BAB II
LANDASAN TEORI

Bab ini berisi pembahasan mengenai teori yang berkaitan dengan penelitian dan pembahasan mengenai penelitian terkait:

2.1. Tinjauan Pustaka

2.1.1. Korpus

Korpus merupakan kumpulan teks berupa kata maupun kalimat dalam ukuran besar dan terstruktur. Korpus dapat berisi *text* dalam satu bahasa (*corpus monolingual*) atau berbagai macam bahasa (*corpus multilingual*) dan dapat disimpan dalam bentuk file *text*. Korpus sangat bermanfaat bagi penelitian bahasa. Korpus bisa digunakan dalam Pengajaran atau Pembelajaran, Statistika, Dialektologi, hingga Linguistik Historis (Prof.Bahren, 2017). Ada empat karateristik korpus menjadi *analisis linguistic*:

a. *Sampling and Representativeness*

   Dalam membangun sebuah korpus dari berbagai bahasa, dapat ditarik dari sebuah sampel yang mewakili dari berbagai pengujian secara maksimal, yaitu menyediakan korpus seakurat mungkin dari kecenderungan yang beragam termasuk proporsi antara korpus dan informasi yang dicari. Jadi, tidak semata-mata berdasarkan pada teks sampel yang dipilih, akan tetapi mencari sampel dari berbagai sumber yang diambil dari sumber dokumen aslinya, sehingga akan memberikan gambaran yang cukup akurat dari seluruh informasi yang akan didapatkan.

b. *Finite Size*

   Selain sampling, istilah korpus juga cenderung menyiratkan suatu isi teks dengan ukuran yang terbatas, misalnya 1.000.000 kata. Teks dapat terus ditambahkan ke dalamnya, sehingga semakin besar karena lebih banyak sampel yang ditambahkan.

   Keuntungan utamanya: (1) teks menjadi tidak statis karena teks yang baru akan selalu ditambahkan dan (2) ruang lingkup akan lebih besar dan jauh lebih luas sehingga akan mencakup dari bahasa yang digunakan.

Kelemahan utamanya adalah bahwa, karena terus berubah dalam ukuran dan kurang ketatnya sampel. Jadi sebaiknya pada awal pembangunan korpus, rencana riset ditetapkan secara rinci bagaimana berbagai bahasa yang digunakan diambil sampelnya, berapa banyak sampel dan kata harus dikumpulkan sehingga jumlah keseluruhan yang sudah ditetapkan ini dapat digunakan.

c. *Machine-Readable Form*

Corpora yang dapat dibaca oleh mesin memiliki beberapa keunggulan dibandingkan dengan format tertulis atau lisan. Pertama dan paling penting keuntungan dari *corpora* yang dapat dibaca oleh mesin adalah bahwa dimungkinkan untuk mencari dan memanipulasi dengan cara-cara yang tidak dilakukan dengan format lain. Sebagai contoh, sebuah korpus dalam format buku, akan perlu dibaca dari depan sampai belakang untuk mengambil semua contoh kata, dengan korpus yang dapat dibaca oleh mesin, tugas ini dapat dicapai dalam beberapa menit dengan menggunakan perangkat lunak, atau sedikit lebih lambat dengan menggunakan fasilitas pencarian di pengolah kata. Keuntungan kedua *corpora* yang dapat dibaca oleh mesih adalah bahwa dapat dengan cepat dan mudah diperkaya dengan informasi tambahan.

d. *Standard Reference*

Meskipun tidak termasuk hal yang penting dari definisi suatu korpus, tetapi ada juga pemahaman bahwa korpus merupakan referensi *standard* untuk berbagai bahasa yang diwakilinya. Hal ini mengandaikan ketersediaan yang luas kepada peneliti lain. Keuntungan dari korpus yang tersedia secara luas adalah bahwa akan memberikan tolok ukur yang dapat digunakan sebagai pembanding dalam studi. Misalnya. Secara langsung dibandingkan dengan hasil yang dipublikasikan (selama metodologi sama) tanpa perlu perhitungan ulang. Korpus *standard* juga berarti penggunaan korpus yang sama digunakan. (McEnery dan Wilson, 2001)

2.1.2. Crawling

Crawling adalah proses untuk mengumpulkan informasi dari halaman web berdasarkan indeks. Tujuan dari crawling adalah mengumpulkan banyak informasi dari halaman web yang berguna dengan cepat dan efisien, berikut dengan struktur *link* yang terkoneksi dengan halaman web tersebut. (Sasongko, 2010).

