BAB I.
PENDAHULUAN

1.1 Latar Belakang

Meningkatnya perkembangan teknologi informasi banyak membawa dampak baik positif maupun negatif dalam kehidupan masyarakat. Sejak ditemukannya komputer pada tahun 1955, kemajuan teknologi diberbagai belahan dunia telah memasuki era informasi. Pemrosesan informasi berbasis komputer mulai dikenal orang hingga saat ini, sudah banyak *software* serta teknologi informasi lainnya yang dapat digunakan orang sebagai alat pengelolaan data untuk menghasilkan informasi yang dibutuhkan oleh masyarakat. Jumlah data digital yang berisi informasi teks dokumen yang dapat berupa teks artikel berita dengan cepat bertambah setiap tahunnya.

Pesatnya perkembangan teknologi informasi yang ada membuat banyak instansi-instansi baik dari lembaga pemerintahan maupun swasta memiliki sistem informasi mereka sendiri, sehingga memungkinkan masyarakat luas untuk mendapatkan informasi dengan mudah dan cepat. Situs atau laman *web* merupakan salah satu bentuk atau bagian dari sistem informasi.

Laman *web* atau yang biasa disebut sebagai *website* yang merupakan terdiri dari teks, *image*, *audio*, *video* atau lain-lain. Biasanya laman *web* hanya dapat diakses dengan menggunakan jaringan internet. Pada setiap laman *web* diperboleh untuk diakses oleh orang awam karena bersifat publik dan dapat diakses secara kolektif sehingga dapat membentuk *World Wide Web* (www). Pada *World Wide Web*, terdapat berbagai macam bentuk situs web yang berisi banyak dokumen. Dari banyaknya situs web ini, ada yang merupakan situs web spam atau situs web yang berisi konten-konten yang tidak berguna dengan menyebarkan informasi-informasi yang tidak benar hanya untuk mendapatkan jumlah kunjungan pada situs web tersebut. Ada yang baik, ada yang buruk dan ada yang dapat dipercaya dan ada yang tidak aman karena memberikan halaman situs web yang berisi virus *trojan* atau *malware*. demikian pula pada dunia digital, ada banyak laman situs *web* yang bagus sementara lainnya tidak berguna.

Situs-situs tersebut belum tentu memberikan sumber berita yang jelas, sehingga pengguna harus memilah berita dari beberapa situs yang terpercaya. Untuk itu, diperlukan suatu sistem pencarian berita dari situs berita kesehatan nasional yang sudah terpercaya dan populer di masyarakat. Sistem pencarian tersebut menghasilkan beberapa berita yang diurutkan berdasarkan jumlah pencarian terbanyak. Satu persatu isi berita yang disajikan

sehingga membutuhkan waktu yang banyak. dibutuhkan peringkasan berita untuk mempersingkat waktu membaca dan mempermudah pengguna memilah berita yang diinginkan tanpa harus membaca keseluruhan isi berita. Sistem yang dibutuhkan adalah Sistem Temu-Kembali Informasi (*Information retrieval* / IR).

Beberapa langkah awal dilakukan sebelum melakukan ekstraksi pada suatu laman *web* yang nantinya akan diekstrasi dan dijadikan cluster berdasarkan *topic modeling*, yaitu dengan menggunakan metode *pre-processing* untuk memudahkan ekstraksi fitur pada suatu laman web. Karena setiap dokumen harus melalui langkah *pre-processing* untuk membuat *search engine* mudah dalam melakukan pencarian dokumen secara otomatis, dan untuk menjadikannya sebagai dokumen untuk dilakukan proses berikutnya. Langkah-langkah seperti *tokenization*, *stemming*, normalisasi, *stopword removal* dll akan diterapkan pada penelitian ini untuk mempermudah ekstraksi setiap dokumen (Mishra and Vishwakarma, 2016).

Maka dari itu pengembangan dilakukan sebagai peningkat kualitas dari sisi layanan dan kualitas dari sisi informasi yang akan disajikan agar penyampaian tetap benar dan akurat ketika sampai ke penerima informasi.

