PEMBANGUNAN ALGORITMA LEMMATIZER BAHASA MELAYU

Sara Syakirah Hashim

Fakulti Teknologi dan Sains Maklumat,
Universiti Kebangsaan Malaysia,
43600 Bangi Selangor, Malaysia.

Lailatul Qadri Zakaria

Fakulti Teknologi dan Sains Maklumat,
Universiti Kebangsaan Malaysia,
43600 Bangi Selangor, Malaysia.

ABSTRAK

Lemmatization dan stemming merupakan salah satu teknik Pemprosesan Bahasa Tabii (NLP) dan digunakan secara meluas dalam perlombongan teks. Perlombongan teks adalah proses menganalisa teks yang ditaip dalam bahasa tabii dan mengekstarak informasi yang berkualti daripada teks tersebut. Algoritma stemmer tertua yang pernah dibina adalah untuk teks berbahasa Inggeris iaitu Porter Stemming Algorithm pada tahun 1979. Kini, lemmatizer dan stemmer bagi bahasa-bahasa lain seperti Bahasa Arab, Bahasa Perancis, dan Bahasa Itali juga telah dibangunkan. Bahasa Melayu juga mempunyai algoritma lemmatizer dan stemmer yang tersendiri. Algoritma lemmatizer pertama bagi Bahasa Melayu telah dibina pada tahun 1993 dan hingga kini banyak algoritma baru telah dibangunkan yang menggunakan pelbagai jenis pendekatan serta teknik. Namun, algortima yang telah dibangunkan masih belum mencapai tahap ketepatan yang setanding dengan aloritma Bahasa Inggeris dan masih memerlukan banyak pengajian dari segi morfologi Bahasa Melayu. Hal ini bagi memastikan algoritma lemmatizer yang dibangunkan mampu mencantas imbuhan dan mencari kata dasar dengan baik serta sesuai untuk diaplikasikan ke teks berbahasa Melayu. Permasalahan juga timbul kerana kekurangan akses kepada algoritma yang boleh dimanipulasi oleh pengkaji sekarang dan masa depan. Kebanyakan pembangun algoritma terdahulu tidak menyediakan penerangan terperinci tentang komponen-komponen yang adalah di dalam setiap gerak kerja ketika membangunkan algoritma tersebut dan pengaturcaraannya tidak diterbitkan secara umum di atas talian. Oleh itu,

Algoritma Lemmatizer Bahasa Melayu akan dibangunkankan semula berdasarkan peraturan penyelidik terdahulu. Oleh itu, diharapkan projek ini mampu membantu pengkaji NLP Bahasa Melayu yang akan datang terutamanya mahasiswa Universiti Kebangsaan Malaysia untuk meneruskan kajian tentang algoritma lemmatizer. Projek ini juga didokumentasi dengan terperinci serta kodnya boleh diolah oleh pengkaji lain secara terbuka dan percuma.

1 PENGENALAN

Teks merupakan sumber data mentah terbesar di dunia berbanding aset multimedia lain. Teks yang dikumpul sebagai data perlulah melalui proses analisis bagi mendapatkan informasi yang khusus mengikut kesesuaian organisasi yang menguruskannya. Bagi menjalankan analisis ini, data terlebih dahulu perlu diproses dengan teknik-teknik Pemprosesan Bahasa Tabii (NLP). Antara teknik yang terdapat dalam NLP adalah pensegmenan ayat, pentokenan, *part-of-speech-tagging*, *lemmatization* atau *stemming*, pembuangan kata henti, penghuraian tanggungan, rangkai kata bagi kata nama, pengecaman entiti bernama dan resolusi rujukan. *Lemmatization* dan *stemming* merupakan salah satu teknik yang penting dalam NLP serta selalu digunakan oleh para penganalisa teks bagi tujuan untuk mencantas kata imbuhan daripada perkataan bagi menghasilkan kata dasar.

Bidang morfologi amat berkait rapat dengan NLP dan merupakan nadi bagi pembangunan teknik-teknik NLP. Menurut Dewan Bahasa dan Pustaka (DBP, 2016) morfologi ialah kajian bentuk kata atau pembentukan kata. Menurut Salinah Ja'afar morfologi juga mengkaji tentang bentuk kata (Salinah, 1995). Bentuk kata ialah rupa unit tatabahasa, sama ada berbentuk tunggal atau hasil daripada proses pengimbuhan, pemajmukan dan penggandaan. NLP turut menggunakan pengetahuan morfologi bagi membangunkan algoritma pemprosesan seperti *lemmatization*.