Fitur *crawler* yang perlu disediakan:

a. Terdistribusi

Crawler harus memiliki kemampuan untuk dapat dijalankan dalam berbagai macam model *hardware* dan *software* yang berbeda.

b. Scalable

Arsitektur *crawler* dimungkinkan untuk dapat meningkatkan kemampuan dengan menambahkan *hardware* dan *bandwith*.

c. Kinerja dan efisiensi

Crawler dapat membuat penggunaan berbagai sumber daya sistem, termasuk *prosesor*, penyimpanan, dan *bandwidth* jaringan lebih efisien.

d. Kualitas

Crawler dapat melayani kebutuhan permintaan anggota dengan mengambil yang berguna dari halaman *web* yang diketemukan.

e. Freshness

Crawler harus dapat mendapatkan informasi dari halaman *web* yang belum dikunjungi dan tidak mengambil informasi yang sudah pernah diambil.

f. Extensible

Crawlers harus dirancang untuk dapat disesuaikan dengan perkembangan teknologi yang ada sekarang dan yang akan datang, baik format data, *protocol* yang digunakan, dan seterusnya. (Sasongko, 2010).

2.1.3. Tokenisasi

Tokenisasi merupakan proses pemisahan suatu rangkaian karakter berdasarkan karakter spasi, dan mungkin pada waktu yang bersamaan dilakukan juga proses penghapusan karakter tertentu, seperti tanda baca [1]. Proses ini cukup rumit untuk sebuat program komputer karena beberapa karakter dapat dijadikan sebagai

pembatas (*delimiter*) dari token-token itu sendiri. Pembatas dari token tersebut antara lain spasi, tab dan baris baru, sedangkan karakter ( ) < > ! ? “ . , terkadang dianggap sebagai pembatas dan juga bukan pembatas tergantung pada kondisi pemakainya. (Septiandri, 2012).

Tokenisasi pada Bahasa Indonesia merupakan proses pemisahan suatu rangkaian karakter berdasarkan karakter spasi, dan mungkin pada waktu yang bersamaan dilakukan juga proses penghapusan karakter tertentu, seperti tanda baca. Sebagai contoh, kalimat “Harga beras terus naik” menjadi “Harga”, “beras”, “terus”, “naik”.

2.1.4. Parsing

Parser adalah sebuah *sistem* dasar yang memungkinkan pemahaman yang lebih baik terhadap kalimat dalam bahasa tertentu. Proses yang dilakukan oleh *parser* disebut *parsing*. *Parsing* merupakan proses pengambilan kata-kata dari kumpulan dokumen. Tujuan utama dari *parsing* adalah memeriksa apakah urutan *token* yang dihasilkan sesuai dengan tata bahasa dari bahasa yang bersangkutan. (Gumita dan Manurung, 2008)

2.1.5. POS Tagging

Part-of-Speech (POS) *tagging*, yang juga disebut sebagai pelabelan kelas kata, adalah suatu proses yang memberikan label kelas kata secara otomatis pada suatu kalimat. Pelabelan kata dapat dilakukan berbasis aturan (*rule based*) dan probabilitas (*probability-based*) dari sebuah model yang dibangun. (Jurafsky, 2000)

Tabel 1: Indonesian tagset (Dinakaramani,2012)

