

LAPORAN AKHIR PROYEK SEARCH ENGINE

Dosen Pengampu : Bapak Herdiesel Santoso, S.T., S.Kom., M.Cs



Disusun Oleh :

Kelompok 2

Agung Prayuga	: 11200465
Ayyu Dian Ariyyati	: 11200483
Nabilah Desyarifah Anwar	: 11200462
Wahyu Kusnanda	: 11200474

SISTEM INFORMASI (SI)

**SEKOLAH TINGGI MANAJEMEN INFORMATIKA DAN ILMU
KOMPUTER EL-RAHMA YOGYAKARTA
TAHUN 2023**

Abstrak

Perkembangan Mesin Pencari saat ini menghasilkan tingkat recall dan level yang tinggi presisi rendah. Ingat mana yang tinggi untuk diartikan dokumen yang dihasilkan penelusuran banyak dokumen, sementara tingkat penyimpanan akurasi rendah dapat diinterpretasikan dokumen yang diharapkan itu dapat ditemukan sedikit/sedikit atau lebih sedikit. Solusi untuk mengatasi permasalahan di atas dengan membuat sistem informasi pengembalian pertemuan menggunakan metode Vector Model Luar Angkasa (VSM). Metode VSM dipilih karena cara modelnya efisien, mudah digunakan dalam representasi dan mendapatkan implementasi pada pencocokan dokumen.

Keywords: *Vektor Space Model, Web Scraping, TF-IDF, Search Engine*

KATA PENGANTAR

Alhamdulillah Puji dan syukur kehadiran Tuhan Yang Maha Esa yang telah melimpahkan rahmat dan hidayah-Nya kepada kami. Sehingga laporan ini dapat diselesaikan sebagai salah satu tugas akhir dari mata kuliah Sistem Temu Kembali Informasi. Adapun penulisan makalah ini bertujuan agar kita mengetahui cara membuat mesin pencari (Search Engine) sederhana dengan pengumpulan data dari metode web scraping. Tak lupa juga kami ucapkan terima kasih kepada Bapak Herdiesel Santoso S.T S.Kom. M.Cs selaku dosen pengampu mata kuliah Sistem Temu Kembali Informasi, yang telah memberikan arahan kepada kami selama masa perkuliahan berlangsung.

Kami menyadari jika pada laporan ini terdapat banyak kesalahan dan kekurangan, mungkin disebabkan karena keterbatasan pengetahuan maupun pengalaman. Oleh karena itu, kami sangat mengharapkan saran dan kritik yang membangun dari pembaca demi kesempurnaan makalah ini. Akhir kata semoga laporan ini dapat memberikan manfaat maupun inspirasi terhadap pembaca.

Daftar Isi

Abstrak.....	2
KATA PENGANTAR	3
Daftar Isi	4
BAB I PENDAHULUAN	5
1. Latar Belakang.....	5
1. Rumusan Masalah.....	7
2. Tujuan Makalah	7
3. Tools dan Alat yang Digunakan	7
BAB II DASAR TEORI	8
1. Search Engine	8
1.1. Pengertian Search Engine	8
1.2. Fungsi yang Dimiliki Search Engine	9
1.3. Bagaimana Cara Kerja Search Engine?	10
1.4. Macam-macam Search Engine	13
2. Web Scrapping	15
2.1. Pengertian web scrapping	15
2.2. Jenis-jenis Web Scrapping	16
2.3. Manfaat Web Scrapping	16
2.4. Kendala Web Scrapping	17
BAB III PEMBAHASAN	18
1. Metode Penelitian	18
1.1. Studi Literatur	18
1.2. Pengumpulan Data.....	18
1.2.1. Preprocessing	19
1.2.1. TF-IDF	21
1.2.2. Cosine and Similarity	22
1.2.3. Python.....	24
1.2.4. PHPMyAdmin.....	25
Fungsi dari PHPMyAdmin.....	25
1. Pelaksanaan Proyek	26
BAB IV PENUTUP.....	36
DAFTAR PUSTAKA.....	37

BAB I

PENDAHULUAN

1. Latar Belakang

Penerapan teknologi digital dan jaringan komputer telah menyebabkan terjadinya “ledakan” informasi yang berkembang eksponensial. Hal ini menyebabkan Sistem temu kembali informasi (information retrieval =IR) mengalami kesulitan. Information Retrieval (IR) merupakan bagian dari computer science yang berhubungan dengan pengambilan informasi dari dokumendokumen yang didasarkan pada isi dan konteks dari dokumen-dokumen itu sendiri. Informasi yang diinginkan pengguna direpresentasikan dalam bentuk query dan mengandung satu atau lebih term yang akan digunakan dalam pencarian. Search engine atau mesin pencari merupakan teknik dari temu-kembali dalam menemukan dokumen dan sekaligus mengeksekusi algoritma peringkat dalam menampilkan dokumen. Pengguna dapat mencari halaman web yang dibutuhkan melalui search engine. Search engine tidak lain sebuah mesin pencari yang ulet dan teliti, yang melakukan eksplorasi atas informasi-informasi yang di-request tanpa memandang kapan, di mana dan oleh siapa itu dilakukan. Mesin pencari menggunakan indeks (yang sudah dibuat dan disusun secara teratur) untuk mencari file setelah pengguna memasukkan kriteria pencarian. Indexing atau pengindeksan merupakan proses membangun basis data indeks dari koleksi dokumen. Indexing dilakukan terhadap dokumen sebelum pencarian dilakukan.

Dalam penelitian yang dilakukan oleh Heninggar Septiantri yang membandingkan metode LSA (Latent Semantic Analysis) dan VSM (Vector Space Model) mengenai sistem penilai jawaban esai otomatis Bahasa Indonesia, didapatkan dari hasil uji coba bahwa secara keseluruhan rata-rata korelasi nilai VSM-manusia lebih tinggi dari LSAMANUSIA. Oleh karena itu metode yang akan digunakan untuk penelitian ini adalah metode Vector Space Model. Vector Space Model adalah suatu metode untuk merepresentasikan sistem temu kembali ke dalam vektor dan memperhitungkan fungsi similarity dalam proses pencocokan beberapa vektor. Untuk melakukan perhitungan fungsi similarity terlebih dahulu akan dilakukan perhitungan pembobotan dari masing-masing dokumen dan keyword menggunakan Metode Pembobotan TF-IDF. Metode TF-IDF (Term Frequency-Inverse Document Frequency) merupakan suatu cara untuk memberikan bobot hubungan suatu kata (term) terhadap dokumen. Metode ini

menggabungkan dua konsep untuk perhitungan bobot yaitu, frekuensi kemunculan sebuah kata didalam sebuah dokumen tertentu dan inverse frekuensi dokumen yang mengandung kata tersebut. Pada studi kasus ini metode Vector Space Model (VSM) digunakan untuk memodelkan kumpulan berita dan keyword dari user dalam bentuk vektor yang telah di beri bobot dengan menggunakan metode pembobotan TF-IDF, kemudian akan di hitung kedekatan dari masing-masing dokumen dengan keyword dari user menggunakan cosine similarity.

Dinas Komunikasi dan Informatika Daerah Istimewa Yogyakarta merupakan sebuah instansi pemerintah yang memiliki tugas membantu Gubernur melaksanakan urusan pemerintahan bidang komunikasi dan informatika dan urusan pemerintahan bidang persandian. Dinas Komunikasi dan Informatika Daerah Istimewa Yogyakarta memiliki beberapa proyek yang dikembangkan salah satunya adalah Jogja Center, sebuah sistem analitik berbasis big data, yang difokuskan pada pengembangan data analitik dan pendukung pengambilan keputusan, serta merujuk pada dimensi-dimensi Jogja Smart Province (JSP).

Sistem analitik yang dikembangkan memerlukan data yang perlu diperbarui secara berkala untuk memberikan hasil analisis yang lebih baik. Sebelum ditampilkan di halaman Jogja Center, data-data tersebut perlu dikumpulkan menggunakan cara yang disebut dengan web scraping. Hal tersebut dilakukan dengan membuat suatu program yang dapat mengambil data dari berbagai sumber, salah satunya dapat diimplementasikan pada web berita detik.com yang kemudian data tersebut dapat disimpan dalam bentuk format dokumen Comma Separated Values atau CSV. Data-data tersebut berupa headline atau judul berita, tanggal, link source, dan lain-lain yang kemudian dapat dianalisis untuk menentukan sebaran mobilitas masyarakat maupun sentimen masyarakat yang kemudian ditampilkan.

Web scraping merupakan teknik yang digunakan untuk mengumpulkan data yang dapat dilakukan secara manual yaitu dengan melakukan copy paste data yang diinginkan maupun secara otomatis yaitu dengan membuat sebuah program atau code yang dapat melakukan proses pengambilan data dari sebuah halaman web. Dalam melakukan web scraping terdapat beberapa metode yang dapat dilakukan yaitu menyalin data secara manual, parsing HTML, regular expression, dan lain-lain. Tidak dipungkiri teknik web scraping memiliki kekurangan yaitu sampai saat ini belum ada teknik web scraping yang 100% efektif. Selain itu hasil yang didapatkan tidak selalu rapih, maka perlu juga untuk memahami struktur halaman website yang dituju, karena

tidak semua data dapat diekstrak dengan mudah, sering kali program harus dijalankan berulang kali yang mengakibatkan akses terhadap halaman web tersebut terblokir. Namun ada pula manfaat dari melakukan web scraping yaitu data-data yang didapatkan akan lebih terfokus yang dapat memudahkan dalam pencarian sesuatu.

