RANCANG BANGUN SISTEM INFORMASI PENDETEKSI KESAMAAN DOKUMEN MENGGUNAKAN METODE *DICE SIMILARITY*



ARIS AKHYAR ABDILLAH

HO71171505

Pembimbing Utama : Dr. Hendra S.Si., M.Kom.

Pembimbing Pertama : Edy Saputra R, S.Si., M.Si.

PROGRAM STUDI SISTEM INFORMASI

FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM

UNIVERSITAS HASANUDDIN

MAKASSAR

2022

# DAFTAR ISI

[DAFTAR ISI ii](#_Toc105956746)

[BAB I PENDAHULUAN 1](#_Toc105956747)

[1.1 Latar Belakang 1](#_Toc105956748)

[1.2 Rumusan Masalah 2](#_Toc105956749)

[1.3 Tujuan Penelitian 3](#_Toc105956750)

[1.4 Manfaat Penelitian 3](#_Toc105956751)

[1.5 Batasan Masalah 3](#_Toc105956752)

[BAB II TINJAUAN PUSTAKA 4](#_Toc105956753)

[BAB III METODE PENELITIAN 5](#_Toc105956754)

# BAB I PENDAHULUAN

## Latar Belakang

Teknologi merupakan sarana untuk menyediakan barang-barang yang diperlukan untuk mempermudah kegiatan atau kehidupan manusia. Penggunaan teknologi oleh manusia dimulai dengan pengubahan sumber daya alam menjadi alat-alat sederhana. Penemuan di masa lampau tentang kemampuan mengendalikan api telah menaikkan ketersediaan sumber-sumber pangan, sedangkan penciptaan roda telah membantu manusia dalam bepergian dari satu tempat ke tempat lainnya. Perkembangan teknologi terbaru dimasa sekarang seperti mesin cetak, telepon, dan internet telah mengubah manusia untuk bekerja maupun berinteraksi dalam menjalankan kehidupannya.

Dengan adanya teknologi di saat ini, dapat dikatakan pekerjaan manusia semakin mudah. Seperti untuk akses informasi, sekarang sudah banyak sekali informasi yang tersebar di internet dan hampir semua yang dicari dapat ditemukan. Tidak seperti dahulu, untuk mengakses informasi, orang biasa pergi ke perpustakaan untuk mendapatkan informasi yang diinginkan. Hal itu memungkinkan karena adanya digitalisasi dokumen yang dulunya hanya berupa *hardcopy* (salinan cetak) menjadi *softcopy* (salinan digital). Dengan hal tersebut kita bisa mengakses semua dokumen kapan saja dan dimana saja. Dengan semua kemudahan yang diberikan terknologi seperti contoh diatas, timbul suatu masalah yaitu mudahnya dilakukan plagiarisme yang merupakan suatu tindakan yang tidak baik.

Plagiarisme berasal dari bahasa Latin “plagiare” yang berarti mencuri. Plagiarisme berasal dari kata plagiat yang berarti 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. Sehingga dapat diartikan plagiarisme merupakan tindakan mencuri gagasan hasil penelitian orang lain, untuk kemudian disajikan seolah-olah milik sendiri (Ridhatillah, 2013). Tindakan plagiarisme merupakan salah satu tindakan yang melanggar hak cipta. Hak Cipta itu sendiri merupakan hak eksklusif untuk Pencipta ataupun penerima hak buat mengumumkan ataupun perbanyak Ciptaannya ataupun membagikan izin buat itu dengan tidak kurangi pembatasan- pembatasan bagi peraturan perundang- undangan yang berlaku. Jika terjadi pelanggaran tersebut, dapat dikenai pelanggaran hak cipta dPasal 72 ayat UUHC dengan dipidana dengan pidana penjara pendek selama 1 bulan serta/ ataupun denda sangat sedikitnya Rp1.000.000,00, ataupun pidana penjara lama 7 tahun serta/ ataupun denda sangat sebanyak Rp5.000.000. 000,00.