Tag	Description	Example
CC	Coordinating conjunction	dan, tetapi, atau
CD	Cardinal number	dua, juta, enam, 7916, sepertiga, 0,025, 0,525, banyak, kedua, ribuan, 2007, 25
OD	Ordinal number	ketiga, ke-4, pertama
DT	Determiner / article	Para, Sang, Si
FW	Foreign word	climate change, terms and conditions
IN	Preposition	dalam, dengan, di, ke, oleh, pada, untuk
JJ	Adjective	bersih, panjang, hitam, lama, jauh, marah, suram, nasional, bulat
MD	Modal and auxiliary verb	boleh, harus, sudah, mesti, perlu
NEG	Negation	tidak, belum, jangan
NN	Noun	monyet, bawah, sekarang, rupiah
NNP	Proper noun	Boediono, Laut Jawa, Indonesia, India, Malaysia, Bank Mandiri, BBKP, Januari, Senin, Idul Fitri, Piala Dunia, Liga Primer, Lord of the Rings: The Return of the King
NND	Classifier, partitive, and measurement noun	orang, ton, helai, lembar
PR	Demonstrative pronoun	ini, itu, sini, situ
PRP	Personal pronoun	saya, kami, kita, kamu, kalian, dia, mereka
RB	Adverb	sangat, hanya, justru, niscaya, segera
RP	Particle	pun, -lah, -kah
SC	Subordinating conjunction	sejak, jika, seandainya, supaya, meski, seolah-olah, sebab, maka, tanpa, dengan, bahwa, yang, lebih ... daripada ..., semoga
SYM	Symbol	IDR, +, %, @
UH	Interjection	brengsek, oh, ooh, aduh, ayo, mari, hai
VB	Verb	merancang, mengatur, pergi, bekerja, tertidur
WH	Question	siapa, apa, mana, kenapa, kapan, di mana, bagaimana, berapa
X	Unknown	statemen
Z	Punctuation	"...", ?, .

2.1.6. WEB dengan Framework

Website merupakan komponen atau kumpulan komponen yang terdiri dari teks, gambar, suara animasi sehingga lebih merupakan media informasi yang menarik untuk dikunjungi. Website adalah halaman informasi yang disediakan melalui jalur internet sehingga bisa diakses di seluruh dunia selama terkoneksi dengan jaringan internet. (Arif, 2012).

Dalam proses pembangunan website perlu adanya *framework* agar website yang dibangun menjadi lebih terstruktur, ditandai dengan adanya MVC (Model, View, dan Controller). Pengertian dari *framework* sendiri adalah sekumpulan library yang diorganisasikan pada sebuah rancangan arsitektur untuk memberikan kecepatan, ketepatan, kemudahan dan konsistensi didalam pengembangan aplikasi dari definisi tersebut" (Arif, 2012).

Framework MVC terdiri dari:

a. Model

   Model mencakup semua proses yang terkait dengan pemanggilan struktur data baik berupa pemanggilan fungsi, *input processing* atau mencetak *output* ke dalam browser.

b. View

View mencakup semua proses yang terkait *layout output*. Bisa dibilang untuk menaruh *template interface* website atau aplikasi.

c. *Controller*

Controller mencakup semua proses yang terkait dengan pemanggilan database dan kapsulisasi proses-proses utama. Jadi semisal dibagian ini ada file bernama member.php, maka semua proses yang terkait dengan *member* akan dikapsulisasi/dikelompokan dalam file ini. Kelebihan dengan adanya *framework* akan lebih mempermudah memahami mekanisme kerja dari sebuah applikasi. Ini tentunya akan sangat membantu proses pengembangan sistem yang dilakukan secara team.