2. Rumusan Masalah

- a. Apa itu Search engine?
- b. Bagaimana teknik web scraping data?
- c. Bagaimana mengolah data hasil scraping?
- d. Bagaimana membuat dokumen dari hasil scraping?
- e. Bagaimana membuat search engine dengan data dokumen hasil scraping?

3. Tujuan Makalah

- a. Memahami cara scraping data pada website
- b. Memahami teknik pengolahan data dengan bahasa pemrograman python
- c. Memahami struktur pembentuk search engine

4. Tools dan Alat yang Digunakan

- a. Python (proses scraping)
- b. Bootstrap 5 (proses pembuatan search engine)
- c. Phpmyadmin (basis data)

BAB II

DASAR TEORI

1. Search Engine

1.1. Pengertian Search Engine

Search engine atau apabila diartikan dalam bahasa Indonesia adalah mesin pencari pada dasarnya merupakan program berbasis web yang diperuntukkan untuk mencari informasi di dalam World Wide Web (www). Pencarian informasi melalui search engine bisa didapatkan dengan menyesuaikan dengan kata kunci yang pengguna masukkan.

Teknologi dari search engine dapat memberikan informasi yang diinginkan pengguna dengan daftar pencarian terbaik yang tersedia. Proses menghasilkan informasi tersebut biasa disebut sebagai SERP atau search engine result page. Sejak pertama kali diciptakan, banyak sekali search engine yang ada di dunia saat ini, salah satu yang sangat terkenal tentu saja Google. Pengguna yang ingin menggunakan Google bisa melalui berbagai perangkat yang dimiliki, dari browser yang ada smartphone, tablet, komputer, dan berbagai perangkat browser lainnya.

Selanjutnya, setelah memasukkan kata kunci yang Kamu ingin ketahui, hasil yang ditampilkan oleh gagal tidak akan sama dengan apa yang ditampilkan oleh Bing, begitupun juga dengan Yahoo, terlebih lagi urutan daftar yang dihasil. Setiap search engine menggunakan algoritma yang berbeda-beda pada saat mengindex dan mengumpulkan data, secara otomatis hasil yang diberikan belum tentu sama.

Pada intinya, setiap search engine dilengkapi dengan sistem yang berbeda-beda dalam menangkap apa yang Kamu inginkan. Hasil yang akan diberikan oleh search engine akan disesuaikan dengan lokasi pencarian, apa yang sedang dicari hingga keinginan yang paling banyak diinginkan pengguna. Masing-masing search engine menggunakan sistem dan proses pencarian yang unik dan menjadi karakteristik dari setiap search engine.

1.2. Fungsi yang Dimiliki Search Engine

Search engine pada dasarnya memiliki fungsi utama sebagai alat untuk menyediakan informasi bagi semua orang. Ketika hampir semua orang menggunakan mesin pencarian, lantas apa sebenarnya fungsi yang ditawarkan oleh search engine itu sendiri. Ketika menggunakan search engine, pengguna yang ingin atau membutuhkan suatu informasi hanya perlu masukkan kata kunci dalam sistem mesin pencarian. Selanjutnya, berbagai daftar web yang berkaitan dengan kata yang dimasukkan akan ditampilkan kepada pengguna. Langkah seperti biasa disebut dalam dunia komputer sebagai crawling atau proses mengumpulkan data atau mengindeks. Dapat dilihat dari fungsi pertama yang sudah dijelaskan, semua pengguna bisa melakukan akses melalui search engine untuk mendapatkan informasi apapun. Mulai informasi tentang cuaca, media sosial, barang yang ingin dibeli, bahkan berbagai jenis barang, dengan catatan telah dimuat di sistem WWW.

Membahas tentang pencarian dan penjualan sebuah produk, search engine bukan hanya alat yang dapat digunakan untuk menemukan informasi saja. Perkembangan saat ini sekaligus fungsi kedua dari adanya mesin pencarian yakni dapat digunakan untuk memaksimalkan mengoptimalkan sebuah bisnis, seperti memasarkan produk. Semakin bertambahnya hari, semakin banyak orang yang menggunakan search engine untuk memaksimalkan dan memenuhi kebutuhan sehari-harinya. Hal inilah yang pada akhirnya memunculkan pengoptimalan mesin pencari sebagai ladang untuk melakukan bisnis online secara luas. Apabila sebelum munculnya search engine, seseorang berjualan dengan jangkauan sekitar dan terbatas. Kini, seseorang bisa berjualan dengan jangkauan yang sangat luas. Sekarang ini, pengguna yang membutuhkan sebuah barang, bisa langsung bertemu secara online dengan pengguna lain yang menjual barang tersebut. Kegiatan transaksi ini tidak lagi terbatas oleh wilayah, karena semua orang bisa terhubung dengan menggunakan internet dan menemukan dengan mesin pencari.

Hanya dengan modal kata kunci, seseorang bisa mendapatkan banyak informasi tentang produk yang sedang diinginkan dan dibutuhkan. Hal itu pada akhirnya akan sangat memudahkan para penjual untuk mengetahui berapa banyak sebuah produk yang dicari pada suatu daerah, sehingga akan sangat

memudahkan untuk melakukan proses iklan melalui search engine. Sistem periklanan yang disediakan search engine yaitu misalnya seperti Google Ads dari Google. Oleh karena itu, saat ini, search engine menjadi salah satu alat yang memiliki peran yang sangat penting untuk kehidupan manusia, dari yang hanya untuk mencari sebuah informasi hingga untuk memaksimalkan pemasaran sebuah produk ke luar daerah hingga luar negara.

1.3. Bagaimana Cara Kerja Search Engine?

Perannya yang amat sangat penting membuat keberadaan search engine hampir ada di setiap elemen dari kehidupan manusia. Namun, bagaimana sebenarnya cara kerja dari search engine tersebut, dan apabila setiap search engine memiliki hasil pencarian yang berbeda-beda, bagaimana sistem kerjanya. Nah, walaupun search engine memiliki banyak macam dan menghasilkan informasi yang berbeda-beda, pada dasarnya cara kerja dari search engine tidaklah berbeda. Dalam sistem kerja search engine melakukan tiga tahap untuk mendapatkan hasil pencarian, yakni meliputi crawling, indexing dan tentunya menciptakan hasil.

1. Crawling

Dalam sebuah mesin pencari hampir dapat dipastikan memiliki crawlers dan robot pencarinya sendiri-sendiri. Crawling pada dasarnya dapat diartikan sebagai proses yang digunakan mesin pencari yakni bots atau spiders. Kedua proses mesin pencari tersebut bisa dimanfaatkan untuk melakukan kunjungan dan download pada sebuah halaman dan mengekstrak link dengan tujuan menemukan halaman tertentu. Halaman yang sudah diketahui oleh mesin pencari dapat masuk dalam proses crawling secara periodik, hal ini dilakukan agar mesin pencari dapat menentukan apakah terdapat perubahan dari waktu terakhir pada sebuah konten. Apabila mesin pencari menemukan adanya perubahan, maka secara otomatis akan dilakukan update pada halaman tersebut. Dalam melakukan crawling, mesin pencari akan menggunakan angka algoritma dan peraturan. Hal itu dilakukan agar mesin pencari dapat menentukan seberapa sering halaman dalam dijelajahi ulang. Selain itu, mesin pencari juga dapat

melakukan index pada berapa banyak halaman dalam situs web. Algoritma pada mesin pencarian sebetulnya dapat digunakan untuk menghasilkan pencarian yang relevan dan yang berkualitas. Berkat adanya algoritma ini, pengguna dapat menemukan berbagai informasi dari pertanyaan dengan bentuk kata kunci yang dimasukkan.

2. Indexing

Setelah melakukan proses crawler dalam melakukan penjelajahan pada jutaan halaman yang ada di web, mesin pencarian akan mengubahnya menjadi sebuah struktur data yang biasa disebut dengan index. Index atau indexing dapat dipahami sebagai proses untuk menemukan URL secara bersamaan dengan sebuah angka yang relevan sesuai bentuk sinyal yang diberikan. Sinyal yang digunakan pada proses index terdiri dari empat jenis yaitu, jenis pertama adalah kata kunci atau biasa disebut dengan keywords. Kata kunci yang ditemukan di dalam sebuah halaman konten web dapat ditemukan berdasarkan pembahasan dari topik pada halaman tersebut.

Selanjutnya, jenis yang kedua adalah tipe konten yang pernah dilakukan penjelajahan dengan menggunakan Schema microdata dan informasi yang masuk ke dalam halaman tersebut. Sementara, jenis yang ketiga merupakan kebaruan dari halaman tersebut atau waktu terakhir dilakukan update. Terakhir, pada jenis keempat yakni kapan terakhir kali halaman tersebut dikunjungi oleh pengguna atau bagaimana proses interaksi pengguna lain dengan menggunakan halaman web. Pengguna dapat memilih hasil manakah yang paling relevan dari web yang di index untuk kemudian nantinya mempengaruhi pencarian berikutnya.