Untuk mengurangi hal tersebut, dapat dilakukan beberapa metode untuk mencegah plagiarisme seperti menggunakan cara manual ataupun menggunakan teknologi yang ada. Jika menggunakan cara manual, dilakukan dengan membandingkan sebuah dokumen yang akan dicek dengan dokumen yang akan dijadikan perbandingan dengan melihat kata yang digunakan ataupun tata penulisan dokumen tersebut. Cara ini kurang efektif karena jumlah dokumen yang akan dijadikan perbandingan jumlah tidak sedikit bisa mencapai ratusan bahkan ribuan dokumen lebih. Sedangkan untuk cepat menggunakan menggunakan mesin sebagi pembanding.

Ada beberapa metode yang dapat digunakan seperti *Cosine Similarity*, TF-IDF, *Jaccard Similarity*, *Dice Similarity*, Word2Vec dan sebagainya. Dengan menggunakan salah satu metode diatas, dapat temukan persentase kesamaan sebuah dokument. Salah satu metode yang dapat digunakan yaitu dengan metode Dice Similarity. Dice Similarity merupakan salah satu metode perbandingan antara dokumen dengan membandingkan kata yang ada diantara dokumen tersebut. Dengan adanya penelitian ini, diharapkan dapat membangun sebuah Sistem Informasi berbasis Web yang mampu memberikan persentase kesamaan dari sebuah dokumen terhadap dokumen lainnya yang terdapat didalam database guna mengurangi tindak plagiarisme yang sering terjadi saat ini.

## Rumusan Masalah

Adapun beberapa point yang menjadi rumusan masalah penelitian ini yaitu :

1. Bagaimana merancang sebuah sistem informasi berbasis web menggunakan metode *Dice Similarity* ?
2. Bagaimana mengimplementasikan metode *Dice Similarity* ?
3. Bagaimana mengukur kesamaan dari sebuah dokumen dengan dokumen lainnya menggunakan metode *Dice Similarity* ?

## Tujuan Penelitian

Adapun beberapa point yang menjadi tujuan penelitian ini yaitu :

1. Merancang sebuah sistem informasi berbasis web menggunakan metode *Dice Similarity*.
2. Mengimplementasikan metode *Dice Similarity*.
3. Mengukur kesamaan dari sebuah dokumen dengan dokumen lainnya menggunakan metode *Dice Similarity*.

## Manfaat Penelitian

Penelitian ini merupakan mengimplimentasian metode *Dice Similarity* pada sebuah sistem informasi berbasis web guna mengetahui persentase kesamaan sebuah dokumen dengan dokumen lainnya sehungga dapat digunakan untuk mengurangi tindak plagiarisme dengan mudah.

## Batasan Masalah

Adapaun beberapa batasan masalah pada penelitian ini yaitu :

1. Sistem ini dibuat menggunakan bahasa pemrograman Javascript dengan runtime environment Nodejs, *framework* Express JS, dengan database MongoDB.
2. Menggunakan data berupa jurnal yang diinput oleh user dan memiliki ekstensi pdf (*Portable Document Format*).
3. Menggunakan metode *Dice Similarity*.

# BAB II TINJAUAN PUSTAKA

## 2.1 Penelitian Terkait

Dalam Penelitian Mufti Ari Bianto, dkk (2018) dengan judul *Perancangan Sistem Pendeteksi Plagiarisme Terhadap Topik Penelitian Menggunakan Metode K-Means Clustering dan Model Bayesian*, hasil kemiripan diuji berdasarkan kesamaan pola kata dalam kalimat, dokumen dengan Nilai similarity tertinggi diperoleh 1 dokumen dengan presentase kemiripan sebanyak 100 %, dan jika berdasarkan kesamaan term, berpengaruh terhadap hasil dokumen mirip yang dihasilkan, sehingga diperoleh 2 dokumen mirip. Meskipun uji berdasarkan kesamaan term, menghasilkan dokumen mirip yang lebih banyak namun belum cukup akurat menunjukkan adanya plagiasi, karena dalam menentukan plagiasi kesamaan rangkaian kalimat merupakan hal yang penting untuk diperhatikan.

