference* yang Akan Di-*merge*

pr_x	Kata	$Y_{f_{start}}$	$Y_{f_{end}}$	$X_{f_{start}}$	$X_{f_{end}}$
pr_{13}	reasoning	1861	1869	3874	3882
pr_{15}	reasoning	1861	1869	3874	3882

Pada tabel 10 terlihat bahwa nilai $X_{f_{start}}$ dan $X_{f_{end}}$ diambil dari titik terendah untuk $X_{f_{start}}$, dan titik terjauh untuk $X_{f_{start}}$ dikarenakan penggabungan dua buah *passage reference*. Dikarenakan *passage reference* yang akan digabung merupakan fitur yang ada pada kata yang sama, sehingga tidak ada perubahan untuk nilai kemunculannya

Tabel 10: Hasil Merge

pr_x	Kata	$Y_{f_{start}}$	$Y_{f_{end}}$	$X_{f_{start}}$	$X_{f_{end}}$
$pr_{13,15}$	reasoning, reasoning	1861	1869	3874	3882

Proses akan diulang hingga tidak ada jarak antar *passsage reference* yang melebihi τ. Tabel 8 menunjukan jarak terpendek antar *passage reference* pada *looping* pertama.

0.1.4 *Filtering*

Pada tahap ini *cluster* akhir akan dipilihi, apabila *cluster* mempunyai jumlah anggota atau *passage reference* $\geq \tau; \tau = 15[?]$ maka $pr_{n..m}$ dianggap menjadi bagian yang diplagiat (r).

0.1.5 *Output*

Dari tahap *filtering* didapat lokasi tindak plagiat, pada tahap *output* ini mengembalikan bagian dari kedua dokumen.

Tabel 11: $r \in R$

r	$Y_{f_{start}}$	$Y_{f_{end}}$	$X_{f_{start}}$	$X_{f_{end}}$
r_1	8	979	3805	4427
r_2	1160	1620	9073	9461
r_3	1719	2134	8269	8580

Tabel 11 menunjukan $r \in R$ pada *pair suspicious-document00044.txt-source-document01326.txt* yang didapat dari hasil akhir *merging* yang didapat dari tahap sebelumnya. Kemudian dibangkitkan nilai kemunculannya mejadi teks pada dokumen sesungguhnya, sehingga hasil yang dimunculkan ditunjukan seperti pada Tabel 12.

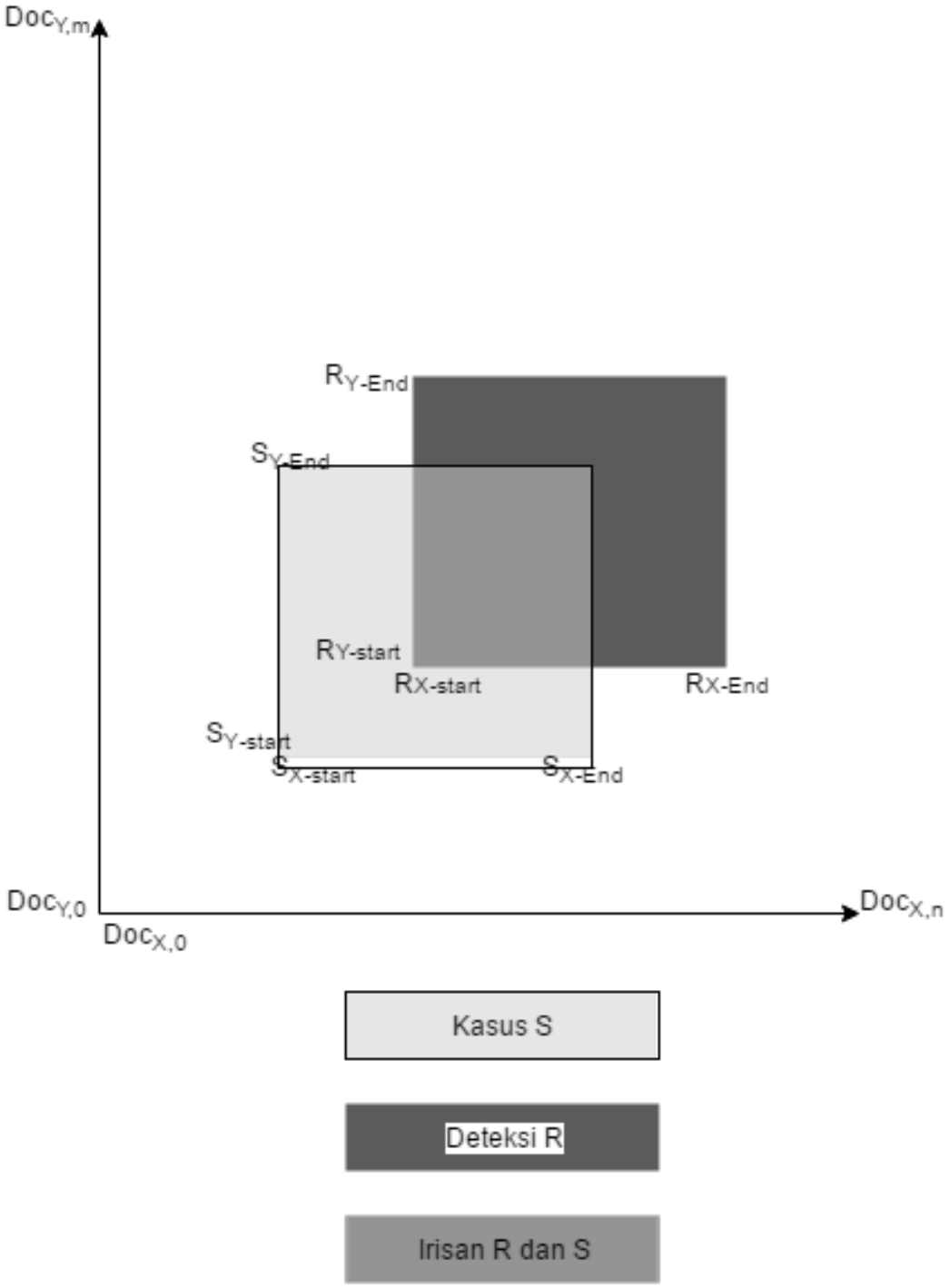
Tabel 12: *Output* akhir berupa teks yang diplagiat