Berikut ini adalah alasan dari beberapa kasus yang dapat membuat sebuah URL tidak masuk index dari search engine, di antaranya adalah:

- a. Adanya robots.txt yang menginformasikan kepada mesin pencari agar tidak melakukan kunjungan pada laman tersebut.
- b. Tidak adanya tag index dapat membuat search engine supaya tidak melakukan indexing pada web tersebut atau halaman yang mirip.
- c. Mesin pencarian memberikan pengkategorian pada laman tersebut dikarenakan memiliki kualitas yang rendah atau memiliki konten yang kurang atau bahkan konten duplikasi.
- d. Munculnya error pada URL.

Adanya beberapa masalah yang perlu untuk segera diperbaiki pada laman hingga web sehingga tidak dapat masuk dalam proses index.

3. Ranking

Setelah menyelesaikan proses index dan mendapatkan hasil dari banyak URL yang terkumpul, selanjutnya maka mesin pencari akan melakukan ranking. Proses ranking berarti adalah proses membuat daftar hasil berdasarkan yang paling relevan dengan kata kunci. Ketika menggunakan search engine, laman yang menjadi peringkat teratas merupakan hasil pencarian yang paling relevan. Hal itu sama halnya bahwa search engine percaya bahwa web atau laman tersebut berkaitan dengan kata kunci. Dalam melakukan pengaturan relevansi, mesin pencarian mengandalkan sebuah sistem algoritma.

Dalam perkembangannya, hampir setiap tahun sebuah algoritma mengalami banyak perubahan yang disesuaikan dan didasarkan pada apa yang didapatkan selama ini. Sebagai contoh, saat ini, setiap hari Google melakukan pembaharuan pada sistem algoritmanya. Pada setiap algoritma yang diatur dan digunakan dilakukan guna memperkecil potensi masalah. Contoh yang paling mudah dilihat adalah penciptaan Penguin untuk mengatasi spam. Algoritma telah berkembang dan akan selalu diikuti oleh para pengguna mesin pencari untuk memaksimalkan dan mengoptimalkan tujuan bisnis mereka.

1.4. Macam-macam Search Engine

Google telah berhasil mendapatkan kepercayaan dari para penggunanya, kualitas dan hasil yang diberikannya selama ini telah membuktikan kehebatannya. Mesin pengolahan algoritma dari Google dikenal memiliki kemampuan untuk meneliti dengan sangat baik dan mampu menyajikan hasil yang sangat akurat. Sekarang ini, search engine yang paling memuaskan pengguna dan paling populer di dunia adalah Google. Google hampir pasti bisa disebut sebagai raja mesin pencarian, hal itu dikarenakan jumlah pengguna yang terlampau banyak. Namun, selain Google, apa Kamu tahu bahwa ada search engine lain yang juga banyak digunakan contohnya adalah sebagai berikut:

1. Bing

Search engine alternatif pertama setelah Google adalah Bing. Saat ini, Bing memiliki pengguna dengan persentase penggunaan melalui desktop sebesar 2,55% dan 12,60% melalui smartphone. Bing pada dasarnya merupakan search engine dari Microsoft yang diciptakan pada tahun 2009. Bing sendiri dibuat oleh Microsoft untuk menghentikan dominasi dari Google.

Bing awalnya merupakan gabungan dari tiga search engine yakni MS search, Windows Live search dan Live Search. Selanjutnya, mesin pencari ini secara otomatis dapat digunakan pada Windows PC.

2. Yahoo

Yahoo adalah search engine kedua setelah Google yang juga sebagai provider email, hingga saat ini, Yahoo masuk dalam jajaran ketiga dengan penguasaan pasar hingga mencapai 2%. Pada Oktober 2011 hingga Oktober 2015, Yahoo berada di bawah kepemilikan Bing. Setelahnya itu, pihak Google juga ingin memiliki share market Yahoo.

Namun, tepatnya pada Oktober 2019, Yahoo akhirnya berhasil diakuisisi kembali secara eksklusif oleh Bing. Yahoo sebenarnya merupakan mesin pencari default dari browser Firefox yang dibuat di Inggris sejak tahun 2014. Berdasarkan Alexa, Yahoo menjadi salah satu portal web yang banyak dikunjungi di dunia.

3. Baidu

Selanjutnya, Baidu merupakan search engine yang hingga saat ini menguasai pasar dengan pengguna sebanyak 0,7% pada desktop dan 11,8% pada smartphone. Search engine yang dibangun pada tahun 2000 ini merupakan search engine yang sangat populer di Cina. Walaupun dapat dijangkau hampir seluruh dunia, tetapi Search engine ini merupakan mesin pencari yang menggunakan bahasa Cina.

Berdasarkan ranking yang dibuat oleh Alexa, pada saat ini, Baidu berhasil berada di peringkat ke 4 sebagai search engine yang paling banyak digunakan. Baidu sendiri menyediakan banyak fitur seperti berita, peta, hingga penyimpanan dengan cloud.

4. Yandex

Setelah mengetahui search engine dari Amerika Serikat, Inggris hingga Cina, selanjutnya adalah search engine yang berasal dari Rusia yaitu Yandex. Yandex sendiri merupakan mesin pencari yang menguasai pasar dengan pengguna sebanyak 0,45% pada perangkat komputer dan 1,41% pada perangkat mobile.

Berdasarkan ranking yang dibuat oleh Alexa, Yandex merupakan search engine yang berada dalam 30 urutan website paling populer dan menjadi ranking keempat. Di negara Rusia, Yandex menjadi search engine terbesar dan terpopuler dengan jumlah presentasi hingga mencapai 65%. Tidak hanya itu, Yandex juga berhasil menjadi presentasi dari perusahaan teknologi yang membuat produk machine learning.

5. Duck Duck Go

Duck Duck Go merupakan search engine yang menguasai pasar hingga sekitar 0,42%. Setiap harinya, mesin pencari ini digunakan sebanyak 47 juta pengguna. Tidak seperti search engine terkenal lainnya, Duck Duck Go tidak melakukan indexing, tetapi search engine tersebut menyajikan hasil pencarian dari berbagai macam sumber.

Hal itu menunjukkan bahwa search engine dari Duck Duck Go tidak dilengkapi penyimpanan data sendiri, tetapi masih bergantung dengan mesin pencarian lain seperti Yahoo dan Bing. Keterbatasan yang dimiliki inilah yang membuat Duck Duck Go kalah apabila dibandingkan dengan Google. Namun, kelebihan dari menggunakan Duck Duck Go yakni tampilan yang bersih, tidak melakukan tracking pada pengguna, dan yang terpenting tidak dipenuhi dengan iklan.

2. Web Scrapping

2.1. Pengertian web scrapping

Web scraping adalah salah satu metode mengumpulkan data yang digunakan untuk mengekstraksi data dari suatu halaman web. Halaman web dibangun menggunakan bahasa berbasis teks (HTML dan XHTML), sering kali berisi banyak data berguna dalam bentuk teks. Sebagian besar halaman web dirancang untuk kemudahan penggunaan oleh manusia, bukan untuk kemudahan penggunaan otomatis. Akibatnya, alat dan perangkat lunak khusus telah dikembangkan untuk melakukan web scraping. (Sahria, 2022)

Web scraping secara hukum merupakan hal yang sah untuk dilakukan selama data yang dikumpulkan digunakan untuk kepentingan pribadi dan dalam tidak melanggar undang-undang yang berlaku. Jika data akan dipublikasikan atau jika konten memiliki hak cipta dan melanggar persyaratan layanan maka ada beberapa preseden hukum yang perlu diperhatikan. Di *Feist Publications, Inc. v. Rural Telephone Service Co.*, Mahkamah Agung Amerika Serikat memutuskan bahwa scraping dan penerbitan ulang data, seperti daftar telepon, diizinkan. Hal ini serupa dengan *Australia, Telstra Corporation Limited v. Phone Directories Company Pty Ltd*, menunjukkan bahwa hanya data dengan penulis yang dapat diidentifikasi dilindungi dengan hak cipta. (Jarmul, 2017)

Kasus lainnya di Amerika Serikat, mengevaluasi penggunaan kembali berita *Associated Press* untuk produk berita gabungan, dinyatakan sebagai pelanggaran hak cipta di *Associated Press v. Meltwater*. Hal ini menunjukkan bahwa ketika data yang didapatkan merupakan data publik seperti lokasi bisnis

dan daftar telepon, data tersebut dapat diterbitkan ulang dengan mengikuti aturan yang berlaku. Namun, jika data tersebut merupakan opini, ulasan atau data pengguna pribadi, kemungkinan besar tidak dapat dipublikasikan ulang karena alasan hak cipta. (Jarmul, 2017)

Web scraping biasanya dibuat untuk menargetkan situs web atau situs tertentu dengan tujuan mengumpulkan informasi spesifik di situs tersebut. Scraper dibuat untuk mengakses halaman spesifik dan perlu dimodifikasi jika situs berubah atau jika informasi yang ada di situs diubah. Sebaliknya, web crawling biasanya dibuat dengan cara umum, yaitu menargetkan situs web dari serangkaian domain tingkat atas atau seluruh web. Crawler dapat dibuat untuk 6 mengumpulkan informasi yang lebih spesifik, tetapi umumnya digunakan untuk mengambil sedikit informasi umum dari berbagai situs atau halaman yang berbeda dan mengikuti tautan ke halaman lain.