Dalam penelitian Joko Priambodo (2018) dengan judul \_Pendeteksian Plagiarisme Menggunakan Algoritma Rabin - Karb dengan Meotde Rolling Hash\_, pendeteksian plagiarisme menggunakan algoritma Rabin-Karp dengan metode rolling hash dari hasil pengujian 30 dokumen teks yang sudah dijelaskan pada bab sebelumnya menghasilkan tingkat akurasi yang terbesar yaitu 47.58 %. Hasil persentase tersebut termasuk dalam kategori tingkat plagiat 15 - 50 %, berarti menandakan dokumen tersebut termasuk plagiat tingkat sedang. Sedangkan tingkat akurasi yang terkecil yaitu 19.28 %, berarti menandakan dokumen tersebut termasuk plagiat tingkat sedang. Selain itu, berdasarkan analisis proses pendeteksian tingkat plagiarisme menggunakan algoritma raibin-karp dengan metode rolling hash bisa membaca karakter berupa huruf, simbol seperti titik (.), koma (,), dan lain - lain.

Dalam penelitian Pavel Stefanovic (2019) dengan judul \_The N-Grams Based Text Similarity Detection Approach Using Self-Organizing Maps and Similarity Measures\_, dengan pendekatan berdasarkan teks yang dipecah menjadi n-gram dan mengevaluasinya menggunakan SOM dan similarity measure. Deteksi teks serupa dilakukan dalam tiga langkah: (1) konversi kumpulan data teks ke numerik ekspresi menggunakan n-gram; (2) perhitungan ukuran kesamaan; (3) visualisasi dataset teks menggunakan SOM dan representasi kesamaan di atasnya. Pada langkah pertama, fokus utamanya adalah membuat sekantong n-gram dari semua dataset. Berbagai jumlah kata dalam n-gram dianalisis. Selain itu, filter yang berbeda diterapkan: penghapusan angka dan tanda baca, frekuensi kata, transformasi huruf besar, stemming algoritma, dll. Analisis menunjukkan filter dan ukuran n-gram mempengaruhi hasil akhir. Untuk ini dataset, ukuran n-gram dipilih dan sama dengan tiga untuk penyelidikan eksperimental. Pada langkah kedua, empat ukuran kesamaan dihitung: cosinus, dadu, Jaccard diperpanjang, dan tumpang tindih. Hasil akhir menunjukkan bahwa persentase kemiripan tertinggi diperoleh dengan menggunakan overlap Pengukuran. Tiga nilai ukuran lainnya selalu sama dan lebih kecil. Penggunaan SOM menunjukkan bahwa SOM membantu untuk melihat hasil ringkasan kesamaan semua teks dalam bentuk visual dengan cepat. Dia sangat mudah untuk memahami teks mana yang mirip satu sama lain atau tidak. Dalam kasus kumpulan data yang dianalisis, SOM membantu mendeteksi kesamaan, dan cluster yang terbentuk dikorelasikan dengan kategoris yang diberikan deskripsi kumpulan data.

Dalam penelitian Uswatun Hasanah, dkk (2019) dengan judul \_Perbandingan Metode Cosine Similarity dan Jaccard Similarity Untuk Penilaian Otomatis Jawaban Pendek\_, dinilai belum mampu memberikan jawaban yang memuaskan i dikarenakan kedua metode hanya menilai kemiripan berdasarkan susunan leksikalnya. Sementara itu, jawaban mahasiswa juga sangat bervariasi dan menggunakan kata-kata yang jauh berbeda dari jawaban kunci, walaupun pada dasarnya memiliki makna semantik yang sama. Oleh karena itu, pada penelitian selanjutnya diperlukan metode lain yang mampu menangani makna semantik pada jawaban. Bagaimanapun, metode Cosine Similarity dan Jaccard Similarity masih dapat dipertimbangkan untuk menilai jawaban pendek secara otomatis, dengan batasan bahwa pertanyaan yang digunakan mengharuskan jawaban dalam format keyword sehingga tidak memunculkan kata-kata lain yang mampu menurunkan nilai kemiripan.