2.1.7. Basis Data

Basis data dilihat dari struktur katanya berasal dari kata data. Data merupakan sesuatu yang nyata, fakta mengenai objek yang dapat mengurangi derajat ketidakpastian tentang suatu keadaan atau kejadian. (Kristanto, 2004).Sedangkan basis data didefinisikan sebagai kumpulan data yang disatukan dalam suatu organisasi atau tempat tertentu.

a. DBMS (Database Management System)

DBMS (Database Management System) merupakan sekumpulan program yang mengatur ataupun mengkoordinasikan semua kegiatan yang berhubungan dengan basis data. Dengan adanya berbagai tingkatan pandangan dalam suatu basis data maka untuk mengakomodasikan masing-masing pengguna dalam piranti lunak manajemen database biasanya terdapat bahasa-bahasa tertentu yang disebut Data Sub Language. (Zain, 2014).

b. NoSQL (Not only SQL)

Pada tahun 1998 pertama kalinya dikembangkan sebuah konsep penyimpanan basis data yaitu NoSQL oleh Carlo Strozzi, yang kemudian pada tahun 2009 Eric Evans memperkenalkan kembali teknologi NoSQL. Kehadiran NoSQL bukan berarti untuk menggantikan model RDBMS yang sudah ada. Awal kemunculannya dilatarbelakangi oleh beberapa masalah yang muncul dari RDBMS. NoSQL dan RDBMS memiliki kelebihan dan

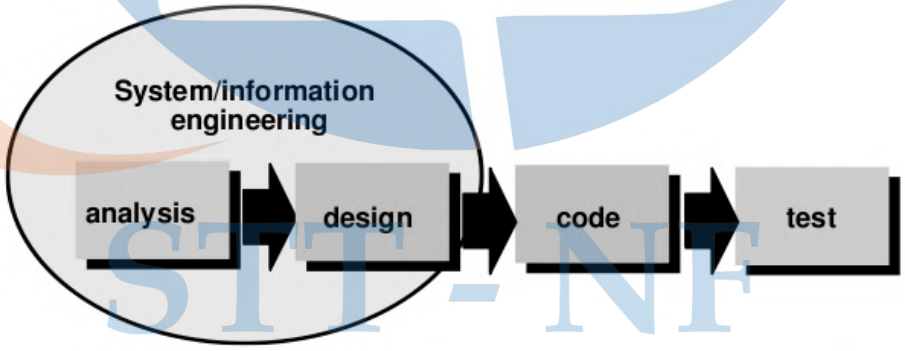
tempat masing-masing sehingga diharapkan dapat saling melengkapi teknologi penyimpanan basis data.

Secara bahasa NoSQL merupakan singkatan dari Not Only SQL yang berarti sistem manajemen basis data tersebut berbeda dengan basis data relasional dalam beberapa aspek. Dalam RDBMS dikenal adanya konsep ACID (Atomicity, Consistency, Isolation, Durability). Sedangkan dalam NoSQL menerapkan konsep BASE (Basically Available, Soft state, Eventually Consistent). (Zain, 2014).

2.1.8. Metode Pengembangan

a. Waterfall

Metode pengembangan sistem yang penulis gunakan adalah model SDLC (System Development Life Cycle) Waterfall. Model waterfall adalah sebuah model pengembangan aplikasi dengan pendekatan sekuensial. Pendekatan model ini terlihat mengalir menurun seperti air terjun (Waterfall) yang dikembangkan oleh Pressman melalui beberapa tahap. Penggunaan istilah waterfall pertama kali dikenal oleh Winston Royce pada tahun 1970.



gambar 1 : Model Pengembangan Sistem dengan Waterfal

Model ini bisa juga disebut dengan linier sequensial model, menggunakan pendekatan sistematis dan sekuensial dalam pengembangan aplikasi, dimulai melalui proses analisis, desain, pengkodean, uji coba dan pemeliharaan. (Dendi Ramdani, 2014).

b. Testing aplikasi menggunakan Black-Box Testing

Black-box testing merupakan pendekatan pengujian yang proses ujinya diturunkan dari spesifikasi program atau komponen. Sebuah sistem merupakan ‘kotak hitam’ yang perilakunya hanya dapat ditentukan dengan mempelajari input dan output yang berkaitan. Nama lain untuk pengujian ini ialah pengujian fungsional karena penguji hanya berkepentingan dengan fungsionalitas dan bukan implementasi perangkat lunak. (Adidaya, 2015).

