

**APLIKASI DETEKSI PLAGIARISME MENGGUNAKAN  
METODE *COSINE SIMILARITY***

**SKRIPSI**

**Oleh :  
RITO PUTRIWANA PRATAMA  
NIM. 14650044**



**JURUSAN TEKNIK INFORMATIKA  
FAKULTAS SAINS DAN TEKNOLOGI  
UNIVERSITAS ISLAM NEGERI MAULANA MALIK IBRAHIM  
MALANG  
2018**

**APLIKASI DETEKSI PLAGIARISME MENGGUNAKAN  
METODE *COSINE SIMILARITY***

**SKRIPSI**

**Diajukan kepada:  
Fakultas Sains dan Teknologi  
Universitas Islam Negeri (UIN)  
Maulana Malik Ibrahim Malang  
Untuk Memenuhi Salah Satu Persyaratan Dalam  
Memperoleh Gelar Sarjana Komputer (S.Kom)**

**Oleh :  
RITO PUTRIWANA PRATAMA  
NIM. 14650044**

**JURUSAN TEKNIK INFORMATIKA  
FAKULTAS SAINS DAN TEKNOLOGI  
UNIVERSITAS ISLAM NEGERI MAULANA MALIK IBRAHIM  
MALANG  
2018**

**LEMBAR PERSETUJUAN**

**APLIKASI DETEKSI PLAGIARISME MENGGUNAKAN  
METODE *COSINE SIMILARITY***

**SKRIPSI**

Oleh:  
**RITO PUTRIWANA PRATAMA**  
NIM. 14650044

Telah Diperiksa dan Disetujui Untuk diuji:  
Tanggal : .....

**Pembimbing I**



**Dr. M. Faisal, MT**  
NIP. 1974501 200501 1 007

**Pembimbing II**



**Ajib Hanani, M.T**  
NIDT. 19840731201608011076

Mengetahui,  
Ketua Jurusan Teknik Informatika  
Fakultas Sains dan Teknologi  
Universitas Islam Negeri Maulana Malik Ibrahim



**Dr. Cahyo Crysdian**  
NIP. 19740424 200901 1 008

LEMBAR PENGESAHAN

APLIKASI DETEKSI PLAGIARISME MENGGUNAKAN  
METODE *COSINE SIMILARITY*

SKRIPSI

Oleh:  
**RITO PUTRIWANA PRATAMA**  
NIM. 14650044

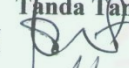



Telah Dipertahankan di Depan Dosen Penguji Skripsi dan  
Dinyatakan Diterima Sebagai Salah Satu Persyaratan Untuk  
Memperoleh Gelar Sarjana Komputer (S.Kom)

Tanggal : November 2018

Susunan Dewan Penguji

Penguji Utama : Dr. Suhartono, M. Kom  
NIP. 19680519 200312 1 001  
Ketua Penguji : Hani Nurhayati, M.T  
NIP. 19780625 200801 2 006  
Sekretaris : Dr. Muhammad Faisal, M.T  
Penguji : NIP. 19740510 200501 1 007  
Anggota Penguji : Ajib Hanani, M.T  
NIDT. 19840731 20160801 1 076

Tanda Tangan

(  )  
(  )  
(  )  
(  )

Mengetahui dan Mengesahkan  
Ketua Jurusan Teknik Informatika  
Fakultas Sains dan Teknologi  
Universitas Islam Negeri Maulana Malik Ibrahim Malang



Dr. Cahyo Crysdian  
NIP. 19740424 200901 1 008

## PERNYATAAN KEASLIAN TULISAN

Yang bertanda tangan di bawah ini :

Nama : Rito Putriwana Pratama

NIM : 14650044

Fakultas : Sains dan Teknologi

Jurusan : Teknik Informatika

Judul Skripsi : **APLIKASI DETEKSI PLAGIARISME  
MENGUNAKAN METODE *COSINE SIMILARITY***

Menyatakan dengan sebenar-benarnya bahwa hasil penelitian saya ini tidak mengandung unsur-unsur penjiplakan karya penelitian atau karya ilmiah yang pernah dilakukan atau dibuat oleh orang lain, kecuali yang secara tertulis dikutip dalam naskah ini dan disebutkan dalam sumber kutipan dan daftar pustaka.

Apabila ternyata hasil penelitian ini terbukti mengandung unsur-unsur penjiplakan, maka saya bersedia untuk mempertanggungjawabkan, serta diproses sesuai peraturan yang berlaku.

Malang, November 2018  
Yang membuat pernyataan,



Rito Putriwana Pratama  
NIM. 14650044

## MOTTO

---

“Saat Allah mendorongmu ke tebing, yakinlah kalau hanya ada dua hal yang mungkin terjadi. Mungkin saja Ia akan menangkapmu, atau Ia ingin kau belajar bagaimana caranya terbang.”

-Anonim-



## PERSEMBAHAN

---

**Puji syukur kehadiran Allah SWT, shalawat dan salam selalu terlimpahkan kepada Rasulullah SAW.**

**Saya persembahkan karya sederhana ini kepada :**

Kedua orang tua yang sangat saya cintai, Bapak Suhadak dan Ibu Supriasih

Saudara yang sangat saya sayangi, Adik Mochammad Amirul Wicaksono

Dosen pembimbing saya Bapak Dr. Muhammad Faisal, M.T dan Bapak Ajib

Hanani, M.T, serta seluruh dosen dan staf Teknik Informatika UIN Maulana

Malik Ibrahim Malang

Sahabat-sahabat saya yang selalu memberi doa dan dukungan untuk saya agar segera menyelesaikan tugas akhir ini

Keluarga besar Teknik Informatika B, keluarga Teknik Informatika 2014 (Biner),

dan seluruh keluarga besar Teknik Informatika UIN Maulana Malik Ibrahim

Malang dari semua angkatan yang telah membantu saya

Orang-orang yang saya sayangi yang tidak bisa saya sebut satu persatu

Terimakasih atas doa, dukungan, dan motivasinya untuk saya. Semoga silaturahmi tetap terjaga dan selalu di Ridhoi Allah SWT. Aamiin

**Rito Putriwana Pratama**

## KATA PENGANTAR

*Assalamu'alaikum Wr. Wb.*

Segala puji bagi Allah yang telah melimpahkan Rahmat dan Karunia-Nya kepada penulis, sehingga penulis dapat menyelesaikan tugas akhir ini. Serta shalawat dan salam tercurahkan kepada Nabi Muhammad SAW yang membawa umatnya dari zaman gelap menuju zaman yang terang benderang yakni *addinul islam wal iman*.

Dan tidak lupa penulis sampaikan terimakasih yang sebesar-besarnya kepada pihak-pihak yang membantu dan mendukung penulis dalam menyelesaikan tugas akhir ini, baik secara langsung maupun tidak langsung.

1. Kepada kedua orang tua yang telah mendidik dan membesarkan penulis selama ini. Serta memberikan doa, dukungan dan pengertiannya selama penulis menyelesaikan studi.
2. Kepada Bapak Dr. Muhammad Faisal, MT, selaku Dosen Pembimbing I yang telah banyak meluangkan waktu, memberi motivasi, mengarahkan dan selalu sabar membimbing penulis, sehingga penulis dapat menyelesaikan tugas akhir ini.
3. Kepada Bapak Ajib Hanani, M.T, selaku Dosen Pembimbing II yang selalu mendukung, mengarahkan dan memberi masukan dalam pengerjaan tugas akhir ini.
4. Kepada Ibu Roro Inda Melani, M.T., M.Sc, selaku Dosen Wali yang telah banyak memberikan saran serta masukan kepada penulis selama studi.



5. Kepada seluruh Dosen Teknik Informatika UIN Malang yang telah memberikan ilmu kepada penulis selama studi.
6. Kepada seluruh Laboran dan Staf Teknik Informatika UIN Malang yang telah memberikan bimbingan.
7. Dan kepada semua pihak yang tidak bisa penulis sebut satu persatu, terimakasih atas segala yang diberikan kepada penulis baik pengertian, perhatian dan dukungannya selama ini.

Akhirnya, penulis berharap bahwa tugas kahir ini dapat memberikan kontribusi keilmuan untuk melanjutkan dan mengembangkannya. Namun, penulis menyadari bahwa tugas ini masih banyak kekurangan dan masih jauh dari sempurna. Semoga kekurangan dapat disempurnakan oleh penelitian selanjutnya. Semoga tugas akhir ini dapat bermanfaat. Aamiin.

*Wassalamu'alaikum Wr. Wb.*

Malang, November 2018

Penulis

## DAFTAR ISI

<b>LEMBAR JUDUL</b> .....	<b>ii</b>
<b>LEMBAR PERSETUJUAN</b> .....	<b>iii</b>
<b>LEMBAR PENGESAHAN</b> .....	<b>iv</b>
<b>PERNYATAAN KEASLIAN TULISAN</b> .....	<b>v</b>
<b>MOTTO</b> .....	<b>vi</b>
<b>PERSEMBAHAN</b> .....	<b>vii</b>
<b>KATA PENGANTAR</b> .....	<b>viii</b>
<b>DAFTAR ISI</b> .....	<b>x</b>
<b>DAFTAR GAMBAR</b> .....	<b>xii</b>
<b>DAFTAR TABEL</b> .....	<b>xiii</b>
<b>ABSTRAK</b> .....	<b>xiv</b>
<b>BAB I PENDAHULUAN</b> .....	<b>1</b>
1.1 Latar Belakang.....	1
1.2 Rumusan Masalah.....	5
1.3 Tujuan Penelitian.....	5
1.4 Manfaat Penelitian.....	5
1.5 Batasan Masalah.....	6
<b>BAB II KAJIAN PUSTAKA</b> .....	<b>7</b>
2.1 Penelitian Terkait.....	7
2.2 Landasan Teori.....	10
2.2.1 Plagiarisme.....	10
2.2.2 Jurnal.....	11
2.2.3 Jurnal Online.....	13
2.2.4 Web Crawler.....	15
2.2.5 Cosine Similarity.....	16
<b>BAB III METODOLOGI PENELITIAN</b> .....	<b>20</b>
3.1 Prosedur Penelitian.....	20
3.2 Pengumpulan Data.....	21
3.3 Perancangan Sistem.....	22
3.4 Desain Sistem.....	26
3.5 Pengujian Sistem.....	34

<b>BAB IV IMPLEMENTASI DAN PENGUJIAN SISTEM .....</b>	<b>35</b>
4.1 Implementasi.....	35
4.1.1 Implementasi Perangkat Lunak .....	35
4.1.2 Implementasi Perangkat Keras .....	37
4.1.3 Analisis Kebutuhan.....	38
4.1.4 Implementasi Struktur Tabel .....	44
4.1.5 Implementasi Interface .....	49
4.2 Pengujian.....	63
4.2.1 Kasus dan Hasil Pengujian .....	63
4.3 Integrasi Islam.....	68
<b>BAB V PENUTUP.....</b>	<b>71</b>
5.1 Kesimpulan .....	71
5.2 Saran .....	71
<b>REFERENSI.....</b>	<b>72</b>
<b>LAMPIRAN.....</b>	<b>75</b>

## DAFTAR GAMBAR

Gambar 2.1 Arsitektur Web Crawler .....	16
Gambar 3. 1 Prosedur Penelitian.....	20
Gambar 3. 2 Perancangan Sistem.....	23
Gambar 3. 3 Context Diagram .....	24
Gambar 3. 4 CDM.....	25
Gambar 3. 5 PDM .....	26
Gambar 3. 6 Desain Sistem.....	27
Gambar 3. 7 Flowchart proses perhitungan Algoritma.....	29
Gambar 4.1 Tampilan Login.....	50
Gambar 4. 2 Tampilan Pengguna.....	50
Gambar 4. 3 Tampilan Kata Dasar.....	51
Gambar 4. 4 Tampilan Stoplist .....	52
Gambar 4. 5 Tampilan Referensi .....	52
Gambar 4. 6 Tampilan <i>Cosine Similarity</i> .....	53
Gambar 4. 7 Tampilan Mahasiswa.....	54
Gambar 4. 8 Tampilan Menu Grabbing dan Upload Jurnal.....	54
Gambar 4. 9 Tampilan Menu Upload Jurnal.....	55
Gambar 4. 10 Tampilan Jurnal yang Sudah di Upload .....	56
Gambar 4. 11 Tampilan Menghapus Jurnal .....	57
Gambar 4. 12 Tampilan Grabbing Jurnal.....	57
Gambar 4.13 Tampilan Dokumen Hasil Grabbing Data.....	59
Gambar 4. 14 Tampilan Upload Jurnal dan Konversi PDF .....	59
Gambar 4. 15 Tampilan Hasil Grabbing Data .....	60
Gambar 4. 16 Tampilan Dokumen Repository .....	61
Gambar 4. 17 Tampilan Perhitungan Algoritma.....	61

## DAFTAR TABEL

Tabel 3. 1 Perhitungan manual <i>Cosine Similarity</i> .....	31
Tabel 4.1 Kebutuhan Fungsional .....	38
Tabel 4.2 Kebutuhan Non Fungsional .....	41
Tabel 4.3 Struktur Tabel <i>tb_grabbingdata</i> .....	43
Tabel 4.4 Struktur Tabel <i>tb_dokumenrepository</i> .....	44
Tabel 4.4 Struktur Tabel <i>tb_mahasiswa</i> .....	45
Tabel 4.5 Struktur Tabel <i>tb_persentase_similarity_diterima</i> .....	45
Tabel 4.6 Struktur Tabel <i>tb_cosinesimilarity</i> .....	46
Tabel 4.7 Struktur Tabel <i>tb_similarityperpublisher</i> .....	46
Tabel 4.8 Struktur Tabel <i>tb_upload</i> .....	47
Tabel 4.9 Struktur Tabel <i>tb_referensi</i> .....	47
Tabel 4.10 Struktur Tabel <i>tb_stoplist</i> .....	47
Tabel 4.11 Struktur Tabel <i>tb_katadasar</i> .....	48

## ABSTRAK

Rito Putriwana Pratama. 2018. Judul: “**Aplikasi Deteksi Plagiarisme Menggunakan Metode *Cosine Similarity***”. Skripsi. Jurusan Teknik Informatika. Fakultas Sains dan Teknologi. Universitas Islam Negeri Maulana Malik Ibrahim Malang.

Pembimbing : (I) Dr. Muhammad Faisal, M.T (II) Ajib Hanani, M.T

Kata Kunci : Plagiarisme, *Cosine Similarity*

---

Plagiarisme merupakan tindakan mengambil gagasan, mengambil hasil riset, mengakuisisi hasil riset, dan meringkas suatu tulisan tanpa menyebutkan sumbernya. Metode *cosine similarity* merupakan salah satu metode yang dapat digunakan untuk menghitung nilai kemiripan antar dokumen. Tahapan yang dilakukan sistem untuk menghasilkan nilai kemiripan antar dokumen yaitu dengan membandingkan dokumen jurnal yang di upload dengan dokumen repository yang diperoleh dari hasil grabbing data DOAJ dan tersimpan di database. Dalam perhitungan metode yang dilakukan akan diperoleh presentase nilai kemiripan antar dokumen. Setelah itu akan dihitung kembali untuk mencari nilai kemiripan dokumen jurnal antar publisher yang ada di dokumen repository. Berdasarkan skenario uji coba yang dilakukan dengan menghitung jumlah dokumen relevan terambil dibagi dengan jumlah dokumen yang ada dalam database kemudian dikali 100%, maka diperoleh nilai *recall* pada Aplikasi Deteksi Plagiarisme Menggunakan Metode *Cosine Similarity* yaitu 13%. Sedangkan untuk memperoleh nilai *precision* dilakukan skenario pengujian dengan menghitung jumlah dokumen relevan terambil dibagi dengan jumlah dokumen relevan dalam pencarian kemudian dikali 100% diperoleh hasil 8%.

## ABSTRACT

Rito Putriwana Pratama. 2018. Title: "**Application of Detection of Plagiarism Using Methode of *Cosine Similarity***". Thesis. Informatics Engineering. Faculty of Science and Technology. State Islamic University of Maulana Malik Ibrahim Malang.

Advisors: (I) Dr. Muhammad Faisal, M.T (II) Ajib Hanani, M.T

Keywords: Plagiarism, *Cosine Similarity*

---

Plagiarism is an act of taking ideas, taking research results, acquiring research results, and summarizing a writing without mentioning the source. The method of cosine similarity is one of the methods that can be used to calculate similarity among documents. The system stages to produce similarities among documents are comparing the journal documents uploaded with the repository documents obtained from the results of data grabbing of DOAJ and stored in the database. In the calculation method performed will get the percentage of similarity among the documents. After that it will be recalculated to find the similarity of journal documents among publishers in the repository document. Based on the trial scenario performed by calculating the number of relevant documents divided by the number of documents in the database then multiplied by 100%, then it would be obtained the recall value on the Application of Plagiarism Detection Using the Cosine Similarity Method was 13%. While to obtain precision value, a test scenario was done by calculating the number of relevant documents divided by the number of relevant documents in the search then multiplied by 100% and it was obtained the result of 8%.

### الملخص

ربطو فوتريوانا براتاما. 2018. العنوان : "التطبيق كشف الانتحال باستخدام طريقة قوسين التشبه".  
 البحث الجامعي قسم الهندسة المعلوماتية. كلية العلوم والتكنولوجيا. جامعة مولانا مالك الإسلامية الحكومية مالانج.  
 المشرف (I) : الدكتور محمد فيصل، الماجستير (II) عجيب حناني الماجستير  
 كلمات البحث : الانتحال، قوسين التشبه

الانتحال هو فعل أخذ الأفكار، أخذ نتائج البحث، الحصول على نتائج البحث، وتلخيص المقال بدون ذكر المصدر. طريقة قوسين التشبه هي إحدى الطرق التي يمكن استخدامها لحساب قيم التشابه بين المستندات. الخطوات المستخدمة لتوليد تشابه القيم بين المستندات من خلال مقارنة الوثائق التي تحميلها المجالات مع مستودع الوثائق المحسولة عليها من DOAJ المخزونة على البيانات في قاعدة البيانات. عند حساب الطريقة التي تنفيذها سيحصل على نسبة من قيم التشابه بين المستندات. ثم بعد ذلك إعادة حساب للعثور على قيمة التشابه من وثيقة الناشر بين المجلة في الوثيقة المستودعة. على أساس سيناريو الاختبار أجري من خلال حساب العدد من الوثائق ذات الصلة رسمها مقسوما على الوثائق في قاعدة البيانات ثم تضاعفت بنسبة 100%، القيمة المحسولة عليها من قيمة التذكير في التطبيق كشف الانتحال باستخدام طريقة قوسين التشبه 13%. في حين أن الحصول على سيناريو قيمة الدقة أجريت بحساب العدد من الوثائق ذات الصلة رسمها مقسوما على العدد من الوثائق ذات الصلة في البحث ثم تضاعفت بنسبة 100% محسولة نتيجة ل8%.



# BAB I

## PENDAHULUAN

Pada bab ini diuraikan secara singkat mengenai latar belakang masalah, rumusan masalah, tujuan penelitian, manfaat penelitian, dan batasan masalah.

### 1.1 Latar Belakang

Perkembangan dunia teknologi menjadi sangat pesat ke arah serba digital. Era digital telah membuat manusia memasuki gaya hidup baru yang tidak bisa dilepaskan dari perangkat serba elektronik. Teknologi menjadi alat yang sangat membantu kebutuhan manusia untuk melakukan kegiatan sehari-hari dengan sangat mudah, cepat, dan praktis. Begitu pentingnya peran teknologi inilah yang membawa peradaban manusia memasuki ke era serba digital. Era digital membawa berbagai dampak positif yang dapat digunakan sebaik-baiknya, namun juga memiliki dampak negatif sehingga menjadi tantangan di era digital. Era digital memasuki berbagai bidang seperti pendidikan, politik, ekonomi, sosial budaya, pertahanan, keamanan, dan teknologi informasi.

Dokumen digital merupakan dampak dari adanya perkembangan dunia teknologi di era serba digital seperti saat ini. Dalam bidang pendidikan, salah satu dampak dari perkembangan dunia teknologi yaitu adanya jurnal *online*. Jurnal *online* merupakan dokumen digital yang sangat dibutuhkan dalam semua bidang, baik pendidikan, politik, ekonomi, dll. Jurnal sendiri merupakan bagian dari jenis terbitan berseri yang ada di perpustakaan. Jurnal berisi koleksi dan terbitan atau transmisi mengenai berita dan hasil-hasil penelitian mengenai media. Jurnal terdapat dua format yaitu tercetak dan digital. Jurnal *online* merupakan versi digital dari jurnal cetak yang sering dijumpai di perpustakaan. Jurnal *online*

tersedia melalui email, web atau akses internet. Jurnal cetak dan jurnal digital memiliki kesamaan yaitu sama sama dalam terbitan berseri, sedangkan perbedaannya terdapat pada bahan baku yang digunakan, yaitu jika jurnal cetak berbahan baku kertas sedangkan jurnal *online* tanpa dicetak dapat langsung dibaca secara *online* oleh pengguna.