## 2.2 Landasan Teori

### 2.2.1 Sistem Informasi

Sistem Informasi terdiri dari dua kata yaitu Sistem dan Informasi. Sistem adalah diagram aktivitas atau activity diagram menggambarkan workflow (aliran kerja) atau aktivitas dari sebuah sistem atau proses bisnis atau menu yang ada pada perangkat lunak (Sukanto dan M.Salahuddin, 2015). Sedangkan informasi adalah sekumpulan data atau fakta yang diorganisasi atau diolah dengan cara tertentu sehingga mempunyai arti bagi penerima. (Anggerani dan Irvani, 2017). Jadi Sistem Informasi adalah sebuah aktivitas antara manusia dan teknologi yang memanajemen sebuah sistem yang mempermudah kegiatan manusia.

Dengan adanya sebuah Sistem Informasi yang semakin berkembang saat ini, maka akan mempermudah kehidupan manusia dalam mengelola sebuah data bahkan meningkatkan produktivitas serta meminimalisir terjadinya banyak kesalahan yang terjadi. Sebagai contoh pembelajaran E-Learning dimana dengan adanya sebuah sistem informasi ini dapat mempermudah orang untuk memperoleh ilmu dari mana saja dan kapan saja.

### 2.2.2 Hypertext Transfer Protocol Secure (HTTPS)

*Hypertext Transfer Protocol Secure* (HTTPS) adalah sebuah *protocol* komunikasi dalam suatu jaringan internet dengan keamanan yang lebih terjamin. Disebut lebih aman karena suatu perintah atau data yang dikirim melalui HTTPS ini dilindungi dengan sistem enkripsi sehingga menyulitkan hacker untuk membobol atau mencurinya (Fakhri Aziz Firmansyah, 2019). HTTPS merupakan tingkatan dari HTTP dimana yang membedakan di HTTPS terdapat (*Secure Socket Layer*) SSL dan (*Transport Layer Security*) TSL yang digunakan untuk mengamankan data yang disimpan atau yang akan dikirim. Menurut SSL Labs pada April 2018, 33,2% dari 1.000.000 situs web teratas Alexa menggunakan HTTPS sebagai default, 57,1% dari 137.971 situs web paling populer di Internet memiliki implementasi HTTPS yang aman, dan 70% dari pemuatan halaman (diukur oleh Firefox Telemetry) menggunakan HTTPS. Jadi saat ini website sekarang sudah mulai beralih dari sebelumnya HTTP ke HTTPS karena di HTTPS lebih aman daripada HTTP yang sudah lama.

### 2.2.3 Javascript

Javascript adalah sebuah bahasa pemrograman tingkat tinggi yang dinamis, *scripting*, *untyped*, dan *interpreter*. Javascript sendiri dibuat oleh Brendan Eich dari perusahaan Nestcape pada tahun 1994 yang diberi nama Mocha pada saat itu, kemudian berganti menjadi Livescript. Karena saat itu browser yang populer adalah Nestcape, Microsoft berusaha untuk mengalahkan popularitas browser tersebut dengan Internet Explorer dan melakukan Reverse Enginering terhadap Livescript dan terciptalah JScript pada tahun 1996. Karena terdapat dua browser yang besar yang berbeda, maka dibuatlah satu standar agar mempermudah pembuatan website saat itu dan dibuatlah ECMAScript. Menurut Douglas Rockford, "JavaScript, JScript, and ECMAScript 3 Silly Name for 1 Silly Language" yang berarti bahwa ketiga nama tersebut adalah bahasa yang sama yaitu Javascript.

#### 2.2.3.1 Karakteristik Javascript

Untuk Ekstensi dari *file* Javascript menggunakan ekstensi \\*.js. Contoh namaFile.js. Setiap akhir dari kode javascript dapat menggunakan ; ataupun tidak. Sedangkan *block scope* di Javascript menggunakan { }.

Javascript merupakan salah satu bahasa pemrogram yang *untyped* / *dynamicly typed* yang berarti tidak mendefiniskan terlebih dahulu tipe variabel yang akan didefinisikan. Untuk penamaan variabel, terdapat beberapa keyword yaitu var, let, dan const. Untuk var, dan let, nilai variabel dapat berubah atau diisi ulang, sedangkan const nilai variabelnya tidak dapat diubah. Perbedaan var dan let sendiri terletak pada hoistingnya. Sedangkan untuk penamaan variabel, hampir sama dengan beebrapa bahasa pemrograman lainnya seperti karakter pertama variabel tidak boleh angka, menggunakan penulisan camelCase, dan untuk keyword const yang merupakan variabel const biasanya menggunakan snake\_case dan semuanya huruf kapital. Contoh :