2.2. Jenis-jenis Web Scrapping

Dalam melakukan web scraping terdapat beberapa teknik yang umumnya digunakan. Yang pertama, menyalin data secara manual. Melakukan web scraping secara manual merupakan bentuk paling sederhana yang dapat dilakukan. Terkadang teknologi untuk melakukan web scraping terbaik sekalipun tidak dapat menggantikan pemeriksaan manual yang dilakukan oleh manusia. Teknik yang kedua merupakan parsing HTML. Dengan teknik ini, web scraping dapat dilakukan tidak hanya pada halaman website yang bersifat statis, tetapi juga dinamis. Parsing HTML juga memungkinkan untuk menyalin data dalam jumlah yang besar dalam waktu singkat.

2.3. Manfaat Web Scrapping

Dalam penggunaannya web scraping memiliki beberapa manfaat. Yang pertama, dapat membandingkan ulasan dalam skala besar karena web scraping pada dasarnya adalah mengumpulkan data. Yang kedua, dapat digunakan untuk mencari informasi. Sebagai contoh ketika ingin memulai suatu bisnis, pencarian informasi menggunakan web scraping dapat menjadi salah satu cara untuk

mengetahui apakah produk atau jasa yang akan dikembangkan merupakan sesuatu yang dibutuhkan dan diminati oleh masyarakat.

Dengan menggunakan web scraping proses mengumpulkan data dapat dilakukan dengan lebih cepat. Selain itu, jika data yang dikumpulkan jumlahnya sangat besar, proses tersebut juga dapat dilakukan secara otomatis. Web scraping dapat membantu untuk menganalisa data, karena mampu mengumpulkan data secara detail dan efisien. (Josi, 2014)

2.4. Kendala Web Scrapping

Dalam melakukan proses web scraping terdapat beberapa hal yang menjadi kendala. Yang pertama adalah belum adanya teknik web scraping yang efektif. Yang kedua, data yang didapatkan tidak terstruktur dengan baik. Hal tersebut dapat terjadi dikarenakan adanya sisasisa teks yang tidak diinginkan seperti contoh tag HTML. Yang ketiga adalah dapat terjadi 7 pemblokiran terhadap alamat IP baik secara manual maupun berdasarkan geolocation dari alamat IP tersebut.

Terdapat metode lain yang digunakan oleh beberapa situs web untuk mencegah dilakukannya web scraping, seperti mendeteksi dan melarang bot melihat halaman mereka dan juga menonaktifkan API layanan web yang mungkin terekspose oleh sistem.

BAB III PEMBAHASAN

1. Metode Penelitian

1.1. Studi Literatur

Studi literatur penulis gunakan untuk memahami konsep dari search engine atau mesin pencari dan metode Vector Space Model (VSM). Literatur yang digunakan adalah beberapa jurnal paper tentang Information Retrieval, Search engine, pembobotan TF-IDF, Preprocessing, VSM, Recall dan Precision, sehingga menambah pemahaman penulis akan tema yang diambil.

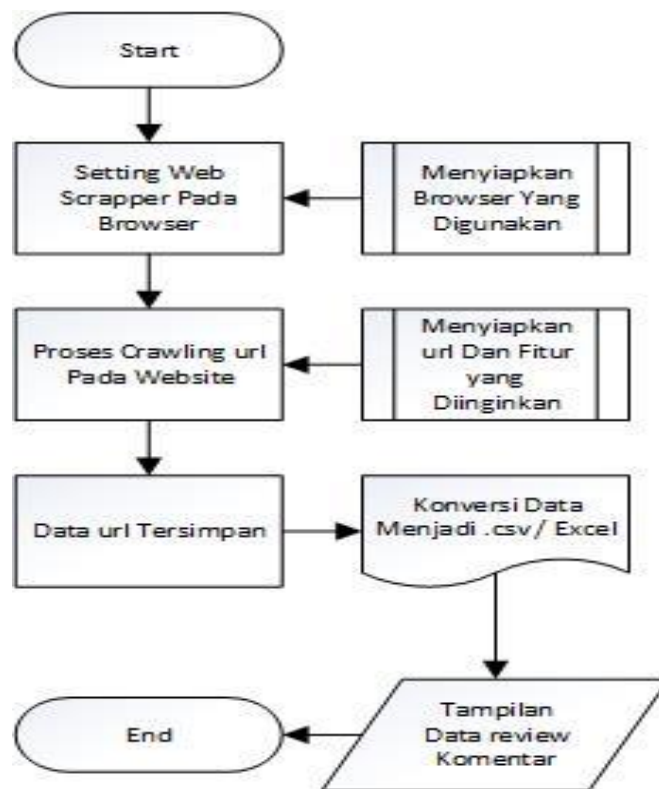
1.2. Pengumpulan Data

Pengumpulan data dilakukan untuk mengumpulkan data-data yang akan digunakan untuk melihat rating sebuah aplikasi shoppe dengan fitur pencarian berbasis search engine. Pengumpulan data dilakukan dengan scraping data dari play store.

Beberapa alasan pengumpulan data menggunakan metode web scraping:

- a. Web scraping adalah salah satu metode mengumpulkan data yang digunakan untuk mengekstraksi data dari suatu halaman web. Halaman web dibangun menggunakan bahasa berbasis teks (HTML dan XHTML), sering kali berisi banyak data berguna dalam bentuk teks. Sebagian besar halaman web dirancang untuk kemudahan penggunaan oleh manusia, bukan untuk kemudahan penggunaan otomatis. Akibatnya, alat dan perangkat lunak khusus telah dikembangkan untuk melakukan web scraping.
- b. Dalam melakukan web scraping terdapat beberapa teknik yang umumnya digunakan. Yang pertama, menyalin data secara manual. Melakukan web scraping secara manual merupakan bentuk paling sederhana yang dapat dilakukan. Terkadang teknologi untuk melakukan web scraping terbaik sekalipun tidak dapat menggantikan pemeriksaan manual yang dilakukan oleh manusia. Teknik yang kedua merupakan parsing HTML. Dengan teknik ini, web scraping dapat dilakukan tidak hanya pada halaman website yang bersifat statis, tetapi juga dinamis. Parsing HTML juga memungkinkan untuk menyalin data dalam jumlah yang besar dalam waktu singkat.

- c. Dengan menggunakan web scraping proses mengumpulkan data dapat dilakukan dengan lebih cepat. Selain itu, jika data yang dikumpulkan jumlahnya sangat besar, proses tersebut juga dapat dilakukan secara otomatis. Web scraping dapat membantu untuk menganalisa data, karena mampu mengumpulkan data secara detail dan efisien.



Gambar Alur Pengambilan Data

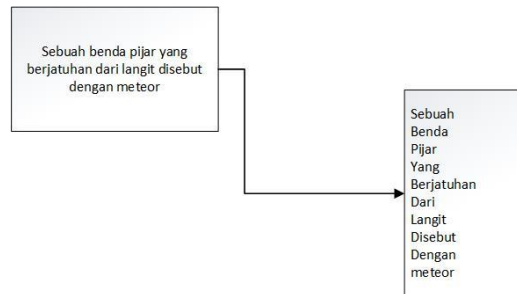
Langkah pengumpulan data scraping sebagai berikut:

1.2.1. Preprocessing

Pra-pemrosesan mengacu pada transformasi yang diterapkan pada data kami sebelum memasukkannya ke algoritme. Data Preprocessing adalah teknik yang digunakan untuk mengubah data mentah menjadi kumpulan data bersih. Dengan kata lain, setiap kali data dikumpulkan dari sumber yang berbeda, data tersebut dikumpulkan dalam format mentah yang tidak layak untuk dianalisis.

1. Tokenizing

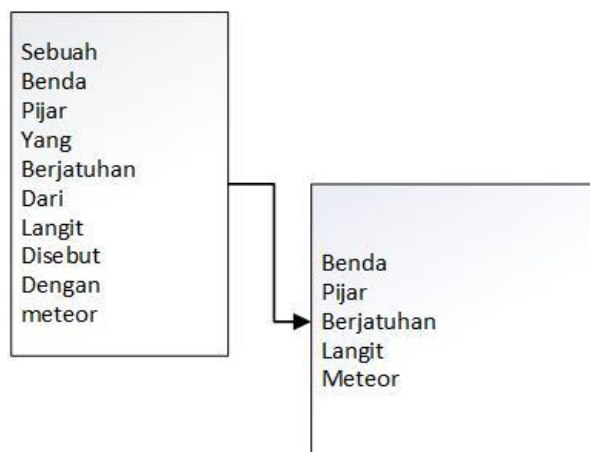
Tokenizing merupakan tahap pemotongan string input berdasarkan tiap kata yang menyusunnya. Contoh dari tahap ini adalah sebagai berikut:



Gambar Proses Tokenizing

2. Filtering

Filtering merupakan tahap pengambilan kata-kata penting hasil token. Bisa menggunakan algoritma stoplist (membuang kata yang kurang penting) atau wordlist (menyimpan kata penting). Stoplist / stopword adalah kata-kata yang tidak deskriptif yang dapat dibuang dalam pendekatan bag-of-words. Contoh stopwords adalah “yang”, “dan”, “di”, “dari” dan seterusnya. Contoh dari tahapan ini adalah sebagai berikut:



Gambar Proses Filtering

1.2.1. TF-IDF

TF-IDF (Term Frequency Inverse Document Frequency) merupakan metode yang digunakan untuk menentukan nilai frekuensi sebuah kata di dalam sebuah dokumen atau artikel dan juga frekuensi di dalam banyak dokumen. Perhitungan ini menentukan seberapa relevan sebuah kata di dalam sebuah dokumen (Evan, 2014). TFIDF adalah sebuah algoritma yang umumnya digunakan untuk pengolahan data besar (Kamath, 2014).