Jurnal *online* memiliki beberapa keuntungan bagi pembaca, diantaranya yaitu mudah dibaca dimana saja tanpa membawa kertas cetakan. Selain memiliki keuntungan, jurnal *online* juga memiliki kekurangan yaitu pembaca sangat mudah untuk melakukan penjiplakan, ini disebabkan karena tidak adanya bentuk cetakan jadi mudah untuk di *copy-paste* tanpa membaca keseluruhan isi jurnal. Perilaku penjiplakan atau biasa disebut plagiat sudah sering terjadi dalam lingkungan akademisi baik sekolah maupun di perguruan tinggi. Plagiat yang dilakukan siswa atau mahasiswa sangat tidak mencerminkan sikap kreatif dan terpelajar sebagai kaum intelektual. Terdapat berbagai bentuk perilaku plagiat yang dapat dilakukan, salah satu yang paling mudah yaitu dengan meng-*copy-paste-edit* suatu isi jurnal *online*. Perilaku plagiat sendiri dapat ditemukan dalam bentuk kutipan pada sebuah dokumen (Firdaus 2003).

Plagiarisme termasuk dalam tindakan pencurian yang harus ditangani serius agar tidak berdampak menjadi kebiasaan tanpa disadari oleh pelakunya. Tindakan plagiarisme dapat menyebabkan pelaku terutama siswa atau mahasiswa menjadi bodoh bahkan dimasukkan kedalam golongan pelanggaran ketentuan dan peraturan pemerintah. Dengan memberi larangan untuk melakukan plagiarisme, diharapkan siswa atau mahasiswa memiliki keterampilan dalam membuat tulisan

dan menghasilkan karya ilmiah sendiri dengan menggunakan kata-kata sendiri (Wibowo 2012).

Perilaku plagiat dapat dikatakan sebagai pelanggaran kode etik karya ilmiah. Di dalam Al-Qur'an terdapat banyak ayat yang menjelaskan tentang dilarangnya memakan harta sesama, dalam kasus ini yaitu perilaku plagiat, seperti pada Surat An-Nisa' Ayat 29 berikut ini :

يَا أَيُّهَا الَّذِينَ آمَنُوا لَا تَأْكُلُوا أَمْوَالَكُمْ بَيْنَكُمْ بِالْبَاطِلِ إِلَّا أَنْ تَكُونَ تِجَارَةً  
عَنْ تَرَاضٍ مِنْكُمْ وَلَا تَقْتُلُوا أَنْفُسَكُمْ إِنَّ اللَّهَ كَانَ بِكُمْ رَحِيمًا

*“Hai orang-orang yang beriman, janganlah kamu saling memakan harta sesamamu dengan jalan yang batil, kecuali dengan jalan perniagaan yang berlaku dengan suka sama-suka di antara kamu. Dan janganlah kamu membunuh dirimu; sesungguhnya Allah adalah Maha Penyayang kepadamu.” (Q.S An-Nisa' : 29).*

Mendeteksi plagiat dapat dilakukan dengan cara manual meski tidak efektif karena harus memeriksa sebuah dokumen dengan ribuan dokumen lainnya dan menafsirkan gaya penulisannya. Cara mudah untuk mendeteksi plagiat yaitu dengan penggunaan *search engine* atau mesin pencarian dengan memasukkan kata kunci tema dokumen dan membiarkan mesin pencarian menemukan dokumen yang dijiplak (Firdaus 2003). *Search engine* sendiri merupakan program komputer yang dapat membantu pengguna untuk menemukan informasi yang relevan dengan kebutuhan. Masukan dari mesin pencarian yaitu berupa kata kunci yang dibutuhkan. Dengan waktu yang relatif singkat, mesin pencarian akan memberi hasil berupa dokumen yang relevan dengan kata kunci yang dimasukkan. Hal

berikut sangat berguna apabila plagiat dilakukan pada seluruh dokumen, tetapi kurang efektif apabila plagiat dilakukan hanya pada sebagian artikel dan menggabungkan beberapa pecahan artikel lainnya.

Untuk dapat menjalankan *search engine* pendeteksi plagiarisme, pertama yang dibutuhkan yaitu sebuah cara untuk mengambil halaman web secara otomatis agar diperoleh dokumen jurnal yang *up to date*. Dalam hal ini, peneliti menggunakan metode *web crawler* karena dianggap mampu menjelajahi serta mengambil halaman web secara rekursif dan otomatis dengan mengikuti *hyperlink* yang tersedia kemudian mengambil URL yang diberikan agar dapat mengunduh dan mengambil *link* dari halaman web lain (Zuliarso 2010).

Cara yang dilakukan untuk mengambil (*filtering*) konten jurnal sebelum dicari kemiripannya yaitu dengan menggunakan *pdf extractor*. Cara ini dianggap mampu menjelajahi dan mengambil semua konten jurnal dalam bentuk pdf. Konten jurnal yang diambil diantaranya berupa metada data serta keseluruhan isi jurnal. Selain menggunakan *pdf extractor* dalam *filtering content*, dibutuhkan metode lain untuk proses pencarian serta menilai tingkat kemiripan sebuah jurnal *online*. Salah satu metode yang digunakan peneliti dalam mesin pencarian yaitu metode *Cosine Similarity*. *Cosine Similarity* merupakan metode yang digunakan untuk menghitung tingkat kemiripan (*similarity*) antar dua buah objek yang berbobot (Sugiyamta 2015). *Cosine Similarity* menggunakan dua parameter vektor. Vektor tersebut didasarkan pada jumlah kata yang mirip pada kedua dokumen teks sehingga *Cosine Similarity* tidak mengenal makna dari suatu kalimat. Peneliti menggunakan metode *Cosine Similarity* karena memiliki nilai keakuratan lebih tinggi dibandingkan dengan metode *Jaccard Similarity*. Hal

tersebut dikarenakan metode *Cosine Similarity* mempunyai konsep normalisasi panjang vektor data dengan membandingkan *N-gram* yang sejajar satu sama lain dari 2 pembanding (Nurdiana, Jumadi, and Nursantika 2016).

## 1.2 Rumusan Masalah

Berdasarkan latar belakang masalah yang telah dipaparkan sebelumnya, pertanyaan penelitian yang akan diangkat adalah bagaimana cara untuk mengukur tingkat akurasi yang ditunjukkan dengan nilai *recall* dan *precision* pada aplikasi deteksi plagiarisme menggunakan metode *cosine similarity*?

## 1.3 Tujuan Penelitian

Dalam penelitian yang dilakukan, adapun tujuan yang ingin dicapai dari penelitian ini adalah untuk mengukur tingkat akurasi deteksi plagiarisme pada dokumen jurnal dengan metode *cosine similarity* dari nilai *recall* dan *precision*.

## 1.4 Manfaat Penelitian

Manfaat yang diperoleh dari penelitian ini adalah sebagai berikut :

1. Mengetahui bahwa metode *cosine similarity* dapat digunakan untuk mendeteksi plagiarisme pada dokumen jurnal.
2. Mengetahui nilai *recall* dan *precision* pada metode pengukuran kemiripan dokumen menggunakan *cosine similarity*.

### 1.5 Batasan Masalah

Untuk menghindari pembahasan yang keluar dari meteri, maka ditentukan batasan masalah. Adapun batasan masalah dalam penelitian ini adalah dokumen repository yang digunakan untuk mengecek kemiripan diambil dari *Directory of Open Access Journal (DOAJ)*.



## BAB II

### KAJIAN PUSTAKA

Pada bab kajian pustaka akan dibahas mengenai penelitian terkait dan landasan teori yang digunakan oleh peneliti sebagai acuan melakukan penelitian.

#### 2.1 Penelitian Terkait

Pada bab ini dibahas penelitian-penelitian yang sudah dilakukan sebagai perbandingan dan acuan untuk penelitian ini.

Adik Wibowo menyatakan dalam penelitiannya mengenai plagiarisme yaitu bahwa plagiarisme merupakan perbuatan salah yang serius karena dapat menurunkan moral dan harkat pelaku serta berdampak pada sivitas akademik. Hal tersebut disebabkan tindakan mengutip tanpa izin tanpa mencantumkan sumber dan nama pengarang aslinya. Adik juga mengatakan bahwa tindakan plagiarisme bukan hanya menyebabkan pelaku menjadi bodoh, tetapi juga termasuk dalam pelanggaran ketentuan dan peraturan pemerintah. Menurut Adik tindakan plagiarisme ada yang disengaja dan tidak disengaja. Plagiarisme disengaja yaitu apabila sudah direncanakan terlebih dahulu sedangkan yang tidak sengaja yaitu apabila pelaku lupa mencantumkan nama pengarang aslinya. Dalam penelitiannya, Adik berhasil mengumpulkan beberapa kategori plagiarisme, yaitu *word by word plagiarism*, *word switch plagiarism*, *style plagiarism*, *metaphor plagiarism*, *idea plagiarism*, *self plagiarism* (Wibowo 2012).

Selanjutnya penelitian berjudul Pencarian Perangkingan Obat Tradisional Berdasarkan Gejala Penyakit Menggunakan Metode *Cosine Similarity*. Dalam penelitian ini *Cosine Similarity* digunakan untuk memberi peringkat atau

perangkingan pada jenis-jenis obat yang paling mirip. Dalam sistem pencarian ini, pengguna akan memasukkan gejala penyakit yang diderita kemudian akan diukur nilai kemiripan gejala dengan manfaat obat tradisional. Dari penelitian yang dilakukan, diperoleh nilai *recall* dan *precision* dalam pencarian perangkingan obat tradisional berdasarkan gejala penyakit menggunakan metode *cosine similarity*. Rata-rata nilai *recall* untuk lima kali uji coba adalah 91% dan rata-rata nilai *precision* untuk lima uji coba adalah 100%. Hasil tersebut membuktikan bahwa metode yang digunakan dapat diimplementasikan untuk pencarian obat menggunakan kata kunci berupa gejala penyakit (Perdana 2014).

Penelitian selanjutnya berjudul Rancang Bangun Aplikasi Pengklasifikasian Halaman Web Berdasarkan *content* Menggunakan Metode *Cosine Similarity* membuktikan bahwa metode *cosine similarity* berhasil mengklasifikasikan dokumen *web* dengan cara mengunduh seluruh halaman *web*. Dalam penelitian yang dilakukan, menghasilkan nilai akurasi pengklasifikasian menggunakan aplikasi hanya berkisar 30%. Ini berarti tingkat akurasi klasifikasi menggunakan aplikasi jauh lebih rendah daripada klasifikasi halaman *web* yang telah ditentukan oleh *web* itu sendiri. Dari hasil penelitian diketahui bahwa Algoritma *cosine similarity* sangat bergantung pada banyak dan tepatnya data *training* yang terdapat pada aplikasi (Prima and Rhosady 2013).

Selanjutnya yaitu penelitian dengan judul Perhitungan Kemiripan Dokumen Bahasa Indonesia Menggunakan Metode *Cosine Simialrity* (Studi Kasus : Abstrak Tugas Akhir Fakultas Informatika Universitas Telkom) menggunakan metode *cosine similarity* untuk menghitung nilai kemiripan antar dokumen teks. Dalam penelitian yang dilakukan Pahlevi dkk terdapat beberapa tahap salah satunya



pengumpulan data uji yang diambil dari seluruh bagian abstrak dari 341 dokumen. Untuk hasil penilaian diambil dari penilaian secara manual dan perhitungan sistem yang kemudian di korelasikan sehingga dihasilkan grafik perbandingan. Perhitungan nilai korelasi akan berada pada nilai -1 sampai 1 dengan keterangan jika mendekati 1 maka nilai korelasi bersifat linier positif yang kuat dan jika mendekati -1 maka nilai korelasi bersifat linier negatif yang kuat. Dari perhitungan pada penelitian ini diketahui nilai korelasinya yaitu 0.5729 dan dinyatakan sebagai nilai korelasi linier positif karena mendekati nilai 1. Dapat disimpulkan bahwa aplikasi yang dibangun pada penelitian ini dinilai sudah cukup baik (Pahlevi, Bijaksana, and Tech n.d.).

Metode *cosine similarity* digunakan dalam penelitian yang berjudul Perancangan dan Pembuatan Aplikasi Pencarian Informasi Beasiswa dengan Menggunakan *Cosine Similarity* untuk pencarian data beasiswa yang relevan satu sama lain yang dibutuhkan oleh *user* berdasarkan *query* yang dimasukkan. Dalam penelitian tersebut terdapat metode *Information Retrieval* (IR) yang digunakan untuk menelusuri kembali dokumen yang terdapat di *database*. Dengan menggunakan sistem *Information Retrieval* (IR), pencarian dilakukan dengan melihat tingkat kemiripan pada setiap dokumen. Untuk mendapatkan informasi beasiswa yang relevan terhadap *query* yang dimasukkan, digunakan metode *cosine similarity* dengan menghitung seberapa besar nilai kemiripan antara dokumen dan *query*. Metode *cosine similarity* memungkinkan adanya perangkingan dokumen. Perangkingan dokumen disesuaikan dengan kemiripan atau relevan *query* yang dimasukkan (Kurniawan, Solihin, and Hastarita 2014).

## 2.2 Landasan Teori

Pada bab ini diuraikan teori-teori dasar berkaitan dengan penelitian yang dilakukan sebagai dasar dalam pemecahan masalah. Teori tersebut meliputi plagiarisme, jurnal, jurnal *online*, *web crawler*, dan *cosine similarity*.

### 2.2.1 Plagiarisme

Dalam Kamus Besar Bahasa Indonesia *online* dikatakan bahwa plagiat adalah pengambilan karangan (pendapat dan sebagainya) orang lain dan menjadikannya seolah-olah karangan (pendapat dan sebagainya) sendiri, misalnya menerbitkan karya tulis orang lain atas nama dirinya sendiri. Tertulis pula dalam Peraturan Menteri Pendidikan RI Nomor 17 Tahun 2010 bahwa plagiat adalah perbuatan sengaja atau tidak sengaja dalam memperoleh kredit atau nilai untuk suatu karya ilmiah, dengan mengutip sebagian atau seluruh karya dan atau karya ilmiah pihak lain yang diakui sebagai karya ilmiahnya, tanpa menyatakan sumber secara tepat.

Selain tertulis dalam Peraturan Menteri Pendidikan RI Nomor 17 Tahun 2010, peraturan menyangkut sanksi bagi pelaku tindakan plagiarisme ditulis juga dalam UU No. 20 Tahun 2003 yang menyatakan bahwa sanksi atas tindakan plagiarisme dalam persoalan karya tulis untuk lulusan perguruan tinggi yang karya ilmiahnya digunakan untuk memperoleh gelar akademik, profesi, atau vokasi, terbukti merupakan sebuah jiplakan, maka akan dicabut gelarnya (pasal 25 ayat 2). Pelaku plagiarisme akan dikenakan hukuman pidana penjara paling lama dua tahun dan atau pidana denda paling banyak Rp. 200.000.000,- (dua ratus juta rupiah) (Zainur 2012).

Plagiarisme merupakan ide atau karya tulisan sendiri dengan menggunakan tulisan orang lain tanpa meng-*edit* tulisan aslinya (Ismail 2014). Seseorang dikatakan plagiarisme apabila mengambil gagasan, mengambil hasil riset, mengakuisisi hasil riset, dan meringkas suatu tulisan tanpa menyebutkan sumbernya. Terdapat beberapa kategori plagiarisme, yaitu *word by word plagiarism* yang berarti mengutip bagian karya seseorang lebih dari 10 halaman. *Word switch plagiarism* yaitu plagiarisme yang dilakukan dengan mengubah kalimat dengan kata-kata dan paragraf baru tetapi masih meniru gaya tulisan pengarang asli. *Metaphor plagiarism* yaitu menggunakan tulisan orang lain untuk memperjelas karya sendiri. *Idea plagiarism* merupakan tindakan plagiarisme dengan mengambil ide atau gagasan orang lain. Dan kategori yang terakhir yaitu *self plagiarism* yang berarti mengutip karya orang lain secara identik tanpa mencantumkan nama pengarang asli kemudian mengirim ke berbagai jurnal untuk di publikasikan (Wibowo 2012).

### 2.2.2 Jurnal

Jurnal adalah catatan peristiwa dari hari kehari dengan penggunaan kata jurnal memberi arti yang bervariasi, misal jurnal dalam bidang ekonomi menunjukkan sistem pembukuan rangkap, jurnal dalam bidang pelayaran diartikan sebagai *logbook* yang berarti buku untuk mencatat semua kejadian selama pelayaran. Jurnal sebenarnya merupakan representasi dari pengetahuan baru tentang perkembangan ilmu pengetahuan yang dilaksanakan secara empiris dan berisi gagasan terbaru (Rusydi 2014).

Jurnal sebagai terbitan berseri yang ditulis oleh para akademisi atau lembaga asosiasi. Jurnal bersifat ilmiah daripada majalah yang harus melalui tahap resensi atau pengujian dari para ahli sebelum diterbitkan. Jurnal berfungsi sebagai sarana menyebarkan perkembangan ilmu dimana seorang ilmuwan akan menyebarkan informasi atau ilmu yang dimiliki melalui artikel jurnal yang telah ditulis (Rusydi 2014).

Keberadaan suatu jurnal harus dilihat dari beberapa hal seperti penulis yang dikenal melalui identitas jurnal yang kuat (*author recognition through strong journal identities*) dan sumber arsip (*archival sources*). Pengawasan terhadap keberadaan satu artikel dilakukan oleh *peer review* yang bertugas menganalisa kelayakan satu artikel untuk dapat dimuat dalam satu jurnal ilmiah. Wells mengungkapkan pentingnya *peer review* sebagai alat pembuktian penelitian yang menyediakan validasi melalui sistem rujukan (Rusydi 2014).

Fungsi jurnal yaitu untuk melestarikan ilmu pengetahuan dengan menyebarkan yang telah *peer review* oleh ilmuwan. Sedangkan karakteristik jurnal berkaitan dengan teori dibandingkan buku, pembahasan lebih ringkas, referensi alternative, aplikasi dan implementasi dunia nyata (Rusydi 2014).

Kelebihan jurnal tercetak adalah sebagai berikut (Rusydi 2014) :

1. Permanen (*permanent*) dalam tampilan dan bentuk fisik.
2. Menyebar (*distributed*) di beberapa perpustakaan atau tempat.
3. Kepercayaan para ilmuwan (*scholars trust*) dan memahami sistem penerbitan (*understand the system*).
4. Memiliki pamor yang berkembang.
5. Mudah dibawa.

Selain kelebihan, jurnal tercetak juga memiliki beberapa kelemahan sebagai berikut :

1. Sistem rujukan tidak sempurna.
2. Membutuhkan biaya untuk membeli.
3. Penundaan pengumpulan untuk penerbitan.
4. Sulit terindeks.
5. Sulit untuk menelusur.
6. Anggaran perpustakaan mengecil.

### 2.2.3 Jurnal *Online*

*Online journals* sebagai versi digital dari jurnal tercetak, atau jurnal seperti dalam bentuk publikasi elektronik tanpa versi tercetaknya, tersedia melalui *email*, *web* atau akses internet. Jurnal tercetak dan *online journals* merupakan jurnal dalam cakupan terbitan berseri dan memiliki sumber informasi yang sama, perbedaan hanya terletak pada media aksesnya (Rusydi 2014).

*Online journals* yaitu jurnal yang telah dibuat dalam bentuk digitalisasi dengan memanfaatkan teknologi informasi dan komunikasi. Tujuan disediakannya *online journals* adalah untuk memudahkan anggota dalam mencari informasi maupun pengetahuan tanpa batasan waktu dan tempat. *Online journals* bisa diakses dimana saja hanya bermodalkan koneksi internet. Selain itu untuk mengantisipasi koneksi internet *down*, maka disediakan pula akses *offline* di perpustakaan (Rusydi 2014).

*Online journals* memiliki kelebihan dan kekurangan dalam pengadaannya. Adapun kelebihan *online journals* adalah sebagai berikut (Rusydi 2014):

1. Menghemat waktu.
2. Penelusurannya mudah (*easily searchable*). Lebih mudah mengetahui penelitian sebelumnya sehingga mengurangi adanya duplikasi penelitian.
3. Interaktif (*interactive*). Mudah dalam mengakses karena artikel dapat dikirim melalui *email*.
4. Aksesibilitas (*accessible*). Akses melalui internet merupakan cara yang mudah sehingga *online journals* sebagai pemecah kendala dalam penelitian yang demokratis (*breaking down the barriers to democratic research*).
5. *Links*. Memiliki kaitan antara satu artikel dengan artikel yang lain.
6. Nilai tambah (*addes value*). Dapat menggunakan animasi, *virtual reality* dan diagram matematik interaktif (*interactive mathematical charts*).
7. Murah (*inexpensive*). Mengurangi biaya percetakan jurnal.
8. Fleksibel (*flexibility*). Tidak tergantung dengan format, *printer* atau jaringan distribusi.