Berdasarkan latar belakang masalah diatas, perlu dirancangnya sebuah sistem untuk meringkas serta menentukan topik dari suatu laman web serta dapat menghitung nilai *similarities* antar dokumen dengan menggunakan model *topic modeling* agar penerima informasi tetap mendapatkan informasi yang memiliki satu tujuan, serta untuk mengukur seberapa layak informasi yang diberikan pada laman web. Penulis mencoba menyelesaikan masalah tersebut dengan membuat skripsi dengan judul "**Analisis dan Representasi Fitur pada Laman Web Menggunakan Metode *Latent Semantic Analysis***".

1.2 Identifikasi Masalah

Berdasarkan latar belakang diatas maka dapat diidentifikasikan beberapa masalah yang didapatkan sebagai berikut:

1. Belum terbentuknya korpus dengan topik pembahasan yaitu penyakit yang sangat berdampak bagi masyarakat dalam bentuk Bahasa Indonesia.
2. Data yang digunakan masih dalam bentuk mentah berupa HTML yang mengharuskan peneliti untuk melakukan ekstraksi terhadap semua dokumen.

1.3 Rumusan Masalah

Beberapa permasalahan yang akan diteliti pada penilitian kali ini dapat dirumuskan sebagai berikut:

1. Bagaimana cara menghasilkan korpus berita mengenai kesehatan sebagai bahan referensi penelitian?
2. Bagaimana menerapkan metode *pre-processing* pada penelitian terhadap suatu laman web?
3. Bagaimana menerapkan ekstraksi fitur pada suatu laman *web* memperoleh data yang detail berdasarkan topik yang akan diteliti menggunakan metode *Latent Semantic Analysis* dengan parameter *TF-IDF*?
4. Bagaimana menerapkan *clustering* berdasarkan *topic modeling* untuk menentukan *similarity* antara *term* & *document*?

1.4 Batasan Masalah

Supaya penelitian yang dilakukan lebih terarah dan terfokus serta tidak meluas dari pembahasan, penelitian memiliki batasan masalah dengan lingkup sebagai berikut:

1. Sistem ini dibuat untuk menentukan topik dari sebuah dokumen, serta memperhitungkan nilai dokumen *similarity*.
2. Data yang akan diekstraksi dalam bentuk laman *web* html. Sumber data berdasarkan top-10 google *search* dengan pemilihan situs yang berkaitan dengan kesehatan dengan rentang tahun 2015-2019.
3. Metode yang digunakan pada penelitian ini adalah dengan menggunakan *Latent Semantic Indexing* dengan parameter *TF-IDF (Term Frequency-Inverse Document Frequency)*.
4. Bahasa pemrograman yang digunakan adalah Python.

1.5 Tujuan Penelitian

Beberapa tujuan yang dilakukan pada penelitian ini adalah:

1. Menghasilkan korpus dengan topik mengenai kesehatan yang diambil dari suatu Laman *Web* sehingga dapat digunakan untuk membangun pemodelan topik dengan *Latent Semantic Analysis* (LSA).
2. Merancang sistem dengan metode LSA sehingga dapat melakukan identifikasi berdasarkan topik pembahasan.

3. Diharapkan proses peringkasan dokumen dengan metode *Information Retrieval System* (IRS) dan menerapkan metode *Latent Semantic Analysis* (LSA) pada sistem representasi data ini dapat menghemat waktu.
4. Menghasilkan *clustering* model berdasarkan hasil dari *topic modeling* dan ekstraksi fitur menggunakan metode *Latent Semantic Analysis*.

1.6 Manfaat Penelitian

Adapun beberapa manfaat yang dapat diperoleh pada penelitian ini diantaranya:

1. Penelitian bisa digunakan sebagai bahan evaluasi pembanding dalam penelitian-penelitian sejenis namun dengan metode perhitungan yang berbeda.
2. Dapat dijadikan acuan untuk merepresentasikan dokumen yang akan digunakan untuk diperhitungkan nilai *similarity* antara kata dengan dokumen.
3. Korpus yang nantinya dihasilkan dapat digunakan dimasa yang akan datang.