Algoritma TF-IDF melakukan pemberian bobot pada setiap kata kunci disetiap kategori untuk mencari kemiripan kata kunci dengan kategori yang tersedia. Sebelum melakukan pembobotan maka akan dilakukan lima tahap pencarian text preprocessing yaitu pemecahan kalimat, case folding, tokenizing, filtering, dan stemming, lalu selanjutnya dilakukan proses menghitung 18 bobot TF-IDF, bobot query relevance dan bobot similarity (Marlinda & Rianto, 2013).

Berdasarkan penelitian-penelitian sebelumnya, yang membahas tentang penerapan metode TF-IDF. Penulis menemukan banyak terdapat variasi formula dalam mengimplementasikan metode TF-IDF pada pembobotan kata. Nilai TF-IDF meningkat secara proporsional berdasarkan jumlah atau banyaknya kata yang muncul pada dokumen, tetapi diimbangi dengan frekuensi kata dalam korpus. Variasi dari skema pembobotan TF-IDF sering digunakan oleh mesin pencari sebagai alat utama dalam mencetak nilai (scoring) dan peringkat (ranking) sebuah relevansi dokumen yang diberikan user. TF-IDF pada dasarnya merupakan hasil dari perhitungan antara TF (Term Frequency) dan IDF (Inverse Document Frequency). Banyak cara untuk menentukan nilai yang tepat dari kedua statistik yang ada. Dalam kasus term frequency $tf(t, d)$, cara yang paling sederhana adalah dengan menggunakan raw frequency di dalam dokumen, yaitu berapa kali term t muncul di dokumen d . Jika menyatakan raw frequency t sebagai $f(t, d)$, maka skema tf yang sederhana adalah $tf(t, d) = f(t, d)$.

Nilai idf sebuah term (kata) dapat dihitung menggunakan persamaan sebagai berikut:

$$IDF = \log_{10}\left(\frac{d}{dfi}\right)$$

D adalah jumlah dokumen yang berisi term (t) dan dfi adalah jumlah kemunculan (frekuensi) kata terhadap D.

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

$$W_{d,t} = tf_{d,t} * IDF_t$$

Keterangan :

d = dokumen ke-d

t = kata ke-t dari kata kunci

W = bobot dokumen ke-d terhadap kata ke-t

tf = term frekuensi/frekuensi kata

Setelah bobot (W) masing-masing dokumen diketahui, maka dilakukan proses pengurutan (sorting) dimana semakin besar nilai W, semakin besar tingkat kesamaan (similarity) dokumen tersebut terhadap kata yang dicari, demikian pula sebaliknya.

1.2.2. Cosine and Similarity

Kesamaan kosinus adalah metrik yang digunakan untuk menentukan seberapa mirip dokumen terlepas dari ukurannya. Secara matematis, Cosine similarity mengukur kosinus sudut antara dua vektor yang diproyeksikan dalam ruang multidimensi. Dalam konteks ini, dua vektor yang saya bicarakan adalah larik yang berisi jumlah kata dari dua dokumen. Ketika diplot pada ruang multi-dimensi, di mana setiap dimensi sesuai dengan kata dalam dokumen, kesamaan kosinus menangkap orientasi (sudut) dokumen dan bukan besarnya. Jika Anda menginginkan besarnya, hitunglah jarak Euclidean sebagai gantinya.

Kesamaan kosinus menguntungkan karena meskipun dua dokumen serupa terpisah jauh dengan jarak Euclidean karena ukurannya (seperti, kata 'kriket' muncul 50 kali dalam satu dokumen dan 10 kali dalam dokumen

lain) mereka masih dapat memiliki sudut yang lebih kecil. diantara mereka. Semakin kecil sudutnya, semakin tinggi kesamaannya. Model ruang vektor dan pembobotan tfidf digunakan untuk merepresentasikan nilai numerik dokumen sehingga kemudian dapat dihitung kedekatan antar dokumen. Kemiripan antar dokumen dihitung menggunakan suatu fungsi ukuran kemiripan (similarity measure). Semakin besar hasil fungsi similarity, maka kedua objek yang dievaluasi semakin mirip, demikian pula sebaliknya. Ukuran ini memungkinkan perankingan dokumen sesuai dengan kemiripan (relevansi)nya terhadap query. Kualitas hasil dari dokumen yang didapatkan sangat tergantung pada fungsi similarity yang digunakan.[6] Jika terdapat dua vektor dokumen d dan query q serta term t diperoleh dari dokumen maka nilai cosinus didefinisikan sebagai:

$$\cos \alpha = \frac{A \cdot B}{|A||B|} = \frac{\sum_{i=1}^n A_i \times B_i}{\sqrt{\sum_{i=1}^n (A_i)^2} \times \sqrt{\sum_{i=1}^n (B_i)^2}}$$

Rumus Cosine Similarity

Keterangan :

A = Vektor A , yang akan dibandingkan kemiripannya

B = Vektor B , yang akan dibandingkan kemiripannya

$A \cdot B$ = dot product antara vektor A dan vektor B

$|A|$ = panjang vektor A

$|B|$ = panjang vektor B

$|A||B|$ = cross product antara $|A|$ dan $|B|$

1.2.3. Python

Python merupakan bahasa pemrograman yang diciptakan oleh Guido Van Rossum, dapat dilihat pada Gambar 2.1. Guido Van Rossum menciptakan python pada awal tahun 1990 untuk keperluan umum di Centrum voor Wiskunde and Informatica (CWI)



Gambar Guido Van Rossum

Sebelum menciptakan bahasa pemrograman Python, Guido Van Rossum telah mengerjakan bahasa pemrograman lainnya, yaitu bahasa pemrograman ABC, yang dikembangkan di CWI sebagai bahasa pengajaran yang menekankan kejelasan. Meskipun proyek ABC pada akhirnya ditutup, van Rossum mengambil banyak pelajaran ketika ia mulai menulis Python sebagai alat untuk digunakan dalam proyek penelitian sistem operasi dan multimedia.

Python adalah bahasa pemrograman yang dinamis. Hirarki nomornya mencakup bilangan bulat asli dengan panjang arbitrer, titik mengambang presisi perangkat keras, bilangan kompleks, dan dukungan library untuk bilangan rasional dan titik mengambang presisi arbitrer. Python juga memiliki string yang kuat, daftar ukuran variabel, set, dan array asosiatif yang sangat fleksibel disebut sebagai dictionaries dalam Python. Jenis-jenis ini memberi Python kosakata yang beragam untuk mengekspresikan banyak pertanyaan algoritmik

kompleks dengan kejelasan dan efisiensi. Python menggabungkan fleksibilitas tingkat tinggi, keterbacaan, dan antarmuka yang terdefinisi dengan baik dengan kemampuan tingkat rendah, termasuk antarmuka C resmi yang memungkinkan untuk memperluas bahasa dengan kode pemrograman C dan menautkan ke library pihak ketiga di C, C++, dan Fortran. Hal ini bermanfaat untuk komputasi ilmiah modern sehingga memberikan fleksibilitas untuk membangun alat dengan keseimbangan yang tepat antara fitur tingkat rendah dan tinggi, dapat memilih dengan tepat antara kinerja dan kemudahan pengembangan atau penggunaan. (Perez, 2011)

1.2.4. PHPMyAdmin

PHPMyAdmin merupakan sebuah aplikasi berbasis web yang berfungsi untuk mengelola database MySQL atau bisa disebut juga sebagai tool database. Siapapun sebenarnya tidak salah jika mempelajarinya, karena aplikasi ini akan sangat berguna dalam pengembangan situs web yang saat ini semakin populer, misalnya WordPress yang memerlukan akses ke database.

Software berbasis web ini akan memudahkan Anda untuk melakukan manipulasi database MySQL tanpa harus mengetikkan perintah pada command line. Aplikasi ini memiliki tampilan yang mudah dipahami dengan fitur lengkap sesuai kebutuhan para pengguna.

Fungsi dari PHPMyAdmin

Setelah kita mengetahui apa itu phpmyadmin, selanjutnya kita akan menjelasnya fungsinya secara garis besar. Sehingga teman-teman akan mengetahui banyaknya manfaat yang bisa didapatkan jika menggunakan phpmyadmin. PHPMyAdmin berfungsi untuk membuat, mengedit, menghapus database, tabel, serta membuat atau menghapus relasi antar tabel, mensortir data, dan lain-lain sesuai dengan kebutuhan Anda. Saat menggunakannya, teman-teman akan mendapatkan kemudahan dengan cara yang lebih efektif dalam pembuatan database menuju web server.