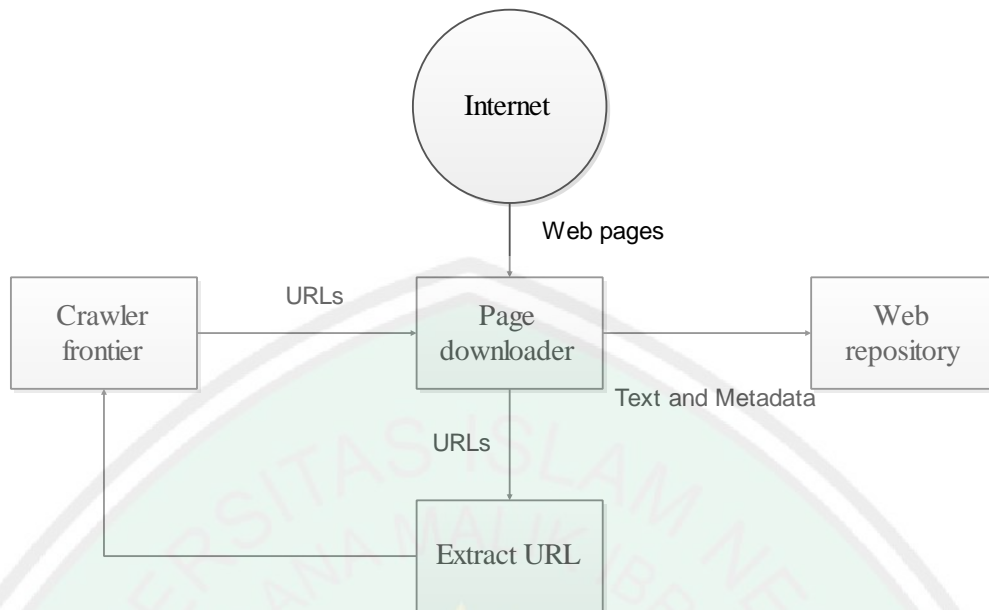
Selain memiliki kelebihan, jurnal *online* juga memiliki kekurangan sebagai berikut :

1. Kesulitan membaca di layar komputer.
2. Tidak memasukkan indeks dan abstrak.
3. Pengarsipan (*archiving*).
4. Keaslian (*authenticity*). Sumber dan otoritas serta kredibilitas pembaca harus diperhatikan.

#### 2.2.4 *Web Crawler*

*Web Crawler* merupakan program yang dapat menelusuri halaman web secara rekursif dan otomatis dengan mengikuti *hyperlink* yang telah disediakan. Pada penelitian yang sebelumnya dilakukan oleh Kumar dkk dikatakan bahwa *web crawler* yang digunakan fokus berbasis *query* yang dapat dijelajah dengan cepat. *Web Crawler* ini dinilai lebih efisien daripada *crawler* BFS sebelumnya karena dapat mengambil halaman web bersama dengan tag meta untuk menentukan relevansi halaman web (Kumar et al. 2018). Untuk menemukan halaman web yang terkait dengan topik di internet, cukup dengan memasukkan satu set kata kunci kemudian memeriksa setiap halamannya. Hasilnya akan diketahui judul, deskripsi, konten dll. *Web Crawler* hanya dapat mengikuti link *hypertext* pada internet dengan menyediakan *crawler* sebagai sarana untuk menentukan relevansi serta menemukan rute terbaik untuk menjelajahi halaman web (Rungsawang and Angkawattanawit 2005).

*Web Crawler* paling banyak digunakan pada *search engine* untuk mengumpulkan data yang ada di halaman web, sehingga *search engine* akan menampilkan data pada website yang relevan dengan kata kunci yang dimasukkan oleh pengguna. Terdapat tiga komponen utama *web crawler* yaitu *frontier*, *page downloader*, dan *web repository* (Haiyan 2017). Berikut merupakan arsitektur *web crawler*.



**Gambar 2.1** Arsitektur *Web Crawler*

Berikut merupakan proses dasar dari ketiga komponen utama *web crawler* (Haiyan 2017):

1. *Crawler Frontier* yang diawali dengan mendaftar URL yang belum dikunjungi kemudian diidentifikasi semua *hyperlink* dari halaman web tersebut.
2. *Page Downloader* yaitu mendownload halaman web sesuai dengan URL yang diterima dari *crawler frontier*.
3. *Web Repository* digunakan untuk menyimpan dan mengelola data yang ada pada halaman web dengan penyimpanan standar halaman HTML.

### 2.2.5 Cosine Similarity

Teknik dasar aplikasi menemukan situs *web* seperti *K-Similarity* telah lama dikenal dan memiliki jumlah data yang sangat banyak dan harus dikumpulkan kedalam *setting* data yang besar. Oleh karena itu dibutuhkan algoritma baru untuk



proses kemiripan dokumen dan waktu singkat untuk menemukan dokumen yang dibutuhkan. Alewiwi dkk mengusulkan *Cosine Similarity* dalam penelitiannya dimana hanya fitur yang paling penting yang akan digunakan untuk menemukan dokumen. Dalam kasus tersebut Alewiwi menggunakan tiga fase pendekatan, yaitu fase deteksi duplikat, fase istilah penting dan umum, dan fase gabungan (Alewiwi, Orencik, and Savaş 2016).

*Cosine Similarity* merupakan metode yang digunakan untuk menghitung tingkat kemiripan antar dua buah objek. Perhitungan yang dilakukan pada *cosine similarity* ini didasarkan pada *vector space similarity measure*. *Cosine similarity* menghitung antara dua buah objek yang dinyatakan dalam dua buah vektor dengan menggunakan kata kunci dari sebuah dokumen (Nurdiana, Jumadi, and Nursantika 2016).

Metode *Cosine Similarity* digunakan untuk menghitung nilai *cosinus* sudut antara dua vektor dan mengukur kemiripan antar dua dokumen dengan menggambarkan suatu kesamaan antara vektor *query* dan vektor dokumen yang menghasilkan sudut *cosinus x* diantara dua vektor tersebut. Nilai sudut *cosinus* antara dua vektor menentukan kesamaan dua buah objek yang dibandingkan dimana nilai terkecil adalah 0 dan nilai terbesar adalah 1. Nilai 0 menandakan bahwa dokumen yang dibandingkan tidak ada kemiripan, dan semakin mendekati nilai 1 maka dokumen tersebut memiliki tingkat kemiripan yang besar (Pahlevi, Bijaksana, and Tech n.d.).

Tiap vektor dalam *cosine similarity* akan merepresentasikan setiap kata dalam setiap dokumen Yng dibandingkan dan akan membentuk sebuah segitiga

sehingga dapat diterapkan hukum *cosinus* untuk menyatakan bahwa (Imbar et al. 2014) :

$$\cos(C) = \frac{a^2 + b^2 - c^2}{2ab}$$

dimana

$$a^2 = a_x^2 + a_y^2, b^2 = b_x^2 + b_y^2$$

dan

$$c^2 = (b_x - a_x)^2 + (a_y - b_y)^2$$

gantikan nilai-nilai tersebut untuk  $a$ ,  $b$ , dan  $c$  sehingga didapatkan :

$$\cos C = \frac{a_x b_x + a_y b_y}{\sqrt{a_x^2 + a_y^2} \times \sqrt{b_x^2 + b_y^2}}$$

dokumen bisa dikatakan identik jika sudutnya adalah nol derajat ( $0^\circ$ ) dan kesamaannya satu (1), sedangkan dokumen dikatakan tidak identik jika sudut  $90^\circ$  dan kesamaan adalah nol (0).

Untuk menentukan jenis plagiarisme, ada 5 penilaian persentase *similarity* (Alamanda et al. 2016):

1. Hasil uji 0% berarti kedua dokumen tersebut benar-benar berbeda baik dari segi isi dan kalimat secara keseluruhan
2. Hasil uji kurang dari 15% berarti kedua dokumen tersebut hanya mempunyai sedikit kesamaan
3. Hasil uji 15% sampai 50% berarti menandakan dokumen tersebut termasuk plagiat tingkat sedang
4. Hasil uji lebih dari 50% berarti dapat dikatakan bahwa dokumen tersebut mendekati plagiarism

5. Hasil uji 100% menandakan bahwa dokumen tersebut adalah plagiat karena dari awal sampai akhir mempunyai isi yang sama persis.



## BAB III

### METODOLOGI PENELITIAN

Bab metodologi penelitian memaparkan tahap-tahap kegiatan dalam melaksanakan penelitian. Penelitian ini mengambil judul Aplikasi Deteksi Plagiarisme Menggunakan Metode *Cosine Similarity*.

#### 3.1 Prosedur Penelitian

Prosedur penelitian adalah serangkaian kegiatan yang dilaksanakan oleh peneliti secara teratur dan sistematis untuk mencapai tujuan penelitian. Adapun prosedur penelitian pada penelitian ini direpresentasikan ke dalam diagram pada gambar 3.1 dibawah ini.



Gambar 3.1 Prosedur Penelitian

Berdasarkan gambar 3.1, terdapat beberapa tahap kegiatan yang dilakukan untuk menyelesaikan penelitian ini. Penelitian dimulai dari identifikasi masalah dengan menentukan pertanyaan penelitian. Tahap selanjutnya yaitu studi literatur yaitu mengumpulkan teori-teori yang mendukung penelitian, seperti melakukan proses *grabbing* dan *Cosine Similarity*. Dilanjutkan pada tahap pengumpulan data, dalam penelitian ini data yang dibutuhkan adalah jurnal untuk dijadikan dokumen repository. Tahap selanjutnya adalah perancangan sistem yaitu untuk memahami alur sistem yang akan dibuat dan mengimplementasikan *web crawler* pada pengambilan konten di setiap jurnal serta metode *cosine similarity* untuk menghitung kemiripan teks dalam proses deteksi plagiarisme. Setelah melakukan perhitungan nilai kemiripan dokumen, akan dilanjutkan tahap pengujian sistem dengan memperhatikan kebenaran sistem dengan metode yang telah digunakan. Dari semua tahap diatas, dapat dilanjutkan dengan melakukan analisis dari perhitungan metode serta sistem yang telah dibuat.

### 3.2 Pengumpulan Data

Pengumpulan data merupakan prosedur yang dilakukan peneliti untuk memperoleh informasi yang dibutuhkan guna mencapai tujuan penelitian. Dalam proses pengumpulan data, terdapat dua jenis data yaitu data primer dan data sekunder. Data primer yaitu data yang dikumpulkan dan diolah sendiri oleh peneliti berdasarkan subyek atau obyek penelitian. Data sekunder yaitu data yang didapat tidak secara langsung dari subyek atau obyek penelitian atau bisa disebut data yang didapat dari penelitian terdahulu.

Dalam penelitian ini terdapat beberapa tahap yang dilakukan peneliti untuk mendapatkan data yang dibutuhkan. Tahapan ini meliputi :

1. *Jurnal Online*

Obyek yang digunakan untuk penelitian ini yaitu jurnal *online*. Tahap pertama yaitu mengumpulkan atau mencari referensi jurnal *online* di berbagai *website* atau *blog*.

2. Konversi PDF

Konversi PDF digunakan untuk mengubah file pdf menjadi teks agar dapat dihitung nilai kemiripannya dengan dokumen repository yang terdapat di database.

3. Isi jurnal

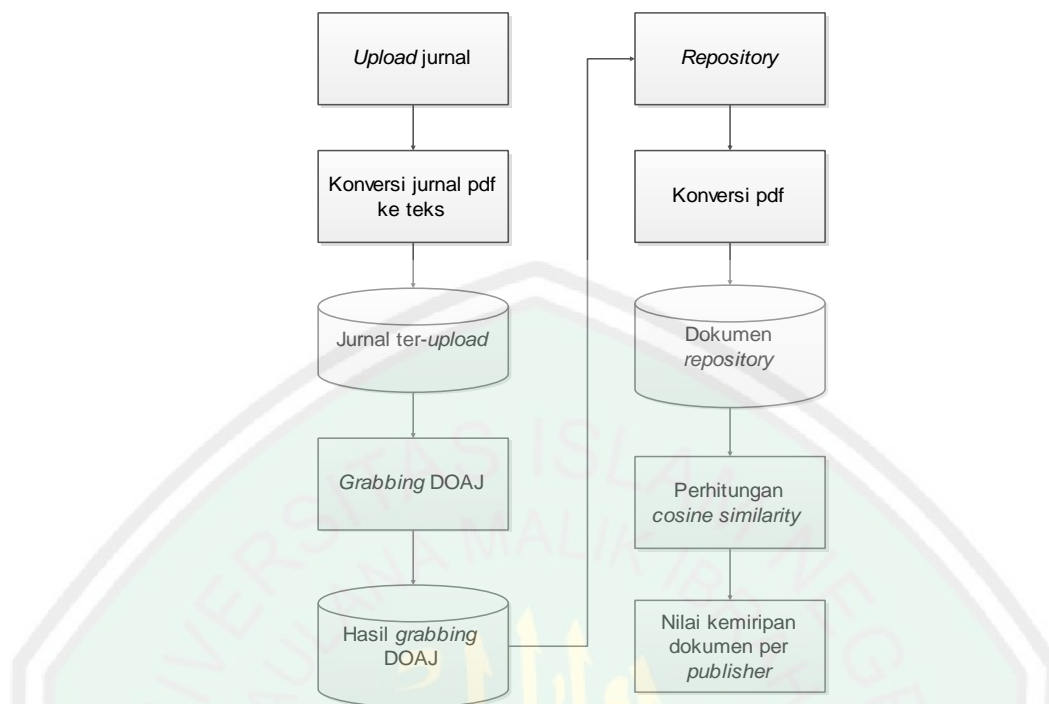
Setelah melakukan konversi pdf, diperoleh semua isi dokumen jurnal berupa teks yang akan di lakukan perhitungan kemiripannya.

4. Database

Tahap terakhir yaitu memasukkan konten yang diperoleh kedalam database yang telah dibuat. Data yang terdapat dalam database inilah yang nantinya akan diproses dalam aplikasi deteksi plagiarisme dokumen jurnal.

### 3.3 Perancangan Sistem

Perancangan sistem yaitu tahap menyusun atau mengembangkan sistem untuk menyelesaikan masalah pada objek penelitian. Pada tahap ini terdapat desain rancangan sistem yang akan memberikan gambaran yang harus dikerjakan serta bagaimana sistem memecahkan masalah deteksi plagiarisme. Perancangan sistem akan direpresentasikan pada gambar dibawah ini.



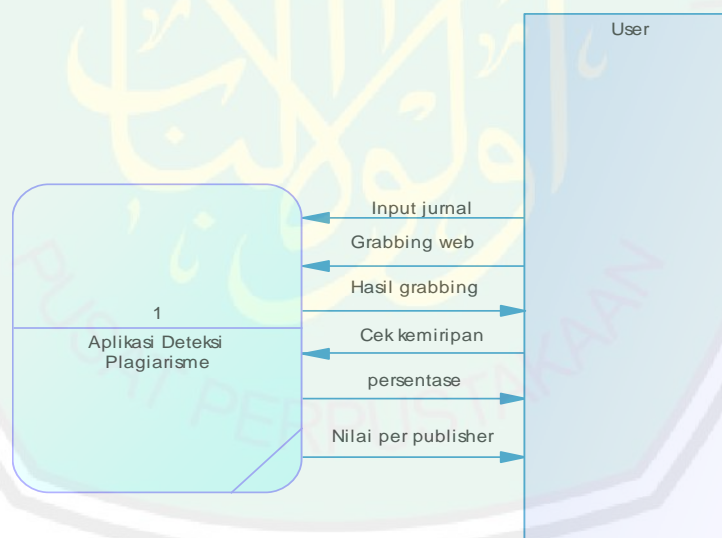
**Gambar 3.2 Perancangan Sistem**

Untuk memperoleh nilai kemiripan antar dokumen, akan dilakukan beberapa proses dalam aplikasi deteksi plagiarisme. Yang pertama yaitu proses *upload* jurnal. Pada proses ini, mahasiswa akan meng-*upload* jurnal dalam bentuk pdf untuk di cek nilai kemiripannya dengan dokumen repository. Jurnal yang di *upload* terlebih dahulu akan di konversi kedalam bentuk teks dan hasil konversi masuk kedalam database. Untuk mencari repository, maka proses selanjutnya yaitu *grabbing* jurnal yang ada di *Directory of Open Access Journal* (DOAJ). *Grabbing* bisa dilakukan sesuai dengan keinginan dengan memasukkan tema dan jumlah jurnal yang akan di ambil. Semua hasil *grabbing* DOAJ akan masuk pada database yang selanjutnya akan di proses kembali untuk dijadikan dokumen repository dengan mengkonversi jurnal dari bentuk pdf ke bentuk teks untuk dilakukan perhitungan algoritma *cosine similarity*. Perhitungan algoritma dilakukan untuk membandingkan jurnal yang di *upload* dengan jurnal yang

terdapat pada dokumen repository dan kemudian diperoleh hasil persentase kemiripannya. Proses terakhir yaitu dengan memunculkan hasil persentase kemiripan dokumen per *publisher*.

### 1. Context Diagram

*Context diagram* atau diagram konteks merupakan diagram yang terdiri dari suatu proses dan menggambarkan ruang lingkup suatu sistem. Diagram konteks merupakan level tertinggi dari DFD yang menggambarkan seluruh *input* ke sistem atau *output* dari sistem. Diagram konteks akan memberi gambaran tentang keseluruhan sistem. Diagram konteks untuk Aplikasi Deteksi Plagiarisme Menggunakan Metode *Cosine Similarity* dipaparkan pada gambar 3.3 dibawah ini.



**Gambar 3.3 Context Diagram**

Pada *context diagram* diketahui bahwa terdapat beberapa kegiatan yang dilakukan oleh aplikasi deteksi plagiarisme dan *user*. Beberapa kegiatan yang dilakukan oleh *user* adalah input jurnal, *grabbing* web, dan cek kemiripan. Sedangkan kegiatan yang dilakukan oleh aplikasi deteksi plagiarisme adalah

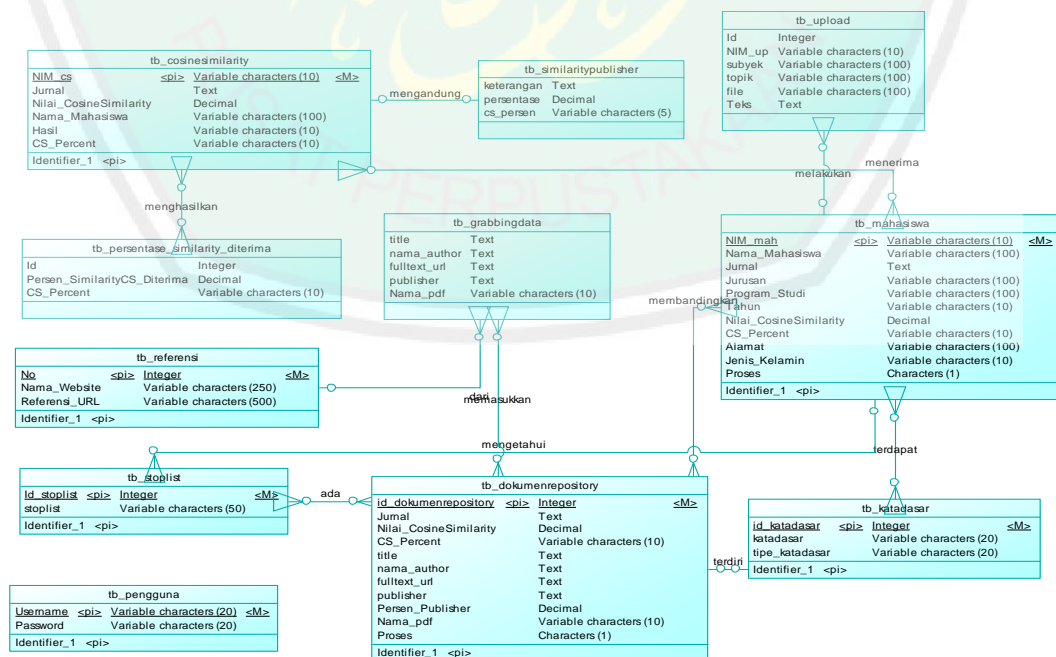


menampilkan hasil *grabbing*, menampilkan persentase kemiripan, dan menampilkan nilai persentase kemiripan per *publisher*.

## 2. CDM

*Entity Relationship Diagram Conceptual* atau biasa disebut CDM digunakan untuk menggambarkan secara detail struktur database dalam bentuk logik. Struktur ini independen terhadap semua software maupun struktur data storage tertentu yang digunakan dalam aplikasi. CDM terdiri dari objek yang tidak diimplementasikan secara langsung kedalam database yang sesungguhnya.