r	Teks Pada Dokumen X	Teks Pada Dokumen Y
r_1	questions of three question types – reading comprehension, critical reasoning, and sentence correction. you are allowed 75 minutes to complete this entire section. verbal section each passage engages with a specialized topic or opinion in either the humanities, social sciences, science, or business, but no specific outside knowledge of the material is required; all questions refer to what is stated or implied in the text. the directions for these questions look like this: each passage is followed by questions about its content. after reading a passage, select the best answer to each question among the five choice	comprehension this is a partial free sample of our prep guide. to view the remainder of this page, purchase the . typically, your verbal test will include 4 reading comprehension passages, with 3 to 4 questions per passage, for a total of 12 to 14 questions of the 41 verbal questions. each passage engages with a specialized topic or opinion in either the humanities, social sciences, science, or business, but no specific outside knowledge of the material is required; all questions refer to what is stated or implied in the text. the directions for these questions look like this: each passage is followed by questions about its content. after reading a passage, select the best answer to each question among the five choices. answer all questions following a passage on the basis of what the passage states or implies.directions: a passage and a corresponding question look like this: the screen will split into two with the passage on the left and the question
r_2	computer-adaptive test) so the questions will begin at an intermediate skill level and . in general, average test takers will get about 50% right of the questions right. as result, higher scorers are effectively taking a completely different test from lower scorers and their strategies will be adjusted accordingly. higher scorers will get longer and more challenging essays and question	computer-adaptive test) so the questions will begin at an intermediate skill level and . in general, average test takers will get about 50% right of the questions right. as result, higher scorers are effectively taking a completely different test from lower scorers and their strategies will be adjusted accordingly. higher scorers will get longer and more challenging essays and questions. this chapter has sections specifically designed to help higher scorer

r_3	adapt to your performance by changing in difficulty if you are extremely good at sentence correction and weak at reading comp and critical reasoning.... guess what? your skill in sentence correction will make the gmat deliver you very hard reading comp and critical reasoning questions. the moral of the stor	adapt to your performance by changing in difficulty if you are extremely good at sentence correction and weak at reading comp and critical reasoning.... guess what? your skill in sentence correction will make the gmat deliver you very hard reading comp and critical reasoning questions. the moral of the story.... be balanced on verbal and skilled at all three question types.how the cat impacts verbal difficult

0.1.6 Evaluasi

Evaluasi dilakukan untuk mengukur perfomansi sistem yang dibangun. Evaluasi pada tahap ini dilakukan pada level *pair*. Cara untuk mengukur perfomansi adalah dengan membandingkan $r \in R$ keluaran sistem, dengan $s \in S$ dari *dataset*. Tiap elemen R akan dicari irisan untuk tiap elemen S untuk mendapatkan nilai ***True Positive***. Sedangkan elemen R dianggap ***Test Outcome Positive*** dan elemen S sebagai ***Condition Positive***. Gambar 4 menunjukan bagaimana perhitungan untuk nilai ***True Positive, Test Outcome Positive, Condition Positive***. Dimana pada Gambar 4 menunjukan area plagiat yang ada pada *dataset* dan area plagiat yang dideteksi oleh sistem.



Gambar 4: Perhitungan Nilai Perfomansi

Tabel 13 merupakan *set case* $s \in S$ atau bagian yang dinyatakan plagiat untuk *pair suspicious-document00044.txt-source-document01326.txt* yang diapat dari *dataset*.

Tabel 13: $s \in S$

s	$Y_{f_{start}}$	$Y_{f_{end}}$	$X_{f_{start}}$	$X_{f_{end}}$
s_1	298	741	3975	4418
s_2	1143	1548	9028	9433
s_3	1713	2032	8240	8559

Sedangkan Tabel 14 menunjukan jumlah luas irisan area plagiat atau *passage* dari $r \in R$ dan $s \in S$.

Tabel 14: $R \cap S$

$\pi(r_x \cap s_y)$	r_1	r_2	r_3
s_1	1772	0	0
s_2	0	1496	0
s_3	0	0	1204

Sehingga perhitungan perfomansi untuk *pair suspicious-document00044.txt-source-document01326.txt* ditunjukan pada Tabel 15.

Tabel 15: Contoh Perhitungan Perfomansi pada Level 1 Dokumen

	True Positive	4474
Condition Positive	True Positive + False Negative	4468
Test Outcome Positive	True Positive + False Positive	6334
Precision	True Positive / Prediction Positive	0.706
Recall	True Positive / Condition Positive	0.958
F_1	$2 \cdot \frac{prec(S,R) \cdot rec(S,R)}{prec(S,R) + rec(S,R)}$	0.813

Dari Tabel 15 didapat informasi bahwa *test accuracy* atau nilai F_1 untuk *pair suspicious-document00044.txt-source-document01326.txt* adalah 0.813.