PHPMyAdmin adalah software yang mempunyai fasilitas import yang bisa Anda manfaatkan untuk membuat database dengan ekstensi .SQL. Lalu pada versi offline, teman-teman hanya cukup memindahkan ke versi web yang

tersedia. Selain itu, saat membuat tabel-tabel pada database, pertimbangkan juga tentang primary key dan foreign key serta relasi datanya. Pada proyek yang kita kerjakan, kami menggunakan PHPMyAdmin sebagai fasilitas open source dalam pembuatan search engine.

1. Pelaksanaan Proyek

Berikut adalah langkah langkah proyek yang kami lakukan

1. Untuk memanggil directory pada google drive silahkan tuliskan script sebagai berikut:

```
[ ] from google.colab import drive
    drive.mount('/content/drive')

Mounted at /content/drive
```

2. Pertama install library python yaitu google scraper

```
Install Google scraper Sebagai Library

[ ] !pip install google-play-scraper

Looking in indexes: https://pypi.org/simple, https://us-python.pkg.dev/colab-wheels/public/simple/
Collecting google-play-scraper
  Downloading google_play_scraper-1.2.2-py3-none-any.whl (28 kB)
Installing collected packages: google-play-scraper
Successfully installed google-play-scraper-1.2.2

[ ] from google_play_scraper import app

import pandas as pd

import numpy as np
```

3. Scraping review komentar aplikasi shopee pada google play store

```
Scraping Review Komentar Aplikasi Shopee Pada Google Play Store

#scrape all available review

from google_play_scraper import Sort, reviews

result, continuation_token = reviews(
    'com.shopee.id',
    lang='id', # defaults to 'en'
    country='id', # defaults to 'us'
    sort=Sort.MOST_RELEVANT, # defaults to Sort.MOST_RELEVANT you can use Sort.NEWEST to get newst reviews
    count=100, # defaults to 100
    filter_score_with=None # defaults to None(means all score) Use 1 or 2 or 3 or 4 or 5 to select certain score
)
```

4. Untuk melihat hasil scrapping bisa lakukan seperti dibawah ini

```
df_busu = pd.DataFrame(np.array(result), columns=['review'])

df_busu = df_busu.join(pd.DataFrame(df_busu.pop('review').tolist()))

df_busu.head()
```

	reviewId	username	userImage	content	score	thumbsUpCount	reviewCreatedVersion	at	replyContent	repliedAt
0	913e34f7-9d07-4268-bd3e-3a8ae5ce92c	Rian Sub	https://play-lh.googleusercontent.com/a/AEoFTp...	Aplikasi Super Lemot, Sering error. 1. Aplikasi...	2	1909	2.96.24	2023-01-17 10:10:25	Hi kak, maaf ya udi buat km ga nyaman terkail...	2023-01-13 07:00:58
1	09e03cab-fb6e-4f54-a876-7677dae2470	achmad wafiq pasya	https://play-lh.googleusercontent.com/a/AEoFTp...	Update bukannya tambah bagus malah tambah ngeb...	1	2222	2.96.24	2023-01-15 11:44:53	Hi kk achmad, maaf ya udah buat ganyaman. Saat...	2023-01-16 12:15:10
2	aa10d0f5-6760-4e04-9067-0c94572e3464	Lee Rosse	https://play-lh.googleusercontent.com/a/AEoFTp...	Harganya pada murah murah, tapi ongkirnya kadia...	1	1508	2.95.52	2023-01-18 08:13:06	hi kak, maaf utk kendala nya, mimin saranin un...	2023-01-18 09:04:32
3	b287339f-6b21-42d3-958c-ad118eb2631a	Budy Bucheq	https://play-lh.googleusercontent.com/a-/ADS-W...	Sekarang semenjak diperbarui ketika dibuka/mem...	3	111	2.95.52	2023-01-17 02:16:46	hi kak, maaf utk kendala nya, mimin saranin un...	2023-01-17 04:07:04
4	f40fb6f8-6836-4116-a21d-4a193245a568	Rizky Oktia Guslian	https://play-lh.googleusercontent.com/a-/ADS-W...	dapat balasan update terbaru lah, jelas ngakt...	1	77	2.95.52	2023-01-17 23:01:41	Hi kak, maaf ya kak bikin kamu ga nyaman, terk...	2023-01-14 04:21:19

5. Kemudian tuliskan script berikut untuk mengambil beberapa kolom data yang diinginkan

```
[ ] df_busu[['userName', 'score', 'at', 'content']].head()
```

	userName	score	at	content
0	Rian Sub	2	2023-01-17 10:10:25	Aplikasi Super Lemot, Sering eror. 1. Aplikasi...
1	achmad wafiq pasya	1	2023-01-15 11:44:53	Update bukannya tambah bagus malah tambah ngeb...
2	Lee Rosse	1	2023-01-18 08:13:06	Harganya pada murah murah, tapi ongkirnya kada...
3	Budy Bucheq	3	2023-01-17 02:16:46	Sekarang semenjak diperbarui ketika dibuka/mem...
4	Rizky Okta Gustian	1	2023-01-17 23:01:41	dapat balasan update terbaru lah , jelas" ngakt...

6. Kemudian berikut adalah script untuk mengubah data yang sudah di scrapping ke dalam bentuk csv

```
▶ my_df = df_busu[['userName', 'score', 'at', 'content']]
```

```
[ ] my_df.to_csv("data-shopee.csv", index = False)
```

7. Kemudian selanjutnya adalah text preprocessing, untuk melakukan itu kita membutuhkan library nltk tidak lupa untuk install sastrawi juga seperti pada gambar dibawah ini

```
import nltk

nltk.download('punkt')
nltk.download('stopwords')

[nltk_data] Downloading package punkt to /root/nltk_data...
[nltk_data] Unzipping tokenizers/punkt.zip.
[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data] Unzipping corpora/stopwords.zip.
True

[11] import re
      from nltk import word_tokenize
      from nltk.corpus import stopwords
      from nltk.stem import PorterStemmer

pip install Sastrawi

Looking in indexes: https://pypi.org/simple, https://us-python.pkg.dev/colab-wheels/public/simple/
Collecting Sastrawi
  Downloading Sastrawi-1.0.1-py2.py3-none-any.whl (209 kB)
    ----- 209.7/209.7 KB 8.9 MB/s eta 0:00:00
Installing collected packages: Sastrawi
Successfully installed Sastrawi-1.0.1
```

8. Kemudian kita lakukan case folding : lowercase , proses ini bertujuan untuk merubah format teks upercase menjadi format huruf kecil semua (lowercase)

▼ Lowercase

```
import pandas as pd

df = pd.read_csv('/content/data-shopee.csv', encoding='utf-8')
def clean_lower(lwr):
    lwr = lwr.lower()
    return lwr

df['Hasil'] = df['content'].apply(clean_lower)
casefolding=pd.DataFrame(df['Hasil'])
casefolding

df.to_excel('/content/drive/MyDrive/STKII/lowecase-data-shopee.xlsx', index=False)
```

9. Untuk membaca data yang sudah kita lowercase bisa kita read pada excel seperti berikut

```
[ ] data = pd.read_excel('/content/drive/MyDrive/STKII/lowercase-data-shopee.xlsx')
```

	userName	score	at	content	Hasil
0	Rian Sub	2	2023-01-17 10:10:25	Aplikasi Super Lemot, Sering error. 1. Aplikasi...	aplikasi super lemot, sering error. 1. aplikasi...
1	achmad wafiq pasya	1	2023-01-15 11:44:53	Update bukannya tambah bagus malah tambah ngeb...	update bukannya tambah bagus malah tambah ngeb...
2	Lee Rosse	1	2023-01-18 08:13:06	Harganya pada murah murah, tapi ongkirnya kada...	harganya pada murah murah, tapi ongkirnya kada...
3	Budy Bucheq	3	2023-01-17 02:16:46	Sekarang semenjak diperbarui ketika dibuka/mem...	sekarang semenjak diperbarui ketika dibuka/mem...
4	Rizky Okta Gustian	1	2023-01-17 23:01:41	dapat balasan update terbaru lah , jelas" ngakt...	dapat balasan update terbaru lah , jelas" ngakt...
...
95	alin tiro	3	2022-12-07 11:35:09	Tepat waktu, sayangnya udah ga bisa dibuka di a...	tepat waktu, sayangnya udah ga bisa dibuka di a...
96	rofiqoh aja	1	2023-01-17 12:17:27	Sekarang apk nya mengecewakan, bikin emosi trs...	sekarang apk nya mengecewakan, bikin emosi trs...
97	Dani Setiawan	1	2022-12-23 05:16:54	Parah banget setelah di update bukannya tambah...	parah banget setelah di update bukannya tambah...
98	Nisa Aulia	1	2023-01-10 03:14:04	Sekarang kok aplikasinya lemot ya, padahal jar...	sekarang kok aplikasinya lemot ya, padahal jar...
99	a k u	1	2023-01-15 12:26:50	Sekarang lemot banget, gara2 ada feed video, f...	sekarang lemot banget, gara2 ada feed video, f...