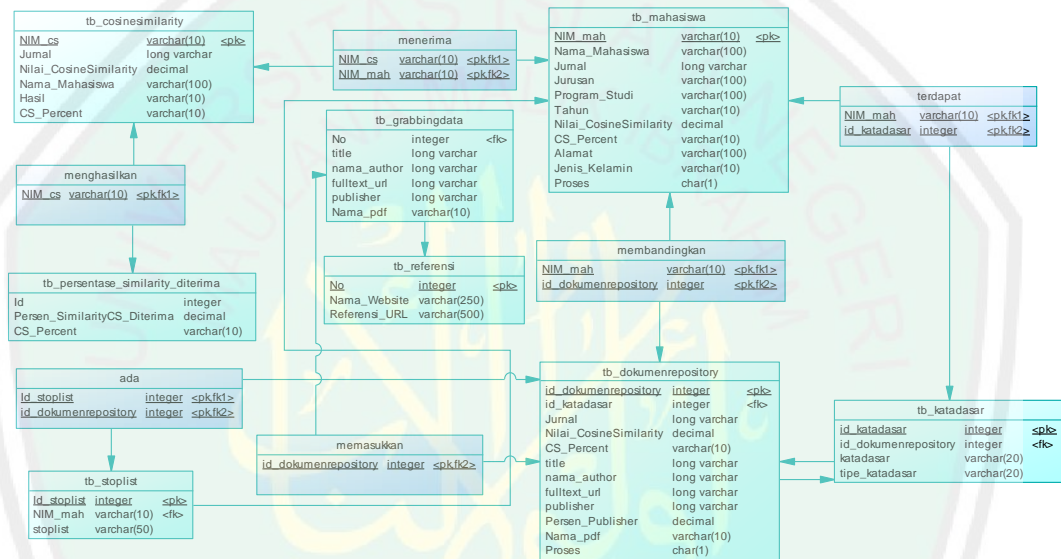
Pada CDM diketahui nama tabel di database yang digunakan untuk membuat aplikasi deteksi plagiarisme, yaitu *tb\_cosinesimilarity*, *tb\_similaritypublisher*, *tb\_upload*, *tb\_persentase\_similarity\_diterima*, *tb\_grabbingdata*, *tb\_mahasiswa*, *tb\_referensi*, *tb\_stoplist*, *tb\_pengguna*, *tb\_dokumenrepository*, *tb\_katadasar*. CDM untuk Aplikasi Deteksi Plagiarisme Menggunakan Metode *Cosine Similarity* dipaparkan pada gambar 3.4 dibawah ini.



Gambar 3.4 CDM

### 3. PDM

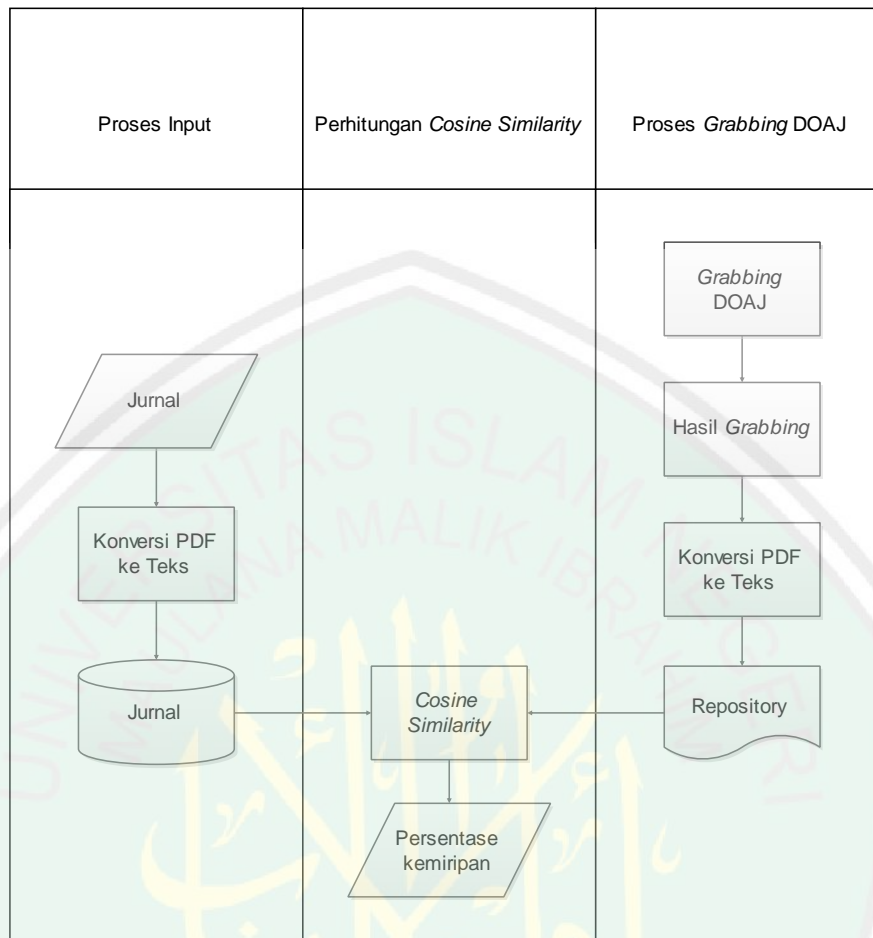
PDM merupakan gambaran secara detail database dalam bentuk fisik. Penggambaran rancangan PDM memperlihatkan struktur penyimpanan data yang benar pada database yang digunakan sesungguhnya. PDM untuk Aplikasi Deteksi Plagiarisme Menggunakan Metode *Cosine Similarity* dipaparkan pada gambar 3.6 dibawah ini.



Gambar 3.4 PDM

### 3.4 Desain Sistem

Pada tahap ini akan dijelaskan bagaimana suatu sistem akan menyelesaikan tugasnya. Desain sistem bertujuan untuk memberikan gambaran yang jelas dan rancang bangun yang lengkap kepada pemakai sistem.



**Gambar 3.6 Desain Sistem**

Ada beberapa tahap yang dilakukan oleh sistem untuk memperoleh nilai kemiripan antar dokumen. tahap tersebut adalah sebagai berikut :

1. *Input* Jurnal

Tahap pertama yaitu melakukan *input* jurnal. *User* meng-*input* jurnal yang akan di deteksi kemiripannya dengan dokumen repository yang telah tersedia di database. Jurnal yang di *upload* merupakan jurnal dalam bentuk pdf.

2. Konversi PDF ke teks

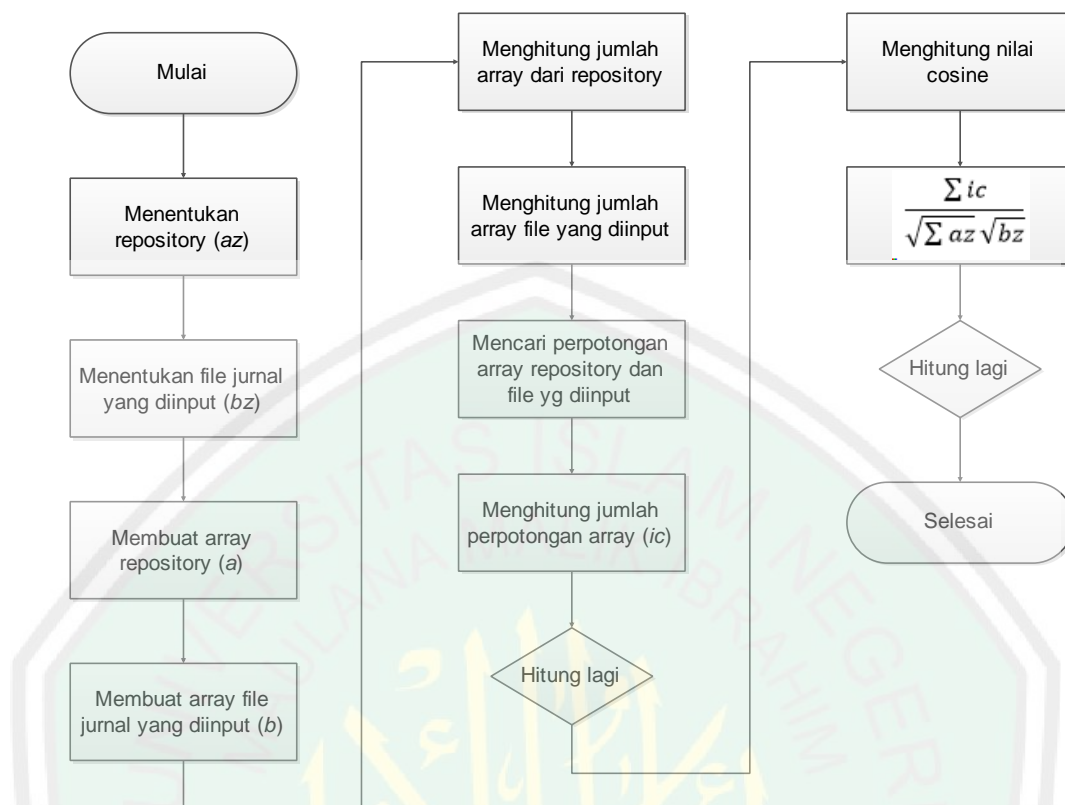
Jurnal yang telah di *upload* akan di konversi dan diubah kedalam bentuk teks agar bisa di hitung nilai kemiripannya. Hasil konversi pdf ke teks dimasukkan dalam database yang sudah disediakan.

3. *Grabbing* DOAJ

*Grabbing* DOAJ dilakukan untuk menjelajah dan mengambil halaman web yang dibutuhkan dengan menelusuri *hyperlink* yang terdapat di DOAJ. Dari proses *grabbing* DOAJ akan didapatkan file jurnal dalam bentuk pdf yang kemudian di konversi kembali kedalam bentuk teks dan disimpan di database sebagai dokumen repository yang akan menjadi dokumen pembanding untuk nilai kemiripan. Proses *grabbing* DOAJ dilakukan dengan memasukkan kata kunci yang ingin dicari dan jumlah jurnal yang diinginkan.

4. Hitung nilai kemiripan

Tahap selanjutnya yaitu menghitung nilai *cosine similarity* untuk mendapatkan persentase nilai kemiripan antara dokumen jurnal dan dokumen repository yang telah didapat.



**Gambar 3.7 Flowchart proses perhitungan Algoritma**

Untuk mengimplementasikan *flowchart* dan rumus metode *cosine similarity* seperti diatas, maka dilakukan perhitungan menggunakan Ms. Excel dengan diketahui:

1. \$mRepository="Ini contoh Repository";
2. \$mSkripsi="Ini contoh Skripsi";
3. \$a yaitu bentuk array dari dokumen repository .
4. \$b yaitu bentuk array dari dokumen uji.
5. \$ic yaitu perpotongan antara array dokumen repository (\$a) dan array dokumen uji (\$b).

Dari beberapa data yang diketahui untuk melakukan perhitungan metode *cosine similarity*, maka dapat dipaparkan alur perhitungan seperti dibawah ini.

1. Menentukan repository (\$a).

Langkah awal untuk melakukan perhitungan yaitu menentukan repository yang akan digunakan sebagai sumber dokumen yang akan dibandingkan nilainya dengan dokumen uji .

2. Menentukan file jurnal yang diinput (\$b).

File jurnal yang diinput disini sebagai dokumen uji yang akan dibandingkan dengan dokumen repository yang telah ditentukan sebelumnya.

3. Membuat array dokumen repository(\$a).

Dokumen repository yang telah didapat akan dibuat dalam bentuk array agar bisa dihitung nilai kemiripannya dengan dokumen uji.

4. Membuat array file jurnal yang diinput (\$b).

Selain dokumen repository, file jurnal atau dokumen uji juga dibuat dalam bentuk array untuk proses perhitungan kemiripannya.

5. Menghitung jumlah array dokumen repository dan array dokumen uji yang sudah didapat.

6. Mencari perpotongan array.

Dari jumlah array yang sudah didapat, maka selanjutnya menghitung perpotongan array dokumen repository dan array dokumen uji.

7. Menghitung jumlah perpotongan array.

8. Menghitung nilai *cosine similarity*

Apabila telah diketahui jumlah perpotongan array dokumen repository dan dokumen uji, maka dilanjutkan dengan menghitung nilai *cosine similarity* dengan menggunakan rumus sebagai berikut.

$$\frac{|X \cap Y|}{|X|^{1/2} \cdot |Y|^{1/2}}$$

(Thada and Jaglan 2013)

Berikut merupakan proses perhitungan metode *cosine similarity* menggunakan

Microsoft

Excel.



Tabel 3.1 Perhitungan Manual Algoritma

\$mRepository="Ini contoh Repository";				len/count		\$mSkripsi="Ini contoh Skripsi";			
\$az	buatarray	\$a	ini contoh repository	21	\$ac	\$bz	buatarray	\$b	ini contoh skripsi
	arr[]		n	i	i+n		arr[]		n
		VALUE	2	0	2			VALUE	2
		Ini	3	1	4			ini	3
		ni c	4	2	6			ni c	4
		i con	5	3	8			i con	5
		conto	6	4	10			conto	6
		contoh	7	5	12			contoh	7
		ontoh re	8	6	14			ontoh sk	8
		ntoh repo	9	7	16			ntoh skri	9
		toh reposi	10	8	18			toh skrips	10
		oh reposito	11	9	20			oh skripsi	11
		h repository	12	10				h skripsi	12
		reposito	13	11				skripsi	13
		repository	14	12				skripsi	14
		epository	15	13				kripsi	15
		pository	16	14				ripsi	16
		ository	17	15				ripsi	17
		sitory	18	16					
		itory	19	17					
		tory	20	18					
				19					
				20					
				21					
		18						15	



Jumlah array pada dokumen repository (\$a) = 18

Jumlah array pada dokumen uji (\$b) = 15

Jumlah (\$is) = FALSE

Jumlah perpotongan array \$ic = 6

Jadi nilai *Cosine Similarity* =  $\frac{6}{19.4422} \times 100\% = 30.86\%$



### 5. *Output* Persentase per *publisher*

Tahap terakhir dalam aplikasi ini yaitu menghitung persentase kemiripan dokumen per *publisher*. Pada tahap ini, hasil persentase nilai kemiripan *cosine similarity* akan dipecah lagi menjadi nilai persentase per *publisher*.

### 3.5 Pengujian Sistem

Menurut data uji dari 12 dokumen jurnal yang telah diambil dari proses *grabbing*, diketahui nilai *recall* 13% dan nilai *precision* 8%. Nilai tersebut diperoleh dari perhitungan berikut ini.

$$\begin{aligned}\text{Nilai Recall} &= \frac{\sum \text{dokumen relevan yang terambil}}{\sum \text{dokumen relevan yang ada dalam database}} \times 100\% \\ &= \frac{12}{91} \times 100\% \\ &= 13\%\end{aligned}$$

$$\begin{aligned}\text{Nilai Precision} &= \frac{\sum \text{dokumen relevan yang terambil}}{\sum \text{dokumen relevan yang ada dalam pencarian}} \times 100\% \\ &= \frac{1}{12} \times 100\% \\ &= 8\%\end{aligned}$$

*Recall* dan *precision* digunakan sebagai parameter untuk mengetahui akurat tidaknya hasil dari proses perhitungan kemiripan yang dilakukan.

### Evaluasi

Evaluasi dalam proses perhitungan pada penelitian ini membuktikan bahwa metode *cosine similarity* efektif digunakan dalam proses pencarian. Hal tersebut dibuktikan dengan hasil perhitungan nilai *recall* dan *precision* yang telah dilakukan.

## BAB IV

### PENGUJIAN DAN UJI COBA SISTEM

#### 4.1 Implementasi

Implementasi merupakan tahap penerapan sistem yang akan dilakukan jika sistem disetujui sebagai program yang telah dibuat pada tahap perancangan. Selain itu, implementasi sistem merupakan sebuah proses pembuatan dan penerapan sistem secara utuh baik dari sisi perangkat keras maupun perangkat lunaknya. Implementasi sistem bertujuan untuk menerapkan perancangan yang telah dilakukan terhadap sistem, sehingga dapat menghasilkan perangkat lunak yang sesuai dengan kebutuhan.

##### 4.1.1 Implementasi Perangkat Lunak

Untuk mendukung aplikasi agar berjalan dengan optimal, maka dibutuhkan perangkat lunak pengolahan data, adapun perangkat lunak yang digunakan untuk mendukung pembuatan aplikasi itu adalah sebagai berikut :

1. PHP

PHP (*Hypertext Preprocessor*) merupakan salah satu bahasa pemrograman *open source* berbasis kode-kode (*script*) yang digunakan untuk pengembangan *web* dan mengolah suatu data untuk dikirim kembali ke *web browser* menjadi kode HTML. Bahasa pemrograman PHP mudah dipelajari dengan menggambarkan bahasa pemrograman yang lain, seperti *C*, *Java*, dan *Perl*. PHP merupakan bahasa pemrograman yang bekerja dalam sebuah *web server*, oleh karena itu *script* PHP akan dieksekusi pada *server side*

kemudian hasilnya dijalankan pada *web browser*. Karena PHP merupakan bahasa pemrograman yang bekerja dalam sebuah *web server*, maka sistem kerja PHP diawali dengan permintaan alamat *website* atau URL untuk menemukan sebuah alamat dari *web server*. Selanjutnya *web server* akan menampilkan kode HTML pada *browser* dan akan langsung ditampilkan pada layar *browser*.

## 2. MySQL

MySQL merupakan *database server* yang bersifat *open source* dan dapat diakses oleh berbagai macam aplikasi komputer dengan menggunakan fasilitas API (*Application Programming Interface*) yang tersedia. MySQL dapat menerima dan mengirim data dengan sangat cepat dengan menggunakan perintah standar SQL seperti memasukkan (*create*), menampilkan (*read*), mengubah (*update*), dan menghapus (*delete*) data.

## 3. XAMPP

XAMPP merupakan *tool* yang menyediakan paket perangkat lunak ke dalam satu buah paket. Dengan menginstall XAMPP maka tidak perlu lagi melakukan instalasi dan konfigurasi *web server Apache*, PHP dan MySQL secara manual. XAMPP akan menginstallasi dan mengkonfigurasikannya secara otomatis.

## 4. Sublime Text Editor

Sublime Text Editor adalah editor teks untuk berbagai bahasa pemrograman termasuk bahasa pemrograman PHP. Sublime Text mendukung *operation system* seperti Linux, Mac Os X, dan juga Windows. Sublime Text memiliki

banyak fitur seperti minimap, membuka *script* secara *side by side*, *bracket highlight*, *drag and drop* direktori ke *sidebar*.

Selain perangkat lunak pengolahan data seperti yang diatas, pembuatan aplikasi juga membutuhkan pihak ketiga atau biasa disebut dengan *Third Party*. *Third Party* merupakan program komputer yang diminta untuk menjadi perantara agar dapat menjalankan program lainnya. *Third Party* yang digunakan adalah sebagai berikut :

1. class.pdf2text.php  
merupakan *third party* yang menjadi perantara dalam mengkonversi *file pdf* menjadi teks.
2. phpMaker  
merupakan *tools* otomatis yang dapat menghasilkan kode PHP untuk membangun aplikasi web dengan sangat cepat dengan menyertakan PHP *framework*, CSS dan *Javascript Framework*.

#### 4.1.2 Implementasi Perangkat Keras

Perangkat keras yang dibutuhkan berdasarkan kebutuhan minimal yang harus dipenuhi adalah sebagai berikut :

1. Processor intel inside core i5
2. Memory (RAM) 4GB
3. Hardisk 500GB
4. Mouse, printer

### 4.1.3 Analisis Kebutuhan

Aplikasi Deteksi Plagiarisme Menggunakan Metode *Cosine Similarity* dibangun menggunakan bahasa pemrograman PHP. Aplikasi ini menggunakan bahasa pemrograman PHP karena dinilai sangat cocok dalam pembuatan website dinamis. Untuk analisis kebutuhan dalam pembuatannya, akan diuraikan sebagai berikut.

#### a. Kebutuhan Fungsional

Kebutuhan fungsional merupakan penjelasan tentang kebutuhan-kebutuhan sistem yang berisi proses-proses yang akan dilakukan. Dalam hal ini akan dipaparkan bagaimana sistem bereaksi pada input tertentu dan bagaimana perilaku sistem pada situasi tertentu. Kebutuhan fungsional pada Aplikasi Deteksi Plagiarisme Menggunakan *Cosine Similarity* adalah sebagai berikut :

1. Mahasiswa Yang Boleh Upload
2. Grabbing dan Upload Jurnal Yang Akan Dideteksi
3. Upload Jurnal dan Konversi PDF ke Teks
4. Hasil Grabbing Data
5. Dokumen Repository
6. Perhitungan Algoritma Cosine Similarity
7. Hasil Cosine Similarity
8. Similarity Per Publisher

Masing-masing kebutuhan fungsional pada Aplikasi Deteksi Plagiarisme Menggunakan *Cosine Similarity* akan dipaparkan pada tabel berikut.