Kata dasar merupakan aspek yang penting bagi proses *lemmatization*. Menurut DBP kata dasar adalah kata akar atau kata terbitan yang diberi imbuhan (DBP, 2016). Kata dasar Bahasa Melayu diperlukan bagi membentuk kata terbitan yang bermakna serta kata terbitan ini mempunyai fungsi mengikut hukum tatabahasanya tersendiri. Kini, DBP telah berjaya menyenaraikan lebih kurang 89,200 patah perkataan termasuk

kata terbitan kata dasar Bahasa Melayu di dalam Kamus Dewan Bahasa dan Pustaka edisi ke empat.

Menurut Sarkar proses *stemming* sangat menyerupai proses *lemmatization*, iaitu dengan mencantas kata imbuhan bagi mendapatkan bentuk dasar perkataan (Sarkar, . Tetapi bagi *lemmatization*, bentuk dasarnya dikenali sebagai kata dasar dan bukannya *stem* dasar. Perbezaannya adalah, *stem* dasar mungkin tidak selalunya tepat dari segi leksikografi, ini bermaksud *stem* dasar tersebut mungkin tidak wujud dalam kamus atau pun *stem* dasar itu bukanlah kata dasar sebenar bagi perkataan imbuhan yang diuji. Misalnya, perkataan 'mengira' akan menghasilkan kata dasar 'ira' melalui proses *stemming*, namun jika *lemmatization* dilakukan, perkataan dasar yang akan dihasilkan adalah 'kira'. Dalam contoh ini proses *stemming* telah menghasilkan *stem* dasar yang tidak tepat kerana menurut DBP definisi 'mengira' adalah menghitung berapa jumlahnya (jawabnya dan lain-lain) dan kata dasarnya adalah 'kira' dan bukannya 'ira' yang bermaksud pangsa durian (limau dan lain-lain).

2 PENYATAAN MASALAH

Pemprosesan Bahasa Tabii (NLP) adalah sangat penting dalam menganalisa teks mentah. Antara kelebihan analisis teks adalah pengekstrakan maklumat daripada teks yang dikumpul. Namun sebelum analisis teks dapat dilakukan, beberapa langkah pra pemprosesan teks perlu dijalankan terlebih dahulu. Proses *lemmatization* yang dilakukan pada teks dapat menjadikan keseluruhan teks lebih bersih tanpa kata imbuhan yang disertakan pada kata dasar. Kata dasar yang dijumpai juga adalah perkataan yang yang wujud di dalam kamus bahasa.

Terdapat beberapa algoritma *stemming* bagi Bahasa Melayu yang telah dibangunkan oleh penyelidik dahulu seperti *Stem* Malaya oleh Husein Zolkepli (Husein, 2020). Namun algoritma ini tidak sempurna, dan penghasilan *stem* dasar bagi sesetengah perkataan yang mempunyai konflik imbuhan adalah tidak tepat. Algoritma *Stemming* Sastrawi untuk Bahasa Indonesia yang telah dibangunkan oleh Hanif Amal Robbani (2016) pula, hampir sempurna buat Bahasa Indonesia, namun algoritma ini tidak boleh dijadikan sandaran utama untuk NLP Bahasa Melayu. Hal ini kerana, terdapat beberapa perbezaan dalam hukum pengibuhan bagi kedua-dua bahasa.

Algoritma yang telah dibangunkan oleh penyelidik terdahulu juga sukar untuk diperolehi, dan pengaturcaraannya kurang didokumentasi. Ini menyukarkan para penyelidik baharu untuk membangunkan semula serta memperbaiki algoritma pencantasan yang tersedia ada kerana kekurangan informasi. Oleh itu, pembangunan algoritma *Lemmatizer* yang baharu perlulah dilakukan supaya dapat diguna pakai oleh penyelidik yang akan datang bagi tujuan penambahbaikan algoritma.

3 OBJEKTIF

Objektif kajian ini, adalah :