100 rows x 5 columns

10. Untuk menghapus kata-kata yang memiliki informasi rendah dari sebuah teks kita lakukan stopwords dengan cara berikut

```
import pandas as pd
import numpy as np
from nltk.corpus import stopwords
import re
import nltk
nltk.download('stopwords')

df = pd.read_csv('/content/lowercase-data-shopee.csv', encoding = 'utf-8')

#clean stopwords
stopword = set(stopwords.words('english'))

add = pd.DataFrame(df['content'])
df['before'] = add.replace(to_replace=['yang', 'dari', 'paling', 'itu', 'nya', 'yg'], value = "", regex = True)

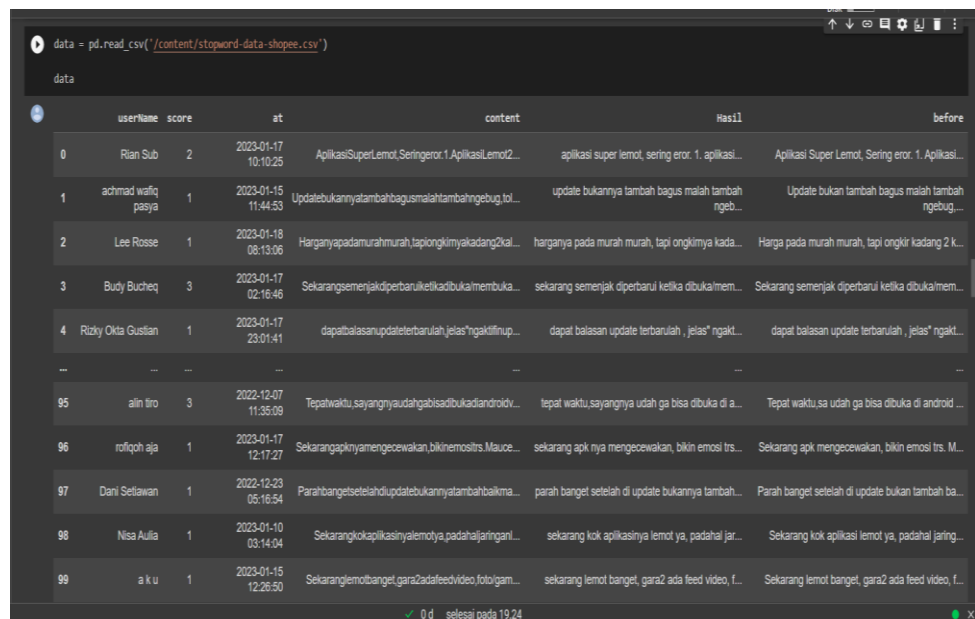
df['before']

def clean_stopwords(text):
    df = ''.join(word for word in text.split() if word not in stopword)
    return df

df['content'] = df['content'].apply(clean_stopwords)
df.to_csv('stopword-data-shopee.csv', index = False)
```

[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data] Package stopwords is already up-to-date!

11. Untuk membaca data yang sudah dilakukan proses stopwords bisa kita read pada excel dengan cara berikut



```
data = pd.read_csv("/content/stopword-data-shopee.csv")
```

	userName	score	at	content	Hasil	before
0	Rian Sub	2	2023-01-17 10:10:25	AplikasiSuperLemot, Sering error. 1. Aplikasi Lemot2...	aplikasi super lemot, sering error. 1. aplikasi...	Aplikasi Super Lemot, Sering error. 1. Aplikasi...
1	ahmad wafiq pasya	1	2023-01-15 11:44:53	Updatebukanntambahbagusmalahambahngebug.tol...	update bukannya tambah bagus malah tambah ngeb...	Update bukan tambah bagus malah tambah ngeb...
2	Lee Rosse	1	2023-01-18 08:13:06	Harganyapadamurahmurah,tapiongkimyakadang2kal...	harganya pada murah murah, tapi ongkimya kada...	Harga pada murah murah, tapi ongkir kadang 2 k...
3	Budy Bucheq	3	2023-01-17 02:16:46	Sekarangsemenjakdiperbaruitakadibuka/membuka...	sekarang semenjak diperbarui ketika dibuka/mem...	Sekarang semenjak diperbarui ketika dibuka/mem...
4	Rizky Okta Guslan	1	2023-01-17 23:01:41	dapatbalasanupdateterbaruah,jelas"ngaklinup...	dapat balasan update terbaruah , jelas" ngakt...	dapat balasan update terbaruah , jelas" ngakt...
...
95	alin tiro	3	2022-12-07 11:35:09	Tepatwaktu,sayangnyaudahgabisdibuka/diandroi...	tepat waktu,sayangny udah ga bisa dibuka di a...	Tepat waktu,sa udah ga bisa dibuka di android ...
96	rotiqoh aja	1	2023-01-17 12:17:27	Sekarangapknymengecewakan,bikinemositrs.Mauce...	sekarang apk nya mengecewakan, bikin emosi trs...	Sekarang apk mengecewakan, bikin emosi trs. M...
97	Dani Seliawan	1	2022-12-23 05:16:54	Parahbangetsetelahdiupdatebukanntambahbaikma...	parah banget setelah di update bukannya tambah...	Parah banget setelah di update bukan tambah ba...
98	Nisa Aulia	1	2023-01-10 03:14:04	Sekarangkokaplikasiylemotya,padahaljarin...	sekarang kok aplikasinya lemot ya, padahal jar...	Sekarang kok aplikasi lemot ya, padahal jaring...
99	a k u	1	2023-01-15 12:26:50	Sekaranglemotbanget,gara2adafeedvideo,foto/gam...	sekarang lemot banget, gara2 ada feed video, f...	Sekarang lemot banget, gara2 ada feed video, f...

0 d selesai pada 19:24

12. Kemudian berikut adalah script untuk mengubah data stopwords yang sudah di scrapping ke dalam bentuk csv

```
[ ] my_df = df_busu[['userName', 'score','at', 'content']]

[ ] my_df.to_csv("stopword-data-shopee.csv", index = False)
```

13. Selanjutnya adalah salah satu bagian penting dari information Retrieval adalah proses stemming. Stemming adalah proses mereduksi kata berimbuhan menjadi kata dasar.

Stemming

```
import nltk
nltk.download('stopwords')
nltk.download('wordnet')
from nltk.stem import PorterStemmer
from nltk.stem import LancasterStemmer

df = pd.read_csv('/content/data-shopee.csv', encoding = 'utf-8')

ps = nltk.PorterStemmer()
ls = nltk.LancasterStemmer()
list_word = pd.DataFrame(df)
print("{0:20}{1:20}{2:20}".format("Word", "Porter Stemmer", "Lancaster Stemmer"))

for word in list_word:
    print("{0:20}{1:20}{2:20}".format(word, ps.stem(word), ls.stem(word)))

df['Hasil'] = df['content'].apply(clean_lower)
casefolding=pd.DataFrame(df['Hasil'])
casefolding

df.to_excel('/content/drive/MyDrive/STKII/stemming-data-shopee.xlsx', index=False)
```

```
[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data] Package stopwords is already up-to-date!
[nltk_data] Downloading package wordnet to /root/nltk_data...
Word          Porter Stemmer    Lancaster Stemmer
username      usernam          usernam
score         score            scor
at            at              at
content       content          cont
```

14. Kemudian berikut adalah script untuk mengubah data stemming yang sudah di scrapping ke dalam bentuk csv

```
[20] data = pd.read_csv('/content/stopword-data-shopee.csv')
```

data

	userName	score	at	content
0	Rian Sub	2	2023-01-17 10:10:25	Aplikasi Super Lemot, Sering error. 1. Aplikasi...
1	achmad wafiq pasya	1	2023-01-15 11:44:53	Update bukannya tambah bagus malah tambah ngeb...
2	Lee Rosse	1	2023-01-18 08:13:06	Harganya pada murah murah, tapi ongkirnya kada...
3	Budy Bucheq	3	2023-01-17 02:16:46	Sekarang semenjak diperbarui ketika dibuka/mem...
4	Rizky Okta Gustian	1	2023-01-17 23:01:41	dapat balasan update terbaru lah , jelas" ngakt...
...
95	alin tiro	3	2022-12-07 11:35:09	Tepat waktu, sayangnya udah ga bisa dibuka di a...
96	rofiqoh aja	1	2023-01-17 12:17:27	Sekarang apk nya mengecewakan, bikin emosi trs...
97	Dani Setiawan	1	2022-12-23 05:16:54	Parah banget setelah di update bukannya tambah...
98	Nisa Aulla	1	2023-01-10 03:14:04	Sekarang kok aplikasinya lemot ya, padahal jar...
99	a k u	1	2023-01-15 12:26:50	Sekarang lemot banget, gara2 ada feed video, f...