Tabel 4.1 Kebutuhan Fungsional

Komponen	Spesifikasi	Penanggung Jawab	Waktu Pengerjaan	Tempat	Teknis Pengadaan
Mahasiswa Yang Boleh Upload	Berisi data mahasiswa yang telah terdaftar dan dapat melakukan proses upload jurnal yang akan dihitung nilai kemiripannya	Pengelola Aplikasi Deteksi Plagiarisme	Saat pembangunan aplikasi	Kampus UIN Malang	Dibangun oleh Rito Putriwana Pratama
Grabbing dan Upload Jurnal Yang Akan Dideteksi	Berisi menu-menu yang digunakan untuk melakukan proses grabbing data dan upload jurnal	Pengelola Aplikasi Deteksi Plagiarisme	Saat pembangunan aplikasi	Kampus UIN Malang	Dibangun oleh Rito Putriwana Pratama
Upload Jurnal dan Konversi PDF ke Teks	Berisi data jurnal pdf yang telah ter-upload dan telah di konversi menjadi teks	Pengelola Aplikasi Deteksi Plagiarisme	Saat pembangunan aplikasi	Kampus UIN Malang	Dibangun oleh Rito Putriwana Pratama
Hasil Grabbing Data	Berisi data yang berhasil diambil melalui proses grabbing data	Pengelola Aplikasi Deteksi Plagiarisme	Saat pembangunan aplikasi	Kampus UIN Malang	Dibangun oleh Rito Putriwana Pratama

<b>Komponen</b>	<b>Spesifikasi</b>	<b>Penanggung Jawab</b>	<b>Waktu Pengerjaan</b>	<b>Tempat</b>	<b>Teknis Pengadaan</b>
Dokumen Repository	Berisi data repository berupa jurnal yang telah di konversi kedalam bentuk teks dan dapat digunakan sebagai dokumen pembanding dalam mencari nilai kemiripan	Pengelola Aplikasi Deteksi Plagiarisme	Saat pembangunan aplikasi	Kampus UIN Malang	Dibangun oleh Rito Putriwana Pratama
Perhitungan Algoritma Cosine Similarity	Berisi proses perhitungan algoritma untuk membandingkan antar dokumen uji dengan dokumen pembanding	Pengelola Aplikasi Deteksi Plagiarisme	Saat pembangunan aplikasi	Kampus UIN Malang	Dibangun oleh Rito Putriwana Pratama
Hasil Cosine Similarity	Data persentase hasil pebandingan kemiripan	Pengelola Aplikasi Deteksi Plagiarisme	Saat pembangunan aplikasi	Kampus UIN Malang	Dibangun oleh Rito Putriwana Pratama
Similarity Per Publisher	Data persentase hasil perbandingan per publisher antar jurnal	Pengelola Aplikasi Deteksi Plagiarisme	Saat pembangunan aplikasi	Kampus UIN Malang	Dibangun oleh Rito Putriwana Pratama



**b. Kebutuhan Non Fungsional**

Kebutuhan non fungsional merupakan kebutuhan properti perilaku yang dimiliki oleh sistem. Kebutuhan non fungsional dalam aplikasi deteksi plagiarisme adalah sebagai berikut :

1. Sistem Operasi
2. Editor
3. Web browser
4. Web service
5. DBMS
6. Bahasa pemrograman

Masing-masing kebutuhan nonfungsional pada Aplikasi Deteksi Plagiarisme Menggunakan *Cosine Similarity* akan dipaparkan pada tabel berikut.

**Tabel 4.2 Kebutuhan Non Fungsional**

<b>Komponen</b>	<b>Spesifikasi</b>	<b>Penanggung Jawab</b>	<b>Waktu Pengerjaan</b>	<b>Tempat Pengerjaan</b>	<b>Teknis Pengadaan</b>
Sistem Operasi	Windows 10	Pembangun dan pengelola Aplikasi Deteksi Plagiarisme	Saat pembangunan Aplikasi Deteksi Plagiarisme	Kampus UIN Malang	Instalasi
Editor	Sublime Text 3	Pembangun dan pengelola Aplikasi Deteksi Plagiarisme	Saat pembangunan Aplikasi Deteksi Plagiarisme	Kampus UIN Malang	Instalasi
Web Browser	Google Chrome	Pembangun dan pengelola Aplikasi Deteksi Plagiarisme	Saat pembangunan Aplikasi Deteksi Plagiarisme	Kampus UIN Malang	Instalasi
Web Service	XAMPP	Pembangun dan pengelola Aplikasi Deteksi Plagiarisme	Saat pembangunan Aplikasi Deteksi Plagiarisme	Kampus UIN Malang	Instalasi

<b>Komponen</b>	<b>Spesifikasi</b>	<b>Penanggung Jawab</b>	<b>Waktu Pengerjaan</b>	<b>Tempat Pengerjaan</b>	<b>Teknis Pengadaan</b>
DBMS	MySQL	Pembangun dan pengelola Aplikasi Deteksi Plagiarisme	Saat pembangunan Aplikasi Deteksi Plagiarisme	Kampus UIN Malang	Instalasi
Bahasa Pemrograman	PHP	Pembangun dan pengelola Aplikasi Deteksi Plagiarisme	Saat pembangunan Aplikasi Deteksi Plagiarisme	Kampus UIN Malang	Instalasi

#### 4.1.4 Implementasi Struktur Tabel

Pada tabel 4.3 merupakan tabel `tb_grabbingdata`. Pengguna mengakses tabel ini untuk menelusuri halaman web DOAJ untuk mengambil dokumen jurnal yang akan dijadikan dokumen repository.

**Tabel 4.3 Struktur Tabel `tb_grabbingdata`**

Nama Tabel	Nama Field	Tipe Data	Key
tb_grabbingdata	- title	- longtext	
	- nama_author	- longtext	
	- fulltext_url	- longtext	
	- publisher	- longtext	
	- Nama_pdf	- varchar(10)	

Pada tabel 4.4 merupakan tabel `tb_dokumenrepository`. Pengguna mengakses tabel ini sebagai wadah penyimpanan dokumen repository yang akan dibandingkan dengan dokumen jurnal yang telah di *upload*.

Tabel 4.4 Struktur Tabel tb\_dokumenrepository

Nama Tabel	Nama Field	Tipe Data	Key
tb_dokumenrepository	- id_dokumenrepository	- int(11)	- PK
	- Jurnal	- longtext	
	- Nilai_CosineSimilarity	- decimal(5,2)	
	- CS_Percent	- varchar(5)	
	- title	- longtext	
	- nama_author	- longtext	
	- fulltext_url	- varchar(200)	
	- publisher	- longtext	
	- Persen_Publisher	- decimal(5,2)	
	- Nama_pdf	- varchar(10)	
	- proses	- varchar(1)	

Pada tabel 4.5 merupakan tabel tb\_mahasiswa. Pengguna mengakses tabel ini untuk menyimpan data mahasiswa atau pengguna yang *input* dokumen jurnal untuk dibandingkan dengan dokumen repository.

Tabel 4.3 Struktur Tabel tb\_mahasiswa

Nama Tabel	Nama Field	Tipe Data	Key
tb_mahasiswa	- NIM	- varchar(10)	- PK
	- Nama_Mahasiswa	- varchar(100)	
	- Jurnal	- longtext	
	- Jurusan	- varchar(100)	
	- Program_Studi	- varchar(100)	
	- Tahun	- varchar(10)	
	- Nilai_CosineSimilarity	- decimal(5,2)	
	- CS_Percent	- varchar(5)	
	- Alamat	- varchar(100)	
	- Jenis_Kelamin	- varchar(10)	
	- Proses	- char(1)	

Pada tabel 4.6 merupakan tabel tb\_persentase\_similarity\_diterima. Pengguna mengakses tabel ini untuk menentukan persentase nilai *cosine similarity* yang dapat diterima.

Tabel 4.6 Struktur Tabel tb\_persentase\_similarity\_diterima

Nama Tabel	Nama Field	Tipe Data	Key
tb_persentase_similarity_diterima	- Id	- tinyint(1)	
	- Persen_SimilarityCS_Diterima	- decimal(5,2)	
	- CS_Percent	- varchar(10)	

Pada tabel 4.7 merupakan tabel *tb\_cosinesimilarity*. Pengguna mengakses tabel ini untuk menghitung nilai *cosine similarity* dan menentukan persentase kemiripan dokumen jurnal dengan dokumen repository.

**Tabel 4.7 Struktur Tabel *tb\_cosinesimilarity***

Nama Tabel	Nama Field	Tipe Data	Key
<i>tb_cosinesimilarity</i>	<ul style="list-style-type: none"> <li>- NIM</li> <li>- Jurnal</li> <li>- Nilai_CosineSimilarity</li> <li>- Nama_Mahasiswa</li> <li>- Hasil</li> <li>- CS_Percent</li> </ul>	<ul style="list-style-type: none"> <li>- varchar(10)</li> <li>- longtext</li> <li>- decimal(5,2)</li> <li>- varchar(100)</li> <li>- varchar(10)</li> <li>- varchar(5)</li> </ul>	

Pada tabel 4.8 merupakan tabel *tb\_similarityperpublisher*. Pengguna mengakses tabel ini untuk mengetahui nilai persentase kemiripan per *publisher*.

**Tabel 4.8 Struktur Tabel *tb\_similarityperpublisher***

Nama Tabel	Nama Field	Tipe Data	Key
<i>tb_similarityperpublisher</i>	<ul style="list-style-type: none"> <li>- keterangan</li> <li>- persentase</li> <li>- cs_persen</li> </ul>	<ul style="list-style-type: none"> <li>- longtext</li> <li>- decimal(5,2)</li> <li>- varchar(5)</li> </ul>	

Pada tabel 4.9 merupakan tabel *tb\_upload*. Pengguna mengakses tabel ini untuk meng-*upload* file jurnal yang akan dibandingkan dengan dokumen repository.

**Tabel 4.9 Struktur Tabel tb\_upload**

<b>Nama Tabel</b>	<b>Nama Field</b>	<b>Tipe Data</b>	<b>Key</b>
tb_upload	- id - NIM - subyek - topik - file - Teks	- int (11) - varchar (10) - varchar (100) - varchar (100) - varchar (100) - longtext	PK

Pada tabel 4.10 merupakan tabel tb\_referensi. Pengguna mengakses tabel ini untuk mengetahui yang digunakan dalam membandingkan dokumen.

**Tabel 4.10 Struktur Tabel tb\_referensi**

<b>Nama Tabel</b>	<b>Nama Field</b>	<b>Tipe Data</b>	<b>Key</b>
tb_referensi	- No - Nama_Website - Referensi_URL	- tinyint (3) - varchar (250) - varchar (500)	PK

Pada tabel 4.11 merupakan tabel tb\_stoplist. Pengguna mengakses tabel ini untuk mengetahui kata apa saja yang masuk kedalam data *stoplist*.

**Tabel 4.11 Struktur Tabel tb\_stoplist**

<b>Nama Tabel</b>	<b>Nama Field</b>	<b>Tipe Data</b>	<b>Key</b>
tb_stoplist	- id_stoplist - stoplist	- int (10) - varchar (50)	PK



Pada tabel 4.12 merupakan tabel `tb_katadasar`. Pengguna mengakses tabel ini untuk mengetahui kata apa saja yang masuk kedalam data kata dasar.

**Tabel 4.12 Struktur Tabel `tb_katadasar`**

<b>Nama Tabel</b>	<b>Nama Field</b>	<b>Tipe Data</b>	<b>Key</b>
<code>tb_katadasar</code>	- <code>id_katadasar</code>	- <code>int (10)</code>	PK
	- <code>katadasar</code>	- <code>varchar (20)</code>	
	- <code>tipe_katadasar</code>	- <code>varchar (20)</code>	

#### 4.1.4 Implementasi *Interface*

*Interface* merupakan tampilan yang dapat melakukan interaksi antara pengguna dengan sistem, dimana *interface* dapat menerima informasi dari pengguna dan memberikan informasi kepada pengguna yang bertujuan untuk menginput pengetahuan baru ke dalam basis pengetahuan sistem pakar, menampilkan penjelasan sistem dan memberikan panduan pemakaian sistem secara menyeluruh sehingga dapat dipahami oleh pengguna.

##### 1. Tampilan Login

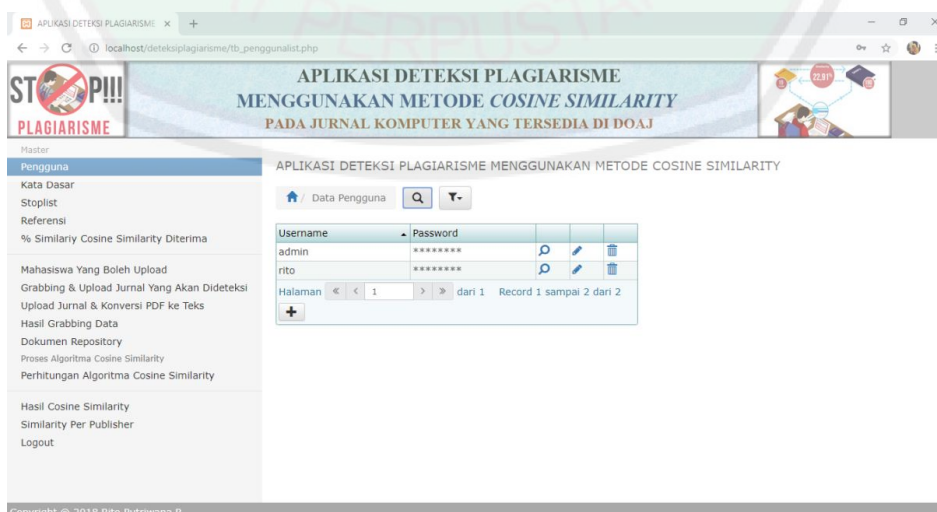
Tampilan *login* merupakan tampilan awal yang akan muncul ketika aplikasi dijalankan. Pada tampilan ini diberikan form untuk *login* pengguna. *Login* pengguna diperlukan untuk menjamin keamanan suatu sistem. Untuk melakukan proses *login*, pengguna harus memasukkan *username* dan *password* sesuai dengan yang sudah ditambahkan pada tabel pengguna. Tampilan *login* ditunjukkan pada gambar 4.1.



**Gambar 4.1 Tampilan Login**

## 2. Tampilan Pengguna

Halaman pengguna merupakan halaman yang menampilkan data pengguna aplikasi deteksi plagiarisme. Data yang akan muncul yaitu *username* dan *password* pengguna. Selain menampilkan data *username* dan *password* pengguna aplikasi, tampilan ini juga dapat berfungsi untuk menambah pengguna baru, serta menghapus dan mengedit data anggota yang sudah ada. Tampilan pengguna ditunjukkan pada gambar 4.2.



**Gambar 4.2 Tampilan Pengguna**

### 3. Tampilan Kata Dasar

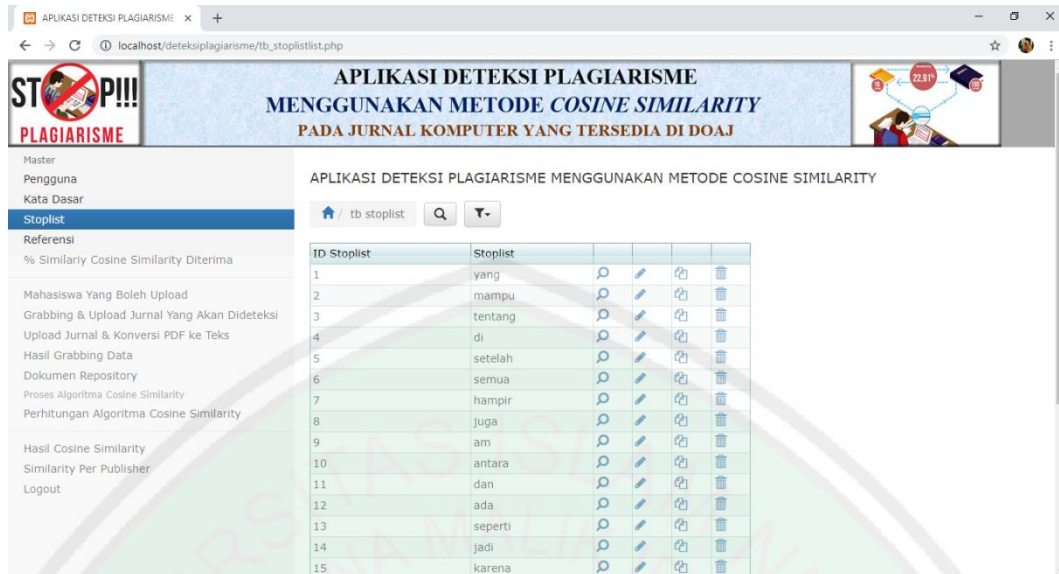
Kata dasar merupakan kata yang menjadi dasar pembentukan suatu kata dan belum mendapat imbuhan. Pada sistem, kata dasar dibutuhkan dalam melakukan proses *preprocessing*. Tampilan kata dasar ini akan menampilkan id kata dasar, kata dasar, dan jenis kata dasar sebanyak 28530 kata. Tampilan kata dasar ditunjukkan pada gambar 4.3.

ID Kata Dasar	Kata Dasar	Tipe Kata Dasar				
1	a	Nomina				
2	ab	Nomina				
3	aba	Nomina				
4	aba-aba	Nomina				
5	abad	Nomina				
6	abadi	Adjektiva				
7	abadiah	Nomina				
8	abah	Nomina				
9	abal	Adjektiva				
10	abaimana	Nomina				
11	abaka	Nomina				
12	abaktinal	Adjektiva				
13	abakus	Nomina				
14	abal-abal	Nomina				
15	aban	Nomina				

Gambar 4.3 Tampilan Kata Dasar

### 4. Tampilan *Stoplist*

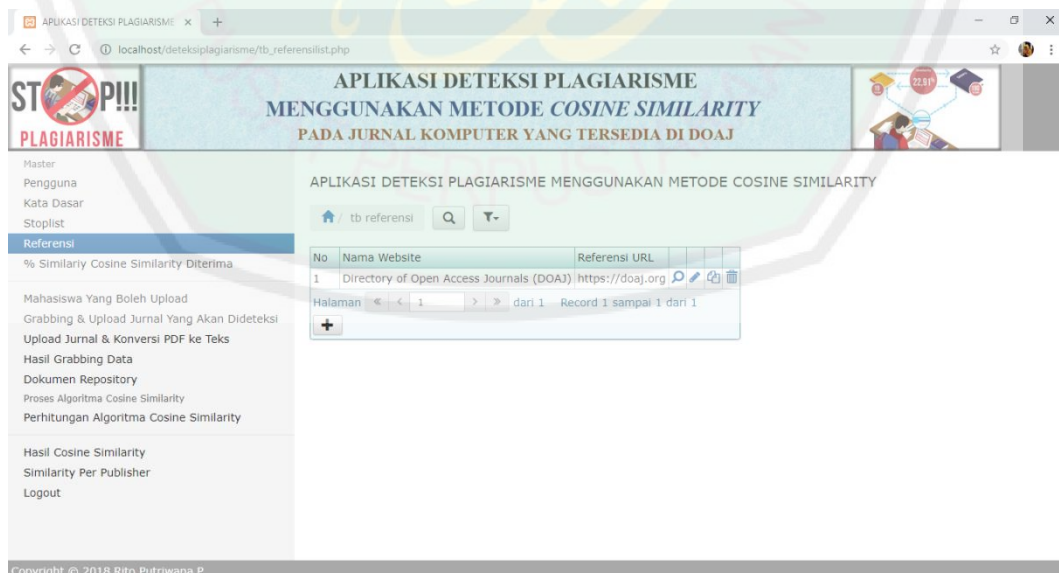
*Stoplist* merupakan pembuangan kata penghubung atau kata dalam bahasa Indonesia yang tidak diperlukan. *Stoplist* merupakan proses yang hanya menyaring kata atau *term* yang dianggap penting. Tampilan *stoplist* ini hanya menampilkan *id stoplist* dan *stoplist* sebanyak 1084 kata. Tampilan *stoplist* ditunjukkan pada gambar 4.4.