2.2. Penelitian Terkait

dalam sebuah penelitian memerlukan banyak sumber bacaan untuk memperkuat penelitian yang dilakukan. Untuk itu penulis perlu mengkaji penelitian penelitian yang sudah ada sebelumnya. Berikut ini penulis berikan table penelitian terkait yang tercantum dalam tabel 2.

Tabel 2 : Penelitian Terkait

No.	Judul Penelitian	Tahun	Peneliti	Kesimpulan
1.	Aplikasi untuk Membangun Corpus dari Data Hasil Crawling dengan Berbagai Format Data Secara Otomatis	2010	Jati Sasongko	a. menghasilkan tiga puluh file baik file dokumen teks maupun image. b. Aplikasi mampu melakukan download dari sebuah alamat web dengan otomatis dengan ketentuan dapat dilakukan oleh user. c. Aplikasi mampu melakukan konversi dari dokumen teks dengan berbagai format data ke dalam bentuk dokumen teks txt, juga dalam melakukan konversi pada

				semua format file image ke dalam bentuk format bmp. d. Aplikasi mampu menyimpan dokumen teks dan image dalam tabel teks dan tabel image secara otomatis dari semua hasil pencarian dan konversi yang telah dilakukan pada proses sebelumnya.
2.	RANCANG BANGUN APLIKASI INFORMATION RETRIEVAL UNTUK MENGKOLEKSI DATA PARALEL KORPUS TEKS BAHASA INGGRIS – BAHASA INDONESIA	2012	Edy Septiandri	1. Sistem mampu mengumpulkan dokumen artikel berita melalui proses crawling website dari situs milik BBC dengan alamat URL http://www.bbc.co.uk/indonesia/topik/dwi bahasa/, sehingga menghasilkan sebuah dokumen yang berisi kumpulan artikel berbahasa Inggris sebagai bahasa sumber dan berbahasa Indonesia sebagai bahasa terjemahan. 2. Sistem dapat melakukan proses tokenisasi untuk menghilangkan semua tanda baca yang tidak diperlukan, dan proses parsing untuk menghilang semua dokumen yang tidak relevan dan memisahkan antara kalimat bahasa Inggris dan bahasa Indonesia ke dalam dua dokumen yang berbeda 3. Sistem temu balik ini telah menghasilkan paralel korpus bahasa

				Inggris dan bahasa Indonesia sebanyak 1541 kalimat, sehingga dapat menambah perbendaharaan paralel korpus bahasa Inggris dan bahasa Indonesia yang sudah ada, yaitu dari 27.326 kalimat menjadi 28.867 kalimat. 4. Pembuatan paralel korpus menggunakan aplikasi Information Retrieval jauh lebih cepat dibanding dengan pembuatan paralel korpus dengan cara manual.
3.	PENYUSUNAN KORPUS BERITA TERBUKA BERBAHASA INDONESIA	2015	Ahmad Rio Adriansyah	Penelitian ini menghasilkan sekumpulan dokumen yang bersifat terbuka untuk digunakan oleh komunitas yang meneliti pemrosesan bahasa natural atau bahasa Indonesia. Korpus yang diambil dengan metode di atas belum diberikan tag untuk PoS (Part-of-Speech). Untuk tagging, dapat digunakan tagger otomatis seperti yang disampaikan oleh [3] atau secara manual. Yang dicantumkan dalam jurnal ini adalah sebagian kecil data yang sudah berhasil diambil. Ke depannya, korpus ini akan dikembangkan ke website lain dan dengan rentang waktu yang lebih lebar. Data tersebut