Javascript memiliki beberapa tipe data seperti *String*, *Number* atau *Integer*, *Boolean*, *Array*, *Object*, dan *Undifined*. Untuk mengetahui tipe data dari sebuah variabel dapat menggunakan *keyword* *typeof*. Contoh :

*Synchronous* dan *Asynchronous*, Secara sederhana, *Syncronus* dan *Asynchronous* merupakan tahapan dalam mengeksekusi sebuah kode dimana Synchronous mengeksekusi sebuah kode perbaris sesuai urutan kode yang dituliskan. Sedangkan *Asynchronous*, tidak selalu seperti *Synchronous*, tapi melihat waktu proses dari kode tersebut. Penggunaan Asynchronous tidak akan menunggu suatu kode selesai dijalankan, tetapi berlanjut ke kode selanjutnya. Contoh :

Jika melihat output dari dua program diatas antara menggunakan *Synchronous* dan *Asynchronous* terdapat perbedaan urutan dari hasil output, dimana Synchronous memiliki output sesuai dari urutan baris kode sedangkan Asynchronous berbeda dimana output yang dihasilkan berdasarkan jumlah waktu pengeksekusian dari kode tersebut yang disimulasikan menggunakan setTimeout() dari pengeksekusian paling cepat hingga yang paling lambat.

#### 2.2.3.2 NodeJS

Node.js adalah *runtime environment* untuk JavaScript yang bersifat *open-source* dan *cross-platform*. Dengan Node.js kita dapat menjalankan kode JavaScript di mana pun, tidak hanya terbatas pada lingkungan *browser*. Seperti diketahui bahwa Javascript hanya dapat berjalan pada sebuah web *browser*, kemudian Ryan Dahl, membuat sebuah *runtime environment* dengan mengeluarkan *engine* Javasscript dari Chrome yaitu V8 Javascript Engine menggunakan bahaasa C agar Javascript dapat dijalankan diluar browser. Akhirnya tercipta NodeJS pada tahun 2009. Dengan Begitu, Javascript yang sebelumnya hanya bisa di *client side* dengan adanya NodeJS bisa juga di *server side*. Beberapa fitur yang terdapat di NodeJS seperti *Asynchronous & Event-driven, Single Threaded but Highly Scalable*.

#### 2.2.3.3 ExpressJS

Menurut Website resmi dari Express JS adalah '*Fast, unopinionated, minimalist web framework for Nodejs*'. Express JS sendiri merupakan salah satu web *framework* khusus untuk NodeJS dimana kita dapat membuat sebuah *website* yang cepat, sederhana, dengan struktur yang tidak ditentukan atau tergantung dari pengguna Express JS sendiri.

### 2.2.4 *Database*

*Database* atau basis data adalah kumpulan informasi yang disimpan di dalam komputer secara sistematik sehingga dapat diperiksa menggunakan suatu program komputer untuk memperoleh informasi dari basis data tersebut (Andry Andaru). Berdasarkan definisi *database*, *database* berfungsi untuk menyimpan catatan atau sebuah data pada sebuah penyimpanan yang kemudian akan digunakan pada waktu lainnya. Pada umumnya, database sudah pasti memiliki *key* dan *value*, walaupun istilah ini berbeda ditiap - tiap jenis database yang ada. Adapun beberapa jenis database seperti *Operational Database*, *Database Warehouse*, *Distributed Database*, *Relational Database*, dan *End User Database*.

#### 2.2.4.1 SQL dan NoSQL

SQL dan NoSQL merupakan salah satu contoh dari *Relational Database* yang populer saat ini. SQL (*Structured Query Language*) merupakan bahasa yang digunakan untuk mengelola data secara *relational*. SQL sendiri memiliki ciri yaitu memiliki tabel yang terdiri dengan *row* atau *record* dan *field* yang bisa memiliki relasi dengan tabel lainnya. Untuk tiap data yang ada didalam tabel, harus memiliki skema yang sama untuk tiap recordnya. Contoh dari SQL seperti MySQL, PostgreSQL, dan MariaDB. Berbeda dengan NoSQL (*Not Only Structured Query Language*) yang merupakan bahasa untuk mengelola data secara *Non Relational*. Berbeda dengan SQL yang menggunakan tabel untuk menyimpan data. NoSQL memilik banyak jenis tempat untuk menyimpan data seperti *Document Database*, *Key-Value Database*, dan *Graph Database*. NoSQL juga tidak memiliki skema sehingga untuk menyimpan data bisa secara flexibel. Contoh dari NoSQL seperti MongoDB, Redis, Neo4j, dan Cassandra.