Gambar 4.4 Tampilan Stoplist

## 5. Tampilan Referensi

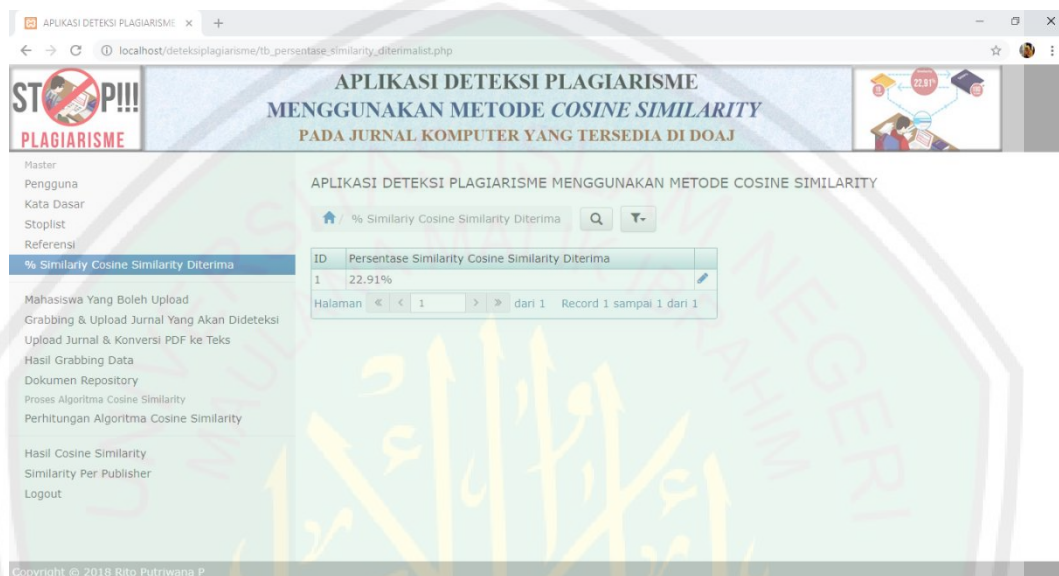
Referensi merupakan rujukan atau sumber suatu informasi. Pada tampilan referensi terdapat url suatu repository jurnal yang digunakan sebagai sumber untuk melakukan proses cek kemiripan. Referensi yang digunakan dalam aplikasi ini yaitu <https://doaj.org>. Tampilan referensi ditunjukkan pada gambar 4.5



Gambar 4.5 Tampilan Referensi

## 6. Tampilan *Cosine Similarity* Diterima

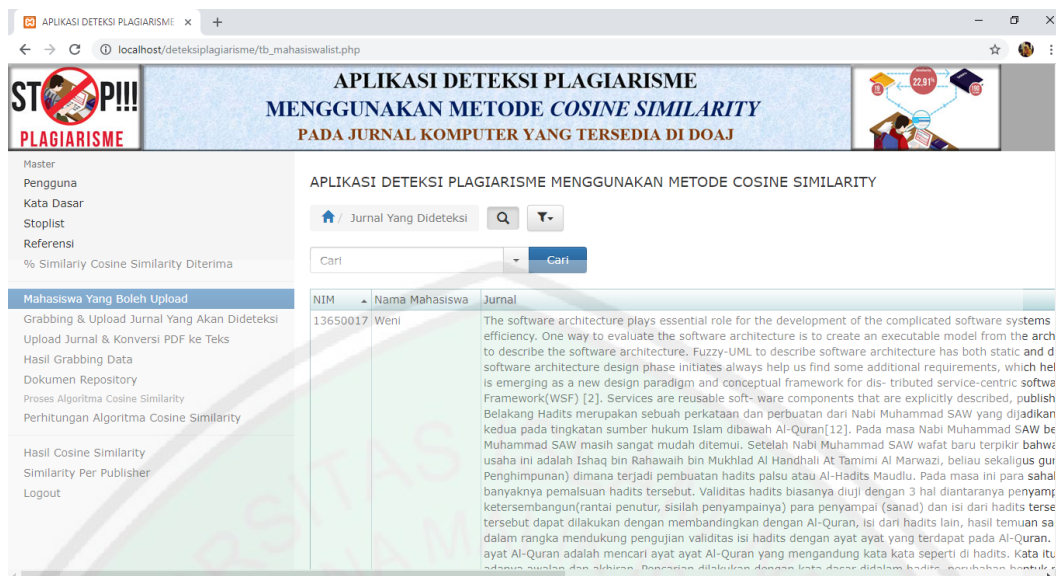
Halaman ini menampilkan batas ketentuan persentase kemiripan suatu dokumen terhadap dokumen repository yang telah disediakan. Nilai persentase dapat diubah sesuai ketentuan masing-masing pengguna aplikasi.



Gambar 4.6 Tampilan *Cosine Similarity*

## 7. Tampilan Mahasiswa

Halaman ini menampilkan data mahasiswa yang sudah meng-*upload* jurnal untuk di hitung nilai *Cosine Similarity* atau nilai kemiripannya. Nilai tersebut didapat dengan cara membandingkan jurnal yang di *upload* dengan dokumen repository yang ada di *database*. Setelah dilakukan perhitungan, akan muncul jumlah persentase kemiripan antar dokumen. tampilan mahasiswa yang boleh *upload* ditunjukkan pada gambar 4.7.



**Gambar 4.7 Tampilan Mahasiswa**

Halaman ini berisi menu-menu yang digunakan untuk upload jurnal yang akan dideteksi serta menu untuk melakukan proses grabbing data. Menu-menu yang terdapat pada halaman ini yaitu menu *home*, menu upload jurnal, menu jurnal yang sudah diupload, menu menghapus jurnal, dan grabbing jurnal.



**Gambar 4.8 Tampilan menu *Grabbing & Upload Jurnal***

## 9. Menu *Upload Jurnal*

Menu ini berisi form yang wajib diisi untuk melakukan upload jurnal. Subyek merupakan nama atau judul jurnal. Jenis file harus berupa file pdf. Nama file yaitu form untuk memilih jurnal yang akan di upload, jurnal bisa didapat dari direktori manapun. Nomor induk mahasiswa diisi dengan NIM yang sudah terdaftar di halaman mahasiswa yang boleh upload.



### 3pdf\_download.php

```

<?php
include 'pdf2text.php';
$repol = mysql_query("SELECT fulltext_url, title,
nama_author,publisher,nama_pdf,Jurnal FROM tb_dokumenrepository where
(length(nama_pdf))>0 and length(trim(Proses))=0");
while($dataRepol = mysql_fetch_array($repol))
{
    $mfulltext_url = $dataRepol['fulltext_url'];
    $mtitle = $dataRepol['title'];
    $mnama_author = $dataRepol['nama_author'];
    $mpublisher = $dataRepol['publisher'];
    $mnama_pdf = $dataRepol['nama_pdf'];
    $filedirektori1="c:/xampp5/htdocs/DeteksiPlagiarisme/Repository/" .
    trim($mnama_pdf) . ".pdf";
    $mteks = pdf2text ($filedirektori1);
    $result =str_replace("", "", $mteks);;
    $sql = "update tb_dokumenrepository set
    Jurnal='".$result."',Proses='S' where
    fulltext_url='".$mfulltext_url.'";
    mysql_query($sql);
}
?>

```

## 10. Tampilan Menu Jurnal yang Sudah Diupload

Menu tersebut menampilkan daftar jurnal yang sudah diupload termasuk subyek, jenis file, dan file yang diupload. Jurnal dapat di download melalui link yang sudah disediakan. Tampilan menu jurnal yang sudah di *upload* ditunjukkan pada gambar 4.10.



Subyek	Jenis File	File Yang Diupload
Jurnal1	pdf	1016-3521-1-PB.pdf
Jurnal2	pdf	4825-3022-1-SM.pdf
jurnal3	pdf	129-438-1-PB.pdf
jurnal4	pdf	881-1901-1-SM.pdf
jurnal5	pdf	113091003_bab1.pdf
jurnal6	pdf	10.1007%2Fs10726-013-9359-1.pdf
jurnal7	pdf	alanzi2016.pdf
jurnal8	pdf	[20] rungsawang.pdf

[Sebelumnya](#)   [Berikutnya](#)  
[Kembali ke menu upload](#)

**Gambar 4. 10 Tampilan jurnal yang sudah di upload**



## 11. Menghapus Jurnal

Pada menu menghapus jurnal, akan ditampilkan kembali jurnal-jurnal yang telah di *upload*, tetapi hanya dapat melakukan proses penghapusan file. Menu menghapus jurnal ditunjukkan pada gambar 4.11.



**Gambar 4. 11** Tampilan menghapus jurnal

## 12. Menu *Grabbing* Jurnal

Menu *grabbing* merupakan menu untuk mencari jurnal yang akan digunakan sebagai dokumen repository yang nantinya untuk menguji kemiripan dengan dokumen yang di *upload*. Menu *grabbing* ditunjukkan pada gambar 4.12.



**Gambar 4. 12** Tampilan *grabbing* jurnal

Untuk mengontrol aksi pada halaman *grabbing* jurnal, dibutuhkan proses yang dinamakan *grabbing-doaj.php*. Berikut merupakan *sourcecode* yang digunakan untuk melakukan *grabbing* data DOAJ.

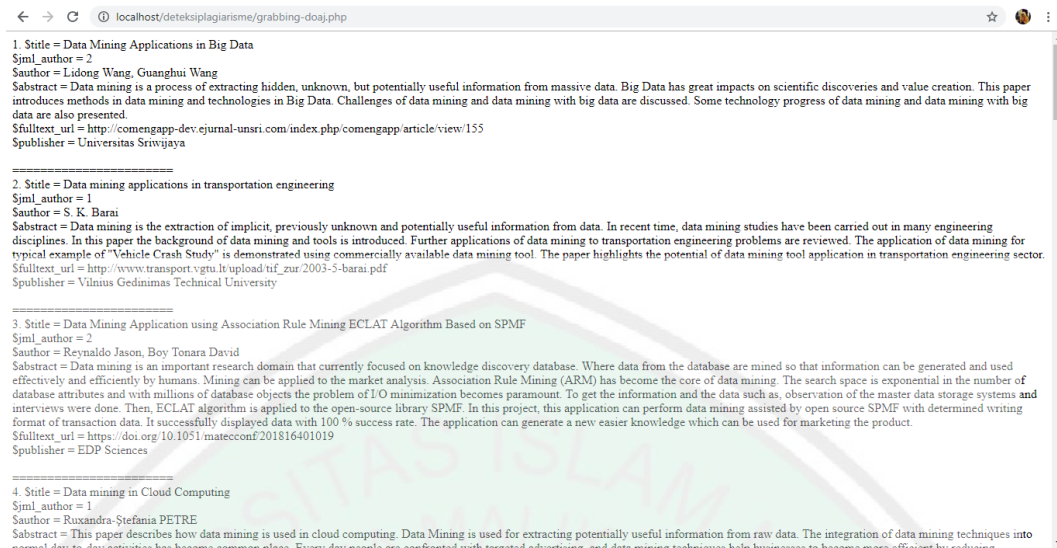
### grabbing-doaj.php

```

<?php
public function doaj(){    $ch = curl_init();
    $no = 1;
    $jml_data = count($data['results']);
    for ($i=0;$i<$jml_data;$i++){
        $title = $data['results'][$i]['bibjson']['title'];
        echo $no." \$title = ".$title;
        echo "<br/>";
        $jml_author = count($data['results'][$i]['bibjson']['author']);
        echo "\$jml_author = ".$jml_author;
        echo "<br/>";
        $author = "";
        for ($j=0;$j<$jml_author;$j++){
            if ($j == 0) {
                $author .= $data['results'][$i]['bibjson']['author'][$j]['name'];
            } else{
                $author .= $data['results'][$i]['bibjson']['author'][$j]['name'];
            }
        }
        echo "\$author = ".$author;
        echo "<br/>";
        $abstract = "";
        if (isset($data['results'][$i]['bibjson']['abstract'])) {
            $abstract = $data['results'][$i]['bibjson']['abstract'];
        }
        echo "\$abstract = ".$abstract;
        echo "<br/>";
        $fulltext_url = $data['results'][$i]['bibjson']['link'][0]['url'];
        echo "\$fulltext_url = ".$fulltext_url;
        echo "<br/>";
        $publisher = $data['results'][$i]['journal']['publisher'];
        echo "\$publisher = ".$publisher;
        echo "<br/>";
        $sql = "INSERT IGNORE INTO tb_grabbingdata
        (title,nama_author,fulltext_url,publisher) VALUES
        ('".$title."','".$author."','".$fulltext_url."','".$publisher."')";
        ;    mysql_query($sql);
        $no++;}
    echo "<pre>";
?>

```

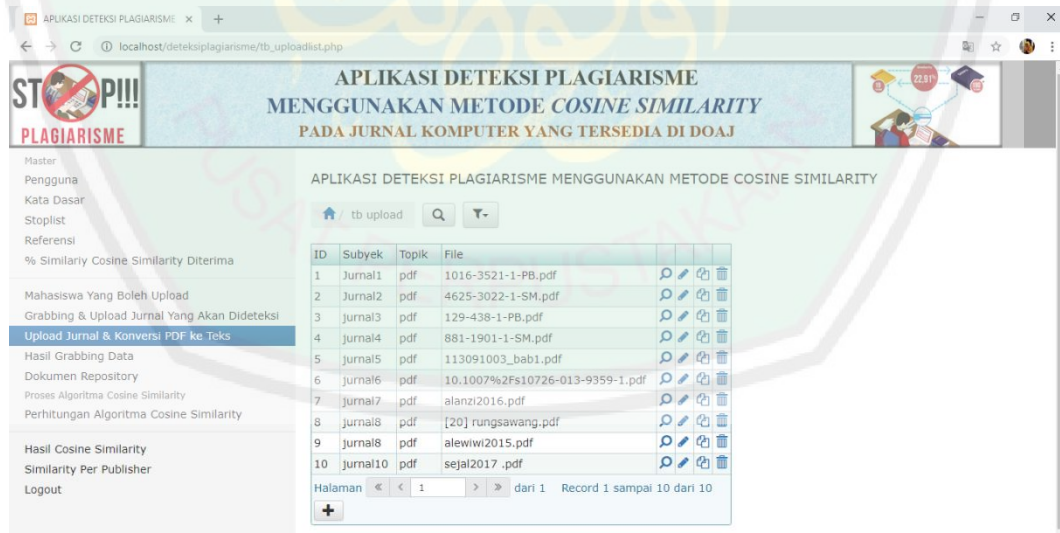
Berikut merupakan tampilan hasil proses *grabbing* data DOAJ. Sesuai dengan *sourcecode* yang digunakan, hasil *grabbing* akan menampilkan *title*, *jml\_author*, *abstract*, *fulltext*, dan *publisher* pada setiap jurnal yang terambil.



Gambar 4.13 Dokumen Hasil Grabbing Data

### 13. Tampilan Upload Jurnal & Konversi PDF ke Teks

Halaman ini berisi data jurnal yang telah di *upload* oleh pengguna. Terdapat ID, subyek, topik, file, serta aksi-aksi yang digunakan untuk mengolah jurnal. Jurnal yang di *upload* tidak boleh memiliki nama subyek dan file yang sama.



Gambar 4.14 Tampilan upload jurnal dan konversi PDF

#### 14. Tampilan Hasil *Grabbing* Data

Halaman ini menampilkan seluruh judul jurnal hasil *grabbing* yang telah dilakukan. Dari proses *grabbing*, akan diperoleh *title*, nama *author*, *fulltext URL*, *publisher*, dan nama pdf. Tidak semua hasil *grabbing* memiliki nama pdf. Tampilan hasil *grabbing* data ditunjukkan pada gambar 4.15.

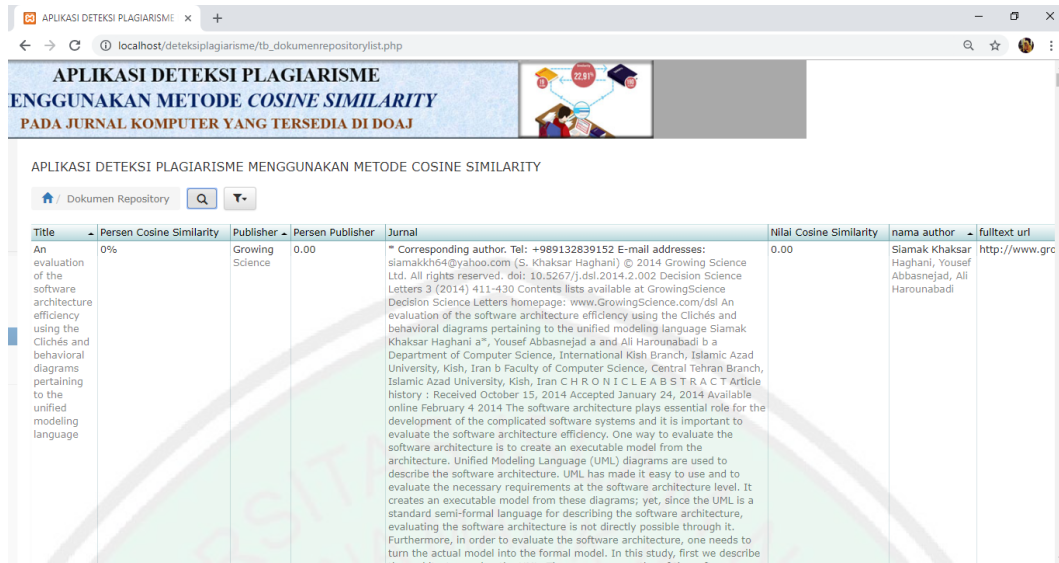


Title	Nama Author	Fulltext URL	Publisher	Nama pdf
A multi-layered software architecture model for building software solutions in an urbanized information system	Sana Guetat, Salem Dakhli, ..	<a href="http://www.sciencesphere.org/ijjspm/archive/ijjspm-010102.pdf">http://www.sciencesphere.org/ijjspm/archive/ijjspm-010102.pdf</a>	SciKA	
A Trustworthiness Evaluation Method for Software Architectures Based on the Principle of Maximum Entropy (POME) and the Grey Decision-Making Method (GDMM)	Rong Jiang	<a href="http://www.mdpi.com/1099-4300/16/9/4818">http://www.mdpi.com/1099-4300/16/9/4818</a>	MDPI AG	
An evaluation of the software architecture efficiency using the Clichés and behavioral diagrams pertaining to the unified modeling	Siamak Khaksar Haghani, Yousef Abbasnejad, Ali Harounabadi	<a href="http://www.growing-science.com/dsl/Vol3/dsl_2014_4.pdf">http://www.growing-science.com/dsl/Vol3/dsl_2014_4.pdf</a>	Growing Science	2

Gambar 4. 15 Tampilan Hasil *Grabbing* Data

#### 15. Tampilan Dokumen Repository

Halaman ini berisi tampilan data jurnal yang akan dijadikan perbandingan antara jurnal yang di *upload* dan dokumen repository. Dokumen repository diperoleh dari hasil *grabbing*, dan hanya di ambil file pdf yang memiliki nama pdf. Tampilan dokumen repository ditunjukkan pada gambar 4.15.



Gambar 4.16 Tampilan Dokumen Repository

## 16. Tampilan Perhitungan Algoritma *Cosine Similarity*

Perhitungan algoritma dilakukan dengan cara membandingkan kemiripan antara dokumen repository dan jurnal yang telah di *upload* oleh pengguna. Hasil dari perhitungan algoritma yaitu berupa persentase kemiripan dan ketentuan apakah jurnal tersebut dapat diterima atau tidak nilai kemiripannya dengan persentase kemiripan yang telah ditentukan.



Gambar 4. 17 Tampilan Perhitungan Algoritma *Cosine Similarity*

Untuk mengontrol aksi pada perhitungan algoritma *cosine similarity*, dibutuhkan proses yang dinamakan *cosinesimilarity.php*. Berikut merupakan *sourcecode* yang digunakan.

### Cosinesimilarity.php

```
<?php
function cosines($a, $b)
{
    $az=buatarray(substr($a,1,1000));
    $ac=count($az);
    $bz=buatarray(substr($b,1,1000));
    $bc=count($bz);
    $is=array_intersect($az, $bz);
    $ic=count($is);
    return $ic / sqrt($ac * $bc); }
function buatarray($str) {
    $a=strtolower($str);
    $len=strlen($a);
    for($n=2;$n<$len;$n++)
        for($i=0 ; $i+$n<=$len ; $i++)
            $arr[]=substr($a,$i,$n);
    return $arr;}
?>
```

Untuk mengetahui jurnal yang di *upload* dapat diterima atau tidak oleh aplikasi deteksi plagiarisme, maka menggunakan *sourcecode* seperti dibawah ini.