				dapat diakses secara umum melalui email penulis.
4.	Part-of-Speech (POS) Tagging Bahasa Indonesia Menggunakan Algoritma Viterbi	2016	1. Nitin Sabloak 2. Bebeto Agung Hardono 3. Derry Alamsyah	4. Algoritma Viterbi dapat digunakan untuk melakukan Part-of-Speech (POS) tagging pada bahasa Indonesia. 5. Probabilitas inisialisasi berpengaruh terhadap label (tag). Nilai probabilitas inisialisasi yang kecil memiliki kemungkinan yang kecil untuk dipilih sebagai label (tag) pada kata. 6. Tingkat akurasi yang dihasilkan pada 10 kali pengujian yang dilakukan menghasilkan akurasi yang tidak jauh berbeda. Dengan rata – rata akurasi adalah 93,23018 % dan standar deviasi sebesar 0,260541273. 7. Penambahan kata ‘zz’ untuk mengelompokkan kata yang tidak terdapat pada korpus (kata asing) tidak berpengaruh terhadap hasil akurasi

Dari penelitian terkait di atas, penulis merepresentasika posisi penelitian yang dilakukan dalam bentuk tabel sebagaimana tabel 2. Berikut ini yang di maksud sebagai tabel 3:

Tabel 3 : Posisi Penelitian

No.	Judul Penelitian	Tagging	Korpus	web	Desktop
1.	Aplikasi untuk Membangun Corpus dari Data Hasil Crawling dengan Berbagai Format Data Secara Otomatis	-	√	-	√
2.	Rancang Bangun Aplikasi Information Retrieval Untuk Mengkoleksi Data Paralel Korpus Teks Bahasa Inggris – Bahasa Indonesia	-	√	√	-
3.	Part-of-Speech (POS) Tagging Bahasa Indonesia Menggunakan Algoritma Viterbi	√	-	-	-
4.	Penyusunan Korpus Berita Terbuka Berbahasa Indonesia	-	√	-	-

2.3 Tabel Daftar Istilah

Penulisan sebuah karya tulis dibidang teknologi terutama Teknik informatika seringkali ditemukan sebuah istilah tersendiri dibidang teknikal yang sulit dipahami oleh orang awam dan hanya dimengerti oleh beberapa orang yang menekuni bidang Teknik informatika. Dengan demikian para pembaca karya tulis akan susah memahami makna makna dari tulisan yang sebenarnya, untuk itu penulis membuat sebuah tabel yang berisikan daftar istilah istilah Teknik yang penulis gunakan dalam penulisan tugas akhir ini. Berikut tabel daftar istilah yang penulis maksudkan.

Tabel 4 : Daftar Istilah Terkait

No.	Nama istilah	Makna
1.	Korpus	Kumpulan kata
2.	Golden korpus	Kumpulan korpus
3.	Data mining	Serangkaian proses untuk menggali nilai tambah berupa informasi yang selama ini tidak diketahui secara manual dari suatu basisdata dengan melakukan penggalian pola-pola dari data dengan tujuan untuk memanipulasi data menjadi informasi yang lebih berharga yang diperoleh dengan cara mengekstraksi dan mengenali pola yang penting atau menarik dari data yang terdapat dalam basisdata.
4.	Analitic linguistic	Penelitian mengenai Bahasa
5.	Tagging	Proses pemberian label(tag)
7.	MYSQL	Database terstruktur
8.	NOSQL	Database tidak terstruktur
9.	Corpora	Bentuk jamak dari korpus
10.	Crawler	Alat yang digunakan untuk mengambil kata dari internet
11.	Parser	Sebuah sistem dasar yang memungkinkan pemahaman yang lebih baik terhadap kalimat dalam bahasa tertentu
12.	Parsing	proses pengambilan kata-kata dari kumpulan dokumen
13.	Tokenisasi	proses pemisahan suatu rangkaian karakter
14.	Token	istilah (term) atau kata
15.	Framework	kerangka kerja adalah sebuah software untuk memudahkan para programmer membuat aplikasi atau web yang isinya adalah berbagai fungsi,

		plugin, dan konsep sehingga membentuk suatu sistem tertentu
16.	Layout	Kerangka dalam Tampilan
17.	Interface	Tampilan