100 rows x 4 columns

15. Berikut adalah proses preprocessing atau clean content

```
def clean(text):
    # Menjadikan Setiap Kalimat Menjadi Huruf Kecil / Lower Text
    a = text.lower()

    # Menghapus Tanda Baca, Huruf dan Kata Yang Tidak di Perlukan
    b = re.sub('[^A-Za-z]+', ' ', a)

    # Melakukan Tokenisasi
    b = word_tokenize(b)

    # Menghapus stop words
    english_stopwords = stopwords.words('english')

    tokens_wo_stopwords = []
    for word in b:
        if word not in english_stopwords:
            tokens_wo_stopwords.append(word)







    # Proses Stemming
    ps = PorterStemmer()

    tokens_wo_stemming = []
    for word in tokens_wo_stopwords:
        tokens_wo_stemming.append(ps.stem(word))

    return " ".join(tokens_wo_stemming)

df_busu['Clean content'] = df_busu['content'].apply(clean)

df_busu
```

	reviewId	username	userImage	content	score	thumbsUpCount	reviewCreatedVersion	at	replyContent	repliedAt	Clean content	
0	913e3477-9d0-4268-bd3e-3aace5ced32c	Rian Sub		https://play-ih.googleusercontent.com/va/AE6FTp...	Aplikasi Super Lemot, Sering error. 1. Aplikasi...	2	2055	2.96.24	2023-01-17 10:10:25	2023-01-13 07:00:58	Hi kak, maaf ya udah buati km ga nyaman A terkait... aplikasi super lemot sene error aplikasi lemot...	
1	09ed20b-8ffe-4f54-a875-7677dae0470	achmad waq pasya		https://play-ih.googleusercontent.com/va/AE6FTp...	Update bukannya tambah bagus malah tambah ngebu...	1	2330	2.96.24	2023-01-15 11:44:53	2023-01-15 12:15:10	Hi kak achmad, maaf ya udah buati ganyaman. Saat... updat bukannya tambah bagus malah tambah ngebug...	
2	aa10c095-6780-4e04-8067-0c94572a3484	Lee Roslie		https://play-ih.googleusercontent.com/va/AE6FTp...	Harganya pada murah murah, tapi ongkirnya kada...	1	1633	2.95.52	2023-01-18 08:13:06	2023-01-18 09:04:32	Hi kak, maaf utk kendala nya, mimin saratin un... harganya pada murah murah tapi ongkirnya kadan...	
3	b2873338-6b21-42d3-958c-ad1f8e2831a	Budy Bucheq		https://play-ih.googleusercontent.com/va/AE6FTp...	Sekarang semenjak diperbaru ketika dibuka/mem...	3	118	2.95.52	2023-01-17 02:15:46	2023-01-17 04:07:04	Hi kak, maaf utk kendala nya, mimin saratin un... sekarang semenjak diperbaru ketika dibuka mem...	
4	4a2b69b-6836-4116-a219-4a193245a598	Riky Oktia Guslan		https://play-ih.googleusercontent.com/va/AE6FTp...	dapat balasan update terbaru lah , jelas" ngakt...	1	83	2.95.52	2023-01-17 23:01:41	2023-01-14 04:21:19	Hi kak, maaf ya kak bikin kamu ga nyaman, terk... dapat balasan updat terbaru lah, jelas ngaklfin...	
95	79b8a4e-bf9f-4628-a53f-70b1c2963010	alin tiro		https://play-ih.googleusercontent.com/va/AE6FTp...	Topat waktu sayangnya udah ga bisa dibuka di a...	3	127	2.95.52	2022-12-07 11:35:09	2022-03-10 04:55:01	Yuhu mantap banget nih ka , makasih ya buati b... topat waktu sayangnya udah ga bisa dibuka di a...	
96	dc3f0b04-8c52-4903-bd00-26b2e21c2060	rofiah aja		https://play-ih.googleusercontent.com/va/AE6FTp...	Sekarang apk nya mengecewakan, bikin emosi trs...	1	4	2.95.52	2023-01-17 12:17:27	None	NaT	sekarang apk nya mengecewakan bikin emosi tr m...
97	a2321960-5ae3-469d-919e-62b6e4c5e6b9	Dani Setiawan		https://play-ih.googleusercontent.com/va/AE6FTp...	Parah banget setelah di update bukannya tambah...	1	639	2.95.52	2022-12-23 05:16:54	2019-02-08 08:53:46	2023-01-09 03:10:21	Hi kak, terima kasih atas rating dan feedback... parah banget setelah di updat bukannya tambah...
98	a4fa4a21-a702-467f-a271-3533e009a536	Nisa Aulia		https://play-ih.googleusercontent.com/va/AE6FTp...	Sekarang kok aplikasinya lemot ya, sudah lah ter...	1	8	2.95.52	2023-01-10 03:14:04	2023-01-09 03:10:21	Hi kak, maaf ya udah bikin km ganyaman, mohon... sekarang kok aplikasinya lemot ya sudah lah ter...	

16. Selanjutnya kita lakukan tf idf pada data atau dokumen yang sudah kita dapatkan sebelumnya

```

from sklearn.feature_extraction.text import TfidfTransformer
from sklearn.feature_extraction.text import CountVectorizer

cv = CountVectorizer()
word_count_vector = cv.fit_transform(data['content'])

tf = pd.DataFrame(word_count_vector.toarray(), columns = cv.get_feature_names_out())

tf.to_csv('/content/tf-data-shopee.csv', index=False)

[26] tfidf_transformer = TfidfTransformer()

X = tfidf_transformer.fit_transform(word_count_vector)
idf = pd.DataFrame({'Name Feature': cv.get_feature_names_out(), 'bobot idf':tfidf_transformer.idf_})

idf.to_csv('/content/idf-data-shopee.csv', index=False)

tf_idf = pd.DataFrame(X.toarray(), columns=cv.get_feature_names_out())

tf_idf.to_csv('/content/tf_idf-data-shopee.csv', index=False)

```

Berikut adalah data tf idf

[illegible]

17. Selanjutnya kita lakukan perhitungan cosine pada data yang ada

```
▼ Cosin Similarity

[ ] from sklearn.feature_extraction.text import TfidfVectorizer
    from sklearn.metrics.pairwise import cosine_similarity

    # Initialize an instance of tf-idf Vectorizer
    tfidf_vectorizer = TfidfVectorizer()

    # Generate the tf-idf vectors for the corpus
    tfidf_matrix = tfidf_vectorizer.fit_transform(df_busu['clean content'])

    # compute and print the cosine similarity matrix
    cosine_sim = cosine_similarity(tfidf_matrix, tfidf_matrix)
    print(cosine_sim)

[[1.          0.11104031 0.10438126 ... 0.08498003 0.03911634 0.16224406]
 [0.11104031 1.          0.07178418 ... 0.24438995 0.06943514 0.06564968]
 [0.10438126 0.07178418 1.          ... 0.03331612 0.0267493  0.07806607]
 ...
 [0.08498003 0.24438995 0.03331612 ... 1.          0.01857305 0.10514046]
 [0.03911634 0.06943514 0.0267493  ... 0.01857305 1.          0.1444266 ]
 [0.16224406 0.06564968 0.07806607 ... 0.10514046 0.1444266  1.          ]]
```

18. Berikut adalah tampilan skrip untuk Search Engine

```
import sys
import os
import glob
import re
import codecs
import nltk
import string
import math
import numpy as np

from nltk.corpus import stopwords
from tqdm import tqdm_notebook
from bs4 import BeautifulSoup, Tag
from google.colab import drive

nltk.download('punkt')
nltk.download('stopwords')

IN_COLAB = 'google.colab' in sys.modules

path = None
if IN_COLAB:
    drive.mount('/content/gdrive')
    path = "/gdrive/My Drive/SearchEngine/"
else:
    path = "./SearchEngine/"

[1] [nltk data] Downloading package punkt to /root/nltk data...
```

BAB IV PENUTUP

1. Kesimpulan

Search engine memiliki miliaran halaman yang terdapat pada databasenya. Sehingga ketika pengguna internet memasukkan kata di search engine, hanya dalam hitungan detik, search engine langsung menampilkan website-website yang berkaitan, dengan penempatan website yang berbeda-beda. Search engine menampilkan website yang berkaitan dari yang paling baik hingga yang dianggap paling buruk.. Untuk itu, kita harus mengerti akan search engine adalah untuk mempermudah manusia memperoleh informasi, bahkan untuk memenuhi beberapa kebutuhannya tanpa batas waktu dan tempat.

2. Saran

Berdasarkan hasil penelitian ini maka saran yang dapat peneliti berikan adalah :

1. Bagi Pembaca

Bagi pembaca pergunakan gadget sesuai dengan k-ebutuhan, kemudian lebih teliti menggunakan search engine dalam mencari informasi, terkhususkan informasi dalam pembelajaran dan menjadikan motivasi belajar remaja menjadi meningkat.

2. Bagi Peneliti Selanjutnya

Mengingat bahwa motivasi belajar sangat penting didalam proses belajar. Dan motivasi juga dipengaruhi oleh banyak faktor, maka hendaklah peneliti selanjutnya meneliti faktor lain, baik dari instrinsik maupun ekstrinsik

DAFTAR PUSTAKA

<file:///C:/Users/ThinkFat/Downloads/1127-Article%20Text-1627-1-10-20131011.pdf>

https://elera.stmikelrahma.ac.id/pluginfile.php/5006/mod_resource/content/1/PENGUNAAN%20METODE%20COSINESIMILARITY%20PADA%20SISTEM%20PENGELOMPOKAN%20KP%20CTA%20dan%20SKRIPSI.pdf

https://elera.stmikelrahma.ac.id/pluginfile.php/5005/mod_resource/content/1/2019_Pembuatan%20Sistem%20Pencarian%20Pekerjaan%20Menggunakan%20TF-IDF.pdf

https://www.gramedia.com/literasi/pengertian-search-engine/#A_Pengertian_Search_Engine

<http://e-journal.uajy.ac.id/10907/4/3TF07164.pdf>

<https://www.linkedin.com/pulse/how-scrape-google-play-reviews-4-simple-steps-using-python-kundi/>

<https://colab.research.google.com/drive/1NvSKY8N3USo9Asy2cIEumS07vIoLIpio#scrollTo=8BUxU3ZyrMQh>