### DiterimaDitolak.php

```
<?php
$repoPSD = mysql_query("SELECT Id, Persen_SimilarityCS_Diterima
                        FROM tb_persentase_similarity_diterima where Id=1");
$dataRepoPSD = mysql_fetch_array($repoPSD);
$mProsen = $dataRepoPSD['Persen_SimilarityCS_Diterima']/100;
$repo = mysql_query("SELECT NIM, Nilai_CosineSimilarity FROM
tb_cosinesimilarity");
while($dataRepo = mysql_fetch_array($repo))
{
    $mNDC = $dataRepo['Nilai_CosineSimilarity'];
    $mNIM = $dataRepo['NIM'];
    if ($mNDC<=$mProsen)
    {
        $tso = mysql_query("update tb_cosinesimilarity set Hasil
        ='Diterima' where NIM='$mNIM'");
    } else
    {
        $tso = mysql_query("update tb_cosinesimilarity set Hasil
        ='Ditolak' where NIM='$mNIM'");}
    $repoMHS = mysql_query("SELECT NIM>Nama_Mahasiswa FROM
    tb_mahasiswa where NIM='$mNIM'");
    $dataRepoMHS = mysql_fetch_array($repoMHS);
    $mNM = $dataRepoMHS['Nama_Mahasiswa'];}
?>
```

## 4.2 Pengujian

Pengujian merupakan bagian penting dalam pembangunan sebuah perangkat lunak. Pengujian ditujukan untuk menemukan kesalahan-kesalahan pada sistem dan memastikan sistem yang dibangun telah sesuai dengan apa yang direncanakan sebelumnya. Pengujian dilakukan untuk menjamin kualitas dan juga mengetahui kelemahan dari perangkat lunak. Tujuan dari pengujian adalah untuk menjamin bahwa aplikasi yang dibangun memiliki kualitas yang handal, yaitu mampu mempresentasikan kajian pokok dari spesifikasi analisis, perancangan dan pengkodean dari perangkat lunak itu sendiri.

### 4.2.1 Kasus dan Hasil Pengujian

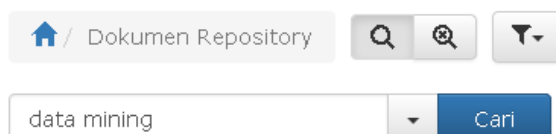
Pada pengujian ini diambil beberapa dokumen uji yang akan dihitung nilai kemiripannya dengan dokumen pembanding yang telah didapat dari hasil pencarian dokumen *repository*, nilai persentase menunjukkan hasil kemiripan antar dokumen yang dihitung menggunakan rumus *cosine similarity*, yaitu

$$\text{similarity}(d_j, q) = \frac{\sum_{i=1}^t (w_{ij} \cdot w_{iq})}{\sqrt{\sum_{i=1}^t w_{ij}^2 \cdot \sum_{i=1}^t w_{iq}^2}}$$

Dalam pengujian dilakukan pencarian dengan nama file 'data mining'.

#### Perhitungan Recall

APLIKASI DETEKSI PLAGIARISME MENGGUNAKAN METODE COSINE SIMILARITY



Home / Dokumen Repository

data mining Cari

Dari proses pencarian, diperoleh hasil 12 dokumen jurnal yang sesuai dengan nama file yang diinputkan pada proses pencarian dalam tabel dokumen repository.

Title	Persen Cosine Similarity	Publisher	Persen Publisher	Jurnal	Nilai Cosine Similarity
A Composite Strategy for the Legal and Ethical Use of Data Mining	0%	International School for Social and Business Studies	0.00		0.00
A Layered Software Architecture for the Management of a Manufacturing Company	0%	Inforec Association	0.00	Informatica Economică vol. 15, no. /2011 Informatica Economică vol. 15, no. /2011 ment where everyone can express own views and make a cooperative workgroup like to Computer Supported Cooperative Work (CSCW [4] Nowadays the world is interconnected, people work in teams, collaborate on projects distributed around the globe. Groupware is a flexible and customizable software that supports and encourages teamwork, exploits the network infrastructure and facilitates the automation of document flows. The Workflow Management System (WFMS) [5] was born in 80s as trend of office systems and it integrated into processes that manage, trace the revisions and the entire production cycle of business documents from the creation to the diffusion at different units software modules are wedged like an ordered assembly	0.02
Application of Data Mining methods in analysis of company's activity	0%	Research Centre of Industrial Problems of Development of NAS of Ukraine	0.00	Business Informatics 2013 www.business-inform.ne 158 c. [Abstract] CEC: http://www.info-library.com.ua/books-text-2488.html	0.00
Based on Similarity Metric Learning for Semi-Supervised Clustering	0%	IFSA Publishing, S.L.	0.00	Sensors & Transducers, Vol. 177, Issue 8, August 2014, pp. 238-245 Abstract: Semi-supervised clustering employs a small amount of labeled data to aid unsupervised learning. The Clustering is a task whose goal is to	0.00
Data mining applications in transportation engineering	0%	Vilnius Gediminas Technical University	0.00		0.00
Data mining in Cloud Computing	0%	Bucharest Academy of Economic Studies	0.00	Database Systems Journal vol. III, no. 3/2012 extracting potentially useful information for day activities has become comm Cloud Computing, Data mining Introduction concepts of recent	0.00
Development of a framework to evaluate service-oriented architecture governance using COBIT approach	0%	Shahrood University of Technology	0.00	Emadi & Deh ghani / Journal of AI and Data Mining, Vol 4, No 2, 2016, 178 The SOA governance maturity models are one of the main tools used to evaluate the SOA governance. A SOA governance maturity model	0.00
Digital Democracy in Knowledge Society: A Proposed Architecture Based on Cloud and Complementary Technologies	0%	Inforec Association	0.00	Informatica Economică vol. 18, no. 4/2014 73 DOI: 10.12948/issn14531305/18.4.2014.07 Digital Democracy in : A Proposed Architecture Based on Cloud and Complementary Technologies Ovidiu TURCOANE University of Economic Studies,	0.00
Evaluation of Classifiers in Software Fault-Proneness Prediction	0%	Shahrood University of Technology	0.00	Journal of AI and Data Mining Vol 5, No 2, 2017, 149 - 167 Evaluation of Classifiers in Software Fault - Proneness Prediction F. Karimian and S. M. Babamir * Department of Computer Engineering, University of Kashan, Kashan, Iran. Received 29 April 2016; Revised 12 June 2016; Accepted 30 October	0.00
Foreword	0%	Institute of Mathematics and Computer Science of	0.00	1 http://www.webology.org/2013/v10n2/a113.pdf Web ology, Volume 10, Number 2, December, 2013 Home Table of Contents Titles & Subject Index Authors Index Perspectives of academic web content managers on the	0.00
Perspectives of academic web content managers on the effectiveness of web publishing and web hosting policies	0%	Webology Center	0.00	1 http://www.webology.org/2013/v10n2/a113.pdf Web ology, Volume 10, Number 2, December, 2013 Home Table of Contents Titles & Subject Index Authors Index Perspectives of academic web content managers on the effectiveness of web publishing and web	0.00
Prioritize the ordering of URL queue in Focused crawler	0%	Shahrood University of Technology	0.00	Journal of AI and Data Mining Vol. 2, No. 1, 2014, 25 - 31. Prioritizing the ordering of URL queue in focused crawler D. Koundal University Institute of Engineering and Technology, Panjab University, Chandigarh, India Received 14 March 2013; accepted 13 July 2013	0.00



					Maedche & Ontology - Focused Crawling of Web Documents Proc. the 2003 ACM symposium on applied computing. [14] Cho, J., H. Garcia - Molina, and L. Page. Efficient 30(17):161172, 1998. [15] Page, L., S. Brin, R. Motwani, T. Winograd. & The PageRank Citation Ranking: Bringing Order to the Web, Stanford Digital Library Technologies Project. [16] Ganesh, S., M. Jayaraj, V. Kalyan, and G. Aghila, & Ontology - based Web Crawler, Proc. of the International Conference on Information Technology: Coding and Computing, Las Vegas, NV, USA, pp.337 - 341, 2004. [17] Deepika Koundal, Mukesh Kumar, Re nu Vig, & Prioritizing the URLs in Ontology based Crawler published and presented at International Conference of IEEE - AICC & 2009 at Thapar University, Patiala. [18] Debashis Hati, Amritesh kumar, 2010. An approach for identifying URLs based on Division score and link score in focused crawler, International journal of computer applications, Volume 2 & No.3. [19] Debashis Hati, Amritesh Kumar, Lizashree Mishra, 2010. Unvisited URL Relevancy Calculation in Focused Crawling Based on Naive Bayesian Classification, International Journal of Computer Applications, Volume 3 - No.9.
--	--	--	--	--	-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

Halaman << < 1 > >> dari 1 Record 1 sampai 12 dari 12

Halaman << < 1 > >> dari 1 Record 1 sampai 12 dari 12

Dari proses pencarian file 'data mining', diketahui jumlah dokumen relevan yang terambil yaitu ada 12 dokumen jurnal.

🏠 / Dokumen Repository 🔍 ⌵

Cari ⌵ Cari

Halaman << < 1 > >> dari 1 Record 1 sampai 91 dari 91

Sedangkan untuk dokumen yang relevan yang ada dalam database yaitu 91 dokumen jurnal.

Dari data dokumen yang diperoleh melalui proses pencarian file 'data mining', dapat dilakukan perhitungan *recall* untuk mengetahui tingkat akurasi suatu sistem.

Perhitungan nilai *recall* dilakukan menggunakan rumus seperti dibawah ini.

$$\begin{aligned}
 Recall &= \frac{\sum \text{dokumen relevan yang terambil}}{\sum \text{dokumen relevan dalam database}} \times 100\% \\
 &= \frac{12}{91} \times 100\% \\
 &= 13\%
 \end{aligned}$$

Perhitungan *recall* merupakan parameter untuk mengukur tingkat akurasi suatu sistem berdasarkan dokumen relevan yang terdapat dalam database. Pada sistem ini, diperoleh nilai *recall* yaitu 13%.

### Perhitungan *Precision*

Title	Persen Cosine Similarity	Publisher	Persen Publisher	Jurnal	Nilai Cosine Similarity
A Composite Strategy for the Legal and Ethical Use of Data Mining	0%	International School for Social and Business Studies	0,00		0,00
A Layered Software Architecture for the Management of a Manufacturing Company	0%	Inforec Association	0,00	Informatica Economică vol. 15, no. /2011 Informatica Economică vol. 15, no. /2011 ment where everyone can express own views and make a cooperative workgroup like to Computer Supported Cooperative Work (CSCW [4] Nowadays the world is interconnected, people work in teams, collaborate on projects distributed around the globe. Groupware is a flexible and customizable software that supports and encourages teamwork, exploits the network infrastructure and facilitates the automation of document flows. The Workflow Management System (WFMS) [5] was born in 80s as trend of office systems and it integrated into processes that manage, trace the revisions and the entire production cycle of business documents from the creation to the diffusion at different units software modules are wedged like an ordered assembly	0,02
Application of Data Mining methods in analysis of company's activity	0%	Research Centre of Industrial Problems of Development of NAS of Ukraine	0,00	Business Informatics, vol. 3, no. 3 — 2013 www.business-inform.net 158 c. [Abstract]. URL: http://www.info-library.com.ua/books-text-2488.html	0,00
Based on Similarity Metric Learning for Semi-Supervised Clustering	0%	IFSA Publishing, S.L.	0,00	Sensors & Transducers, Vol. 177, Issue 8, August 2014, pp. 238-245 Abstract: Semi-supervised clustering employs a small amount of labeled data to aid unsupervised learning. The Clustering is a task whose goal is to	0,00
Data mining applications in transportation engineering	0%	Vilnius Gediminas Technical University	0,00		0,00
Data mining in Cloud Computing	0%	Bucharest Academy of Economic Studies	0,00	Database Systems Journal vol. III, no. 3/2012 extracting potentially useful information for day activities has become common Cloud Computing, Data mining Introduction concepts of recent	0,00
Development of a framework to evaluate service-oriented architecture governance using COBIT approach	0%	Shahrood University of Technology	0,00	Emadi & Dehghani / Journal of AI and Data Mining, Vol 4, No 2, 2016. 178 The SOA governance maturity models are one of the main tools used to evaluate the SOA governance. A SOA governance maturity model	0,00
Digital Democracy in Knowledge Society: A Proposed Architecture Based on Cloud and Complementary Technologies	0%	Inforec Association	0,00	Informatica Economică vol. 18, no. 4/2014 73 DOI: 10.12948/issn14531305/18.4.2014.07 Digital Democracy in : A Proposed Architecture Based on Cloud and Complementary Technologies Ovidiu TURCOANE University of Economic Studies,	0,00
Evaluation of Classifiers in Software Fault-Proneness Prediction	0%	Shahrood University of Technology	0,00	Journal of AI and Data Mining Vol 5, No 2, 2017, 149 - 167 Evaluation of Classifiers in Software Fault-Proneness Prediction F. Karimian and S. M. Babamir * Department of Computer Engineering, University of Kashan, Kashan, Iran. Received 29 April 2016; Revised 12 June 2016; Accepted 30 October	0,00
Foreword	0%	Institute of Mathematics and Computer Science of	0,00	1 http://www.webology.org/2013/v10n2/a113.pdf Webology, Volume 10, Number 2, December, 2013 Home Table of Contents Titles & Subject Index Authors Index Perspectives of academic web content managers on the	0,00

Perspectives of academic web content managers on the effectiveness of web publishing and web hosting policies	0%	Webology Center	0.00	1 <a href="http://www.webology.org/2013/v10n2/a113.pdf">http://www.webology.org/2013/v10n2/a113.pdf</a> Web o logy , Volume 10, Number 2, December, 2013 Home Table of Contents Titles & Subject Index Authors Index Perspectives of academic web content managers on the effectiveness of web publishing and web	0.00
Prioritize the ordering of URL queue in Focused crawler	0%	Shahrood University of Technology	0.00	J ournal of AI and D ata M ining Vol. 2, No. 1 , 2014, 25 - 31 . Prioritizing the ordering of URL queue in focused crawler D. Koundal University Institute of Engineering and Technology, Panjab University, Chandigarh, India Received 14 March 2013; accepted 13 July 2013 Maedche &œOntology - Focused Crawling of Web Documents&œ Proc. the 2003 ACM symposium on applied computing. [14] Cho, J. , H.Garcia - Molina, and L. Page. Efficient 30(17):161172, 1998. [15] Page, L. , S. Brin, R. Motwani, T. Winograd. &œThe PageRank Citation Ranking: Bringing Order to the Web&œ, Stanford Digital Library Technologies Pr oject. [16] Ganesh, S. , M. Jayara, V. Kalyan, and G. Aghila,&œOntology - based Web Crawler,&œ Proc. of the International Conference on Information Technology: Coding and Computing, Las Vegas, NV, USA, pp.337 - 341, 2004. [17] Deepika Koundal, Mukesh Kumar, Re nu Vig, &œPrioritizing the URLs in Ontology based Crawler&œ published and presented at International Conference of IEEE - AICC &œ 2009 at Thapar University, Patiala. [18] Debashis Hati, Amrutesh kumar, 2010. An approach for identifying URLs based on Div ision score and link score in focused crawler, International journal of computer applications, Volume 2 &œ No.3. [19] Debashis Hati, Amrutesh Kumar, Lizashree Mishra, 2010. Unvisited URL Relevancy Calculation in Focused Crawling Based on Naive Bayesian Classification, International Journal of Computer Applications, Volume 3 - No.9.	0.00
Halaman << < 1 > >> dari 1 Record 1 sampai 12 dari 12					
Halaman << < 1 > >> dari 1 Record 1 sampai 12 dari 12					

Jumlah dokumen relevan yang diambil pada saat melakukan proses pencarian file 'data mining' yaitu 12 dokumen jurnal.

A Layered Software Architecture for the Management of a Manufacturing Company	0%	Infotec Association	0.00	Informatica Economic&œ vol. 15, no. /2011 Informatica Economic&œ vol. 15, no. /2011 ment where everyone can express own views a nd make a cooperative workgroup like to Computer Supported Cooperative Work (CSCW [4] Nowadays the world is interconnected, people work in teams, collaborate on projects distributed around the globe. Groupware is a flexible and customizable software that supports and encourages teamwork, exploits the network infrastructure and facilitates the automation of document flows. The Workflow Management System (WFMS) [5] was born in 80s as trend of office systems and it integrated into processes that manage , trace the revisions and the entire production cycle of business documents from the creation to the diffusion at different units software modules are wedged like an ordered assembly line that distribute information [6]. The birth of Enterprise Resource Planning ERP [7] , in 90s, allowed to companies to manage, synchronize and integrate business functions optimizing available resources. The ERP is primarily an organizational and managerial method to optimize operational activities of a manufacturing company. ERP manages customized orders and integrates business cycle and all aspects of the business: planning, manufacturing, sales, finance, procurement, logistics and marketing. ERP working for processes has led the enterprise to re engineer deeply business processes by the Business	0.02
Halaman << < 1 > >> dari 1 Record 1 sampai 12 dari 12					

Sedangkan untuk jumlah dokumen yang relevan dalam pencarian diperoleh 1 dokumen jurnal yang sesuai dengan nama file pada proses pencarian 'data mining' dalam dokumen repository dengan nilai *cosine similarity* diatas 0% dari total jurnal dalam pencarian sejumlah 12 file.

Dari data jumlah dokumen jurnal yang diperoleh dari proses pencarian file 'data mining' seperti dijelaskan diatas, dapat dilakukan perhitungan nilai *precision* dengan menggunakan rumus seperti dibawah ini

$$\begin{aligned} \text{Precision} &= \frac{\sum \text{dokumen relevan yang terambil}}{\sum \text{dokumen relevan dalam pencarian}} \times 100\% \\ &= \frac{1}{12} \times 100\% \\ &= 8\%. \end{aligned}$$

Perhitungan *precision* merupakan parameter untuk mengukur tingkat akurasi sebuah sistem berdasarkan dokumen yang relevan pada pencarian yang dilakukan. Pada sistem ini, diperoleh nilai *precision* yaitu 8%.

### 4.3 Integrasi Islam

Aplikasi deteksi plagiarisme dibangun untuk membantu mendeteksi tingkat plagiarisme terutama dikalangan akademisi. Perilaku plagiarisme dapat dikatakan sebagai pencurian atau mencuri. Mencuri sendiri mengandung makna mengambil tanpa izin milik orang lain. Plagiarisme merupakan penjiplakan atau pengambilan karangan, pendapat, dan sebagainya dari orang lain dan menjadikannya seolah karangan dan pendapat sendiri. Hal tersebut sesuai dengan firman Allah SWT, dalam Al-Qur'an surat An-Nisa' Ayat 29:

يَا أَيُّهَا الَّذِينَ آمَنُوا لَا تَأْكُلُوا أَمْوَالَكُمْ بَيْنَكُمْ بِالْبَاطِلِ إِلَّا أَنْ تَكُونَ تِجَارَةً  
عَنْ تَرَاضٍ مِّنْكُمْ وَلَا تَقْتُلُوا أَنْفُسَكُمْ إِنَّ اللَّهَ كَانَ بِكُمْ رَحِيمًا

“Hai orang-orang yang beriman, janganlah kamu saling memakan harta sesamamu dengan jalan yang batil, kecuali dengan jalan perniagaan yang berlaku dengan suka sama-suka di antara kamu. Dan janganlah kamu membunuh dirimu; sesungguhnya Allah adalah Maha Penyayang kepadamu.” (Q.S An-Nisa’ : 29).

Surat An-Nisa’ menjelaskan tentang hukum transaksi secara umum, lebih khusus kepada transaksi perdagangan, bisnis jual beli. Dalam ayat ini Allah mengharamkan orang beriman untuk memakan, memanfaatkan, menggunakan harta orang lain dengan jalan yang batil, yaitu yang tidak dibenarkan oleh syariat. Plagiarisme merupakan tindakan yang diharamkan karena memanfaatkan hak dan harta orang lain tanpa izin.

‘Ali bin Abi Thalib mengatakan dari Ibnu ‘Abbas : “Ketika diturunkan oleh Allah: *yaa ayyuHal ladziina aamanuu laa ta’kuluu amwaalakum bainakum bil baathili* (“Hai orang-orang yang beriman janganlah kamu saling memakan harta sesamamu dengan jalan yang bathil,”) kaum muslimin berkata, “Sesungguhnya Allah telah melarang kita untuk memakan harta di antara kita dengan bathil. Sedangkan makanan adalah harta kita yang paling utama, untuk itu tidak halal bagi kita makan di tempat orang lain, maka bagaimana dengan seluruh manusia?”.

طَبَّنَا رَسُولُ اللَّهِ صَلَّى اللَّهُ عَلَيْهِ وَسَلَّمَ فَقَالَ أَلَا وَ لَا يَحِلُّ لِأَمْرٍ مِنْ مَالِ أَخِيهِ شَيْءٌ إِلَّا بِطَيْبِ نَفْسٍ  
(مِنْهُ) (رواه أحمد في مسنده)

*“Rasulullah SAW menyampaikan khotbah kepada kami, sabdanya; “Ketahuilah, tidak halal bagi seseorang sedikitpun dari harta saudaranya kecuali dengan kerelaan hatinya.”(HR. Ahmad).*

Ayat dan hadist tersebut mengingatkan agar tidak memakai atau menggunakan hak orang lain, dan tidak pula memakan harta orang lain, kecuali dengan persetujuan. Pelanggaran terhadap hak orang lain termasuk hak cipta bisa termasuk kedalam kategoris *mufliis*, yaitu orang yang bangkrut amalnya nanti di akhirat.