#### 2.2.4.2 MongoDB

MongoDB merupakan salah satu contoh dari NoSQL yang dikembangkan menggunakan bahasa pemrograman C++ yang rilis pertama kali pada tanggal 11 Februari 2009. MongoDB adalah salah satu contoh *Document Database* yang dimana tiap - tiap datanya merupakan sebuah JSON (*Javascript Object Notation*) atau dalam MongoDB disebut BSON (*Binary* JSON). Hingga saat ini, MongoDB sudah digunakan lebih dari 85 Juta pengguna diseluruh dunia dan sudah banyak perusahaan besar yang menggunakan database ini seperti EBay, Google, Adobe, dan EA. Salah satu contoh data yang terdapat di MongoDB :

### 2.2.5 Jurnal

Jurnal merupakan bagian dari jenis terbitan berseri yang ada diperpustakaan, adapun pengertian jurnal menurut *High Beam* “*Journal is the collection and periodic publication or transmission of news and the result of research through media*”, artinya bahwa jurnal merupakan suatu koleksi dan terbitan berkala atau transmisi mengenai berita dan hasil-hasil penelitian mengenai media. Jurnal sendiri terbagi atas dua format yaitu tercetak dan digital (*e-journal*). Untuk format digital jurnal dikemas dalam dua format , yaitu bentuk CD-ROM dan dalam bentuk akses secara online melalui internet. E-Journal dipahami sebagai publikasi ilmiah dalam format elektronik dan mempunyai ISSN (International Standard Serial Number) yang format dokumennya biasanya PDF (Rusydi 2014).

Penggunaan kata jurnal untuk berbagai bidang juga memberi arti yang bervariasi, misalnya jurnal dalam bidang ekonomi menunjukan sistem pembukuan rangkap. Jurnal dalam bidang pelayaran diartikan sebagai logbook berarti buku untuk mencatat semua kejadian selama pelayaran. Jurnal sebenarnya merupakan publikasi ilmiah yang memuat informasi tentang hasil kegiatan dalam bidang ilmu pengetahuan dan teknologi minimal harus mencakup kumpulan atau kumulasi pengetahuan baru, pengamatan empiris dan pengembangan gagasan atau usulan. Dengan demikian jurnal merupakan representasi dari pengetahuan baru tentang perkembangan ilmu pengetahuan yang dilaksanakan secara empriris dan biasanya merupakan gagasan yang terbaru.

### 2.2.6 Plagiarisme

Plagiarisme berasal dari bahasa Latin “*plagiare*” yang berarti mencuri. Plagiarisme berasal dari kata plagiat yang berarti 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. Sehingga dapat diartikan plagiarisme merupakan tindakan mencuri gagasan hasil penelitian orang lain, untuk kemudian disajikan seolah-olah milik sendiri (Ridhatillah, 2013). Tindakan plagiarisme merupakan salah satu tindakan yang melanggar hak cipta. Hak Cipta itu sendiri merupakan hak eksklusif untuk Pencipta ataupun penerima hak buat mengumumkan ataupun perbanyak Ciptaannya ataupun membagikan izin buat itu dengan tidak kurangi pembatasan- pembatasan bagi peraturan perundang- undangan yang berlaku. Jika terjadi pelanggaran tersebut, dapat dikenai pelanggaran hak cipta di Pasal 72 ayat UUHC dengan dipidana dengan pidana penjara pendek selama 1 bulan serta / ataupun denda sangat sedikitnya Rp1.000.000,00, ataupun pidana penjara lama 7 tahun serta / ataupun denda sangat sebanyak Rp5.000.000.000,00.

# BAB III METODE PENELITIAN