## **BAB V**

### **PENUTUP**

#### **5.1 Kesimpulan**

Dari hasil penelitian dan pembahasan tentang aplikasi deteksi plagiarisme menggunakan metode *cosine similarity*, dapat diambil kesimpulan nilai akurasi sistem dengan melakukan perhitungan *recall* dan *precision* dari perhitungan *cosine* dengan mengambil data dan membandingkan dengan repository yang telah ada. Nilai *recall* untuk kasus ini yaitu 13%, diperoleh dari jumlah dokumen relevan yang terambil dibagi dengan jumlah dokumen yang ada dalam database dikali 100%. Sedangkan nilai *precision* yaitu 8%, diperoleh dari jumlah dokumen relevan yang terambil dibagi dengan jumlah dokumen relevan yang ada dalam pencarian dikali 100%.

#### **5.2 Saran**

Berdasarkan pengujian yang telah dilakukan pada aplikasi deteksi plagiarisme menggunakan metode *cosine similarity*, tentunya masih banyak kekurangan dan kelemahan sehingga masih perlu dilakukan pengembangan lebih lanjut. Untuk aplikasi ini diperlukan pengembangan pada proses perhitungan algoritma agar proses perhitungan bisa berjalan lebih cepat.

## REFERENSI

- Alamanda, Rio, Cucu Suhery, Yulrio Brianorman, and Jurusan Sistem Komputer. 2016. "Aplikasi Pendeteksi Plagiat Terhadap Karya Tulis Berbasis Web Menggunakan Natural Language Processing Dan Algoritma Knuth Moriris Pratt." *Jurnal Coding, Sistem Komputer Untan* 4(1).
- Alewiwi, Mahmoud, Cengiz Orencik, and Erkay Savaş. 2016. "Efficient Top-K Similarity Document Search Utilizing Distributed File Systems and Cosine Similarity." *Cluster Computing* 19(1): 109–26.
- Firdaus, Hari Bagus. 2003. "Algoritma Rabin-Karp." *Jurnal Ilmu Komputer dan Teknologi Informasi* III No. 2: 1–5.
- Haiyan, Muhammad Jabir Al. 2017. "PERANCANGAN DAN PEMBUATAN WEB FRAMEWORK LIBRARY."
- Imbar, Radiant Victor et al. 2014. "Implementasi Cosine Similarity Dan Algoritma Smith-Waterman Untuk Mendeteksi Kemiripan Teks." *Jurnal Informatika*: 31–42.
- Ismail, Eka Widhi Yunarso. 2014. "Aplikasi Berbasis Web Pendeteksi Plagiarisme Menggunakan Algoritma Himpunan Kata." *Infotel* 6(2): 2–7.
- Kumar, Manish, Ankit Bindal, Robin Gautam, and Rajesh Bhatia. 2018. "Keyword Query Based Focused Web Crawler." *Procedia Computer Science* 125: 584–90.  
<http://linkinghub.elsevier.com/retrieve/pii/S1877050917328399>.
- Kurniawan, Andry, Firdaus Solihin, and Fika Hastarita. 2014. "Perancangan Dan Pembuatan Aplikasi Pencarian Informasi Beasiswa Dengan Menggunakan Cosine Similarity." 4(2): 115–24.



- Nurdiana, Ogie, Jumadi, and Dian Nursantika. 2016. "Perbandingan Metode Cosine Similarity Dengan Metode Jaccard Similarity Pada Aplikasi Pencarian Terjemah Al-Qur'an Dalam Bahasa Indonesia." *Jurnal Online Informatika (JOIN)* 1(1): 59–63.
- Pahlevi, Irfan, Moch Arief Bijaksana, and M Ir Tech. "Perhitungan Kemiripan Dokumen Bahasa Indonesia Menggunakan Metode Cosine Similarity ( Studi Kasus : Abstrak Tugas Akhir Fakultas Informatika Universitas Telkom )."
- Perdana, Khairul. 2014. "Pencarian Dan Perankingan Obat Tradisional Berdasarkan Gejala Penyakit Menggunakan Metode Cosine Similarity Skripsi."
- Prima, Verry, and Anindya Rhosady. 2013. "Rancang Bangun Aplikasi Pengklasifikasian Halaman Web Berdasarkan."
- Rungsawang, Arnon, and Niran Angkawattanawit. 2005. "Learnable Topic-Specific Web Crawler." *Journal of Network and Computer Applications* 28(2): 97–114.
- Rusydi, Ibnu. 2014. "Pemanfaatan E-Journal Sebagai Media Informasi Digital." *Jurnal Iqra'* 8(2): 200–210. <http://oaji.net/articles/2015/1937-1430103772.pdf>.
- Santoso, Hari. 2015. "PENCEGAHAN DAN PENAGGULANGAN PLAGIARISME DALAM PENULISAN KARYA ILMIAH DI LINGKUNGAN PERPUSTAKAAN PERGURUAN TINGGI Oleh : Drs. Hari Santoso, S.Sos. 1." *Perpustakaan UM Malang* (1): 1–23.
- Sugiyamta. 2015. "Sistem Deteksi Kemiripan Dokumen Dengan Algoritma Cosine Similarity Dan Single Pass Clustering." *Dinamika Informatika* 7(2): 7.

Thada and Jaglan. 2013. "Comparison of Jaccard, Dice, Cosine Similarity Coefficient to Find Best Fitness Value for Web Retrieved Document Using Genetic Algorithm". India. International Journal of Innovations in Engineering and Technology (IJET).

Wibowo, Adik.2012."Mencegah dan Menanggulangi Plagiarisme di Dunia Pendidikan". Departemen Administrasi dan Kebijakan Kesehatan Fakultas Kesehatan Masyarakat Universitas Indonesia.

Zainur, Muhammad.2012."Plagiarisme Di Kalangan Mahasiswa Dalam Membuat Tugas-Tugas Perkuliahan Pada Fakultas Tarbiyah Iain Imam Bonjol Padang". Dosen Fakultas Tarbiyah IAIN Imam Bonjol Padang.

Zuliarso, Eri. 2010. "Aplikasi Web Crawler Berdasarkan Breadth First Search Dan Back-Link." Fakultas Teknologi Informasi, Universitas Stikubank Semarang XV(1): 52–56.

## LAMPIRAN

**3pdf-download.php**

```

<?php
error_reporting(0);
ini_set('display_errors',0);
//error_reporting(E_ALL & ~E_NOTICE);
//error_reporting(E_ALL ^ E_DEPRECATED);
ini_set('memory_limit', '-1');
ini_set('max_execution_time', 30000);
include "koneksi.php";
include 'connection.php';
include ( 'PdfToText.phpclass' ) ;

$repo = mysql_query("SELECT fulltext_url, title,
nama_author,publisher>Nama_pdf FROM tb_grabbingdata where
right(fulltext_url,4)='.pdf'");
$fileke=0;
while($dataRepo = mysql_fetch_array($repo))
{
    $mfulltext_url = $dataRepo['fulltext_url'];
    $mtitle = $dataRepo['title'];
    $mnama_author = $dataRepo['nama_author'];
    $mpublisher = $dataRepo['publisher'];
    $fileke++;

    $nmfile=strval($fileke);
    $path = "c:/xampp5/htdocs/DeteksiPlagiarisme/repository/".$nmfile.".pdf";

    $ch = curl_init($mfulltext_url);
    curl_setopt($ch, CURLOPT_RETURNTRANSFER, true);
    curl_setopt($ch, CURLOPT_REFERER, $mfulltext_url);

    $data = curl_exec($ch);

    curl_close($ch);

    $result = file_put_contents($path, $data);
    $tso = mysql_query("update tb_grabbingdata set
nama_pdf='$nmfile' where fulltext_url='$mfulltext_url'");
}

$tso = mysql_query("INSERT INTO tb_dokumenrepository
(fulltext_url,title,nama_author,publisher,nama_pdf) SELECT fulltext_url,
title, nama_author,publisher,nama_pdf FROM tb_grabbingdata where
length(nama_pdf)>0");

$repo = mysql_query("SELECT fulltext_url, title,
nama_author,publisher,nama_pdf FROM tb_dokumenrepository");
while($dataRepo = mysql_fetch_array($repo))
{
    $mfulltext_url = $dataRepo['fulltext_url'];
    $mtitle = $dataRepo['title'];
    $mnama_author = $dataRepo['nama_author'];
    $mpublisher = $dataRepo['publisher'];
    $mnama_pdf = $dataRepo['nama_pdf'];

    $filedirektori="c:/xampp5/htdocs/DeteksiPlagiarisme/repository/".$
mnama_pdf.".pdf";

```

```

10240)                                     if (filesize($filedirektori) <=
{
    unlink($filedirektori);
    $tso = mysql_query("update
tb_dokumenrepository set          nama_pdf=' ' where
fulltext_url='".$mfulltext_url."'");
}

}

include 'pdf2text.php';
$repol = mysql_query("SELECT fulltext_url, title,
nama_author,publisher,nama_pdf,Jurnal FROM tb_dokumenrepository where
(length(nama_pdf))>0 and length(trim(Proses))=0");
while($dataRepol = mysql_fetch_array($repol))
{
    $mfulltext_url = $dataRepol['fulltext_url'];
    $mtitle = $dataRepol['title'];
    $mnama_author = $dataRepol['nama_author'];
    $mpublisher = $dataRepol['publisher'];
    $mnama_pdf = $dataRepol['nama_pdf'];

    //$filedirektori="c:/xampp5/htdocs/DeteksiPlagiarisme/upload/".$nmfile;
    $filedirektori1="c:/xampp5/htdocs/DeteksiPlagiarisme/Repository/".$
    trim($mnama_pdf).".pdf";
    $mteks = pdf2text
($filedirektori);
    //echo $filedirektori;
    //$pdf = new PdfToText
($filedirektori);
    //$mteks = $pdf -> Text ;
    //echo $mteks;
    $result =str_replace("'", "", $mteks);;
    //$result = str_replace("'", "",
$result);
    //$result = str_replace("[", "",
$result);
    //$result = str_replace("]", "",
$result);
    //$result = str_replace("~", "",
$result);

    $sql = "update
tb_dokumenrepository set Jurnal='".$result."',Proses='S' where
fulltext_url='".$mfulltext_url."'";
    mysql_query($sql);
}

echo ' <a
href="http://localhost/deteksiplagiarisme/menuupload.php">Kembali ke menu
upload</a>';

?>

```

### Grabbing-doaj.php

```

<?php
error_reporting(0);
//error_reporting(E_ALL & ~E_NOTICE);

```

```

//error_reporting(E_ALL ^ E_DEPRECATED);
include 'connection.php';
set_time_limit(0);
ini_set("display_errors",0);
$con = mysql_connect("localhost","root","");

class GrabbingDOAJ
{
    private $host="localhost";
    private $dbname="deteksiplagiarisme";
    private $conn,$url;

    // koneksi ke database mysql
    private $driver="mysql";
    private $user="root";
    private $password="";
    private $port="3306";

    // diload pertama kali
    public function __construct($url)
    {
        $this->url = $url;
        try
        {
            if ($this->driver == 'mysql')
            {
                $this->conn = new PDO("mysql:host=$this->host;port=$this->port;dbname=$this->dbname;charset=utf8",$this->user,$this->password);
            } elseif ($this->driver == 'pgsql')
            {
                $this->conn = new PDO("pgsql:host=$this->host;port=$this->port;dbname=$this->dbname;user=$this->user;password=$this->password");
            }
        } catch (PDOException $e)
        {
            echo "Koneksi gagal";
        }
    }

    public function doaj()
    {
        $ch = curl_init();
        curl_setopt ($ch, CURLOPT_URL, $this->url);
        curl_setopt ($ch, CURLOPT_CONNECTTIMEOUT, 5);
        curl_setopt ($ch, CURLOPT_RETURNTRANSFER, true);
        $response = curl_exec($ch);
        $data = json_decode($response, true);

        $no = 1;
        $jml_data = count($data['results']);
        for ($i=0;$i<$jml_data;$i++)
        {
            $title = $data['results'][$i]['bibjson']['title'];
            echo $no.". \ $title = ". $title;
            echo "<br/>";

            $jml_author =
            count($data['results'][$i]['bibjson']['author']);
            echo "\ $jml_author = ". $jml_author;
            echo "<br/>";

            $author = "";
            for ($j=0;$j<$jml_author;$j++)
            {
                if ($j == 0)
                {
                    $author
                } else
                {
                    $author
                }
            }
            .= $data['results'][$i]['bibjson']['author'][$j]['name'];
            .=",";
        }
    }
}

```

```

    }
    }
    echo "\$author = ".$author;
    echo "<br/>";

    $abstract = "";
    if
(isset($data['results'][$i]['bibjson']['abstract']))
    {
        $abstract
$data['results'][$i]['bibjson']['abstract'];
    }
    echo "\$abstract = ".$abstract;
    echo "<br/>";

    $fulltext_url
$data['results'][$i]['bibjson']['link'][0]['url'];
    echo "\$fulltext_url = ".$fulltext_url;
    echo "<br/>";

    $publisher
$data['results'][$i]['bibjson']['journal']['publisher'];
    echo "\$publisher = ".$publisher;
    echo "<br/>";

    echo "<br/> ===== <br/>";

    // $query = $this->conn->prepare("INSERT IGNORE INTO
data (title,author,abstract,fulltext_url,publisher) VALUES
(:title,:author,:abstract,:fulltext_url,:publisher)");
    // $query-
>execute(array(':title'=>$title,':author'=>$author,':abstract'=>$abstract
,':fulltext_url'=>$fulltext_url,':publisher'=>$publisher));
    // $query = null;
    // ok
    $sql = "INSERT IGNORE INTO tb_grabbingdata
(title,nama_author,fulltext_url,publisher) VALUES
('".$title."','".$author."','".$fulltext_url."','".$publisher."')";
    // mysql_query($sql,$con);
    mysql_query($sql);

    $no++;

//unset($title,$jml_author,$author,$abstract,$fulltext_url);
}

    echo "<br/>##### <br/>";
    echo "<pre>";
    // echo print_r($data);
    // echo print_r($data['results']);
    // echo print_r($data['results'][0]['bibjson']);

    // author
    // echo print_r($data['results'][0]['bibjson']['author']);

    // title
    // echo print_r($data['results'][0]['bibjson']['title']);

    // abstract
    // echo print_r($data['results'][0]['bibjson']['abstract']);

    // fulltext url
    //
print_r($data['results'][0]['bibjson']['link'][0]['url']);
    echo "</pre>";
    //unset($ch,$response,$data,$no,$jml_data,$i);

```

```

    }

    public function __destruct()
    {
        $this->conn = null;
        unset($this->url);
    }
}

// catatan
// ubah "Computer software" dengan subject yang ingin dicari
// ubah "pageSize" sesuai dengan jumlah data yang ingin di download
// $url='https://doaj.org/api/v1/search/articles/Computer
software?pageSize=10';
$con = mysql_connect("localhost","root","");
$sql = "TRUNCATE TABLE tb_grabbingdata";
mysql_query($sql,$con);
$mnamafile = $_POST['fnamafile'];
$mjmlfile = $_POST['fjmlfile'];
$url='https://doaj.org/api/v1/search/articles/'.$mnamafile.'?pageSize='.$
mjmlfile;

$sql = "TRUNCATE TABLE tb_similarityperpublisher";
mysql_query($sql,$con);
$sql = "INSERT INTO tb_similarityperpublisher (keterangan) VALUES
('".$mnamafile."')";
mysql_query($sql,$con);

$a = new GrabbingDOAJ($url);
$a->doaj();
include "3pdf-download.php";
echo
href="http://localhost/deteksiplagiarisme/menuupload.php">Kembali ke menu
upload</a>';?>
upload.php

<html xmlns="http://www.w3.org/1999/xhtml">
<head>
<link rel="stylesheet" href="css/bootstrap.min.css" />
<style type="text/css">
#wrapper {
margin: 0 auto;
float: none;
width:70%;
}
.header {
padding:10px 0;
border-bottom:1px solid #CCC;
}
.title {
padding: 0 5px 0 0;
float:left;
margin:0;
}
.container form input {
height: 30px;
}
body
{
font-size:12;
font-weight:bold;
}

```

```

        </style>
        <meta http-equiv="Content-Type" content="text/html; charset=utf-8"
/>
        <title>Upload File</title>

        <?php
            error_reporting(0);
            //error_reporting(E_ALL & ~E_NOTICE);
            //error_reporting(E_ALL ^ E_DEPRECATED);
            ini_set("display_errors",0);
            if(!empty($_POST))
            {
                include 'connection.php';
                include ( 'PdfToText.phpclass' );
                $con = mysql_connect("localhost","root","");
                if (!$con)
                    echo('Tidak dapat terhubung ke server: '
                . mysql_error());
                else
                {
                    if (file_exists("upload/"
                $_FILES["file"]["name"]))
                    {
                        echo 'Maaf! Nama file tsb. sudah
                <script
                language="javascript">alert("
                //ada!...")</script>';
                    }
                    else
                    {
                        move_uploaded_file($_FILES["file"]["tmp_name"],
                "upload/"
                $_FILES["file"]["name"]);
                        $nmfile=$_FILES["file"]["name"];
                        $filedirektori="c:/xampp5/htdocs/DeteksiPlagiarisme/upload/".$nmfile;
                        //echo $filedirektori;
                        $pdf = new PdfToText
                ($filedirektori);
                        $mteks = $pdf -> Text ;
                        //echo $mteks;
                        $msimpan =str_replace("'", "", $mteks);

                        mysql_select_db("deteksiplagiarisme", $con);
                        $sql = "INSERT INTO
                tb_upload(nim,subyek,topik,file,teks)
                VALUES
                ('".$_POST["nim"]."', '".$_POST["sub"]."', '".$_POST["pre"]."', '".$_$nmfile."
                ', '".$_$msimpan."'");
                        if (!mysql_query($sql,$con))
                            echo('Error
                :
                '
                . mysql_error());
                        else
                        {
                            $tso = mysql_query("update
                tb_mahasiswa set Jurnal = '".$_$msimpan.'" where NIM='".$_POST["nim"]."'");
                            echo 'Terima kasih! File telah
                <script
                language="javascript">alert("
                diupload")</script>';
                        }
                    }
                }
            }
            mysql_close($con);
        ?>
    </head>
    <body>

```



```

<div class="container home">
<br>
    <h3><center> UPLOAD JURNAL/ ABSTRAK </center> </h3> </font>

    <form id="form3" enctype="multipart/form-data" method="post"
action="upload.php">
        <table class="table table-bordered">
            <tr>
                <td> <label for="sub">Subyek: </label></td>
                <td> <input type="text" name="sub" id="sub"
class="input-medium"
placeholder="Judul dari subyek Jurnal"/>
                </td>
            </tr>
            <tr>
                <td> <label for="file">Nama file:</label></td>
                <td> <input type="text" name="pre" cols="50" rows="10" id="pre"
placeholder="Jenis file"
class="input-medium"
required></td>
            </tr>
            <tr>
                <td><label for="file">Nama file:</label></td>
                <td><input type="file" name="file" id="file"
title="Klik disini untuk memilih file yang akan
diupload." required /></td>
            </tr>
            <tr>
                <td><label for="nim">Nomor Induk
Mahasiswa:</label></td>
                <td><input type="text" name="nim" id="nim"
placeholder="Ketik NIM dengan teliti"
title="Ketik NIM dengan teliti" required /></td>
            </tr>
            <tr>
                <td colspan="2" align="center">
                    <input type="submit" class="btn btn-
primary" name="upload" id="upload"
title="Klik disini untuk mengupload file."
value="Upload File" />
                    <a
href="http://localhost/deteksiplagiarisme/menuupload.php"> Kembali ke
menu upload </a></td>
            </tr>
        </table>
    </form>
</div>
</body>
</html>

```

### Cosinesimilarity.php

```

<?php
error_reporting(0);
ini_set('memory_limit', '-1'); //utk mengubah batas ukuran memory menjadi
tidak terhingga
ini_set('post_max_size', '200M');
ini_set('upload_max_filesize', '200M');
ini_set('max_input_time', 1000);
ini_set('session.gc_maxlifetime', 1200);
ini_set('max_excution_time', 86400);

```

```
set_time_limit(86400);

function cosines($a, $b)
{
    $az=buatarray(substr($a,1,1300));
    $ac=count($az);

    $bz=buatarray(substr($b,1,1300));
    $bc=count($bz);

    $is=array_intersect($az, $bz);
    $ic=count($is);

    return $ic / sqrt($ac * $bc);
}

function buatarray($str)
{
    $a=strtolower($str);
    $len=strlen($a);

    for($n=2;$n<$len;$n++)
        for($i=0 ; $i+$n<=$len ; $i++)
            $arr[]=substr($a,$i,$n);
    return $arr;
}
?>
```