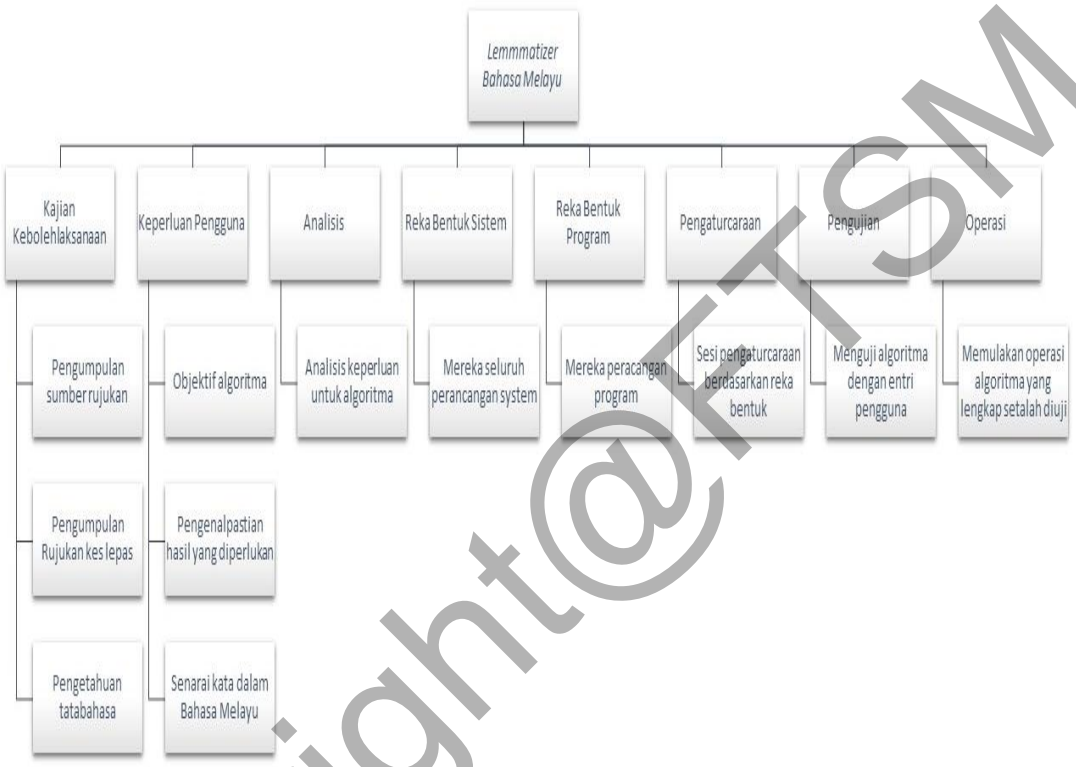
1. Membangunkan algoritma *Lemmatizer* bagi Pemprosesan Bahasa Tabii (NLP) Bahasa Melayu. Algoritma ini akan menggunakan peraturan imbuhan Bahasa Melayu untuk mengenalpasti imbuhan daripada perkataan atau ayat yang diinput oleh pengguna.
2. Menguji keberkesanan algoritma yang dibina dengan memastikan algoritma melaksanakan tugasan pencantasan imbuhan dengan tepat ketika fasa pengujian.

4 METOD KAJIAN

Model yang akan digunakan sepanjang tempoh pembangunan algoritma adalah model Air Terjun (Waterfall) oleh Royce. Model ini dikenali sebagai sebuah model klasik jika dibandingkan dengan model-model pembangunan sistem yang lain. Konsep model ini adalah fasa-fasa pembangunan akan bermula dari atas sehingga ke bawah seperti sebuah air terjun dan ianya bermula daripada fasa kebolehlaksanaan sehingga ke fasa operasi. Menurut Huges dan Cotterell (1999) proses pengulangan semula yang tehad merupakan salah satu kekuatan model ini. Model ini juga membolehkan masa penyiapan sesebuah projek diramal dengan lebih tepat berbanding model lain dan membolehkan projek ini dikawal secara berkesan. Oleh itu, bagi projek yang mempunyai matlamat yang jelas dan masa yang terhad, model ini adalah model yang paling sesuai.

4.1 Fasa Perancangan

Perancangan aktiviti projek adalah berdasarkan metodologi yang dipilih. Aktiviti dibahagikan kepada beberapa sub-aktiviti yang telah dipecahkan dalam Model *Work Breakdown Structure* berikut.



Rajah 1 Model *Work Breakdown Structure*

4.2 Fasa Analisis

Bertujuan menganalisa keselurahan projek bagi membangunkan model dan peraturan logik yang akan digunakan dalam algoritma. Analisa dilakukan pada hasil-hasil kajian yang terdahulu dan telah diringkaskan dalam jadual di bawah. Ini membantu membangunkan suatu algoritma berdasarkan ciri-ciri kajian dahulu.

Ciri-ciri	Pengakar Perkataan	PPM Versi Baharu	UniSZA *Stemmer*	CS *Stemmer*	*Lemmatizer* Bahasa Indonesia

	Melayu (PPM)				
Bahasa	Melayu	Melayu	Melayu	Indonesia	Indonesia
Jenis input diterima	Perkataan	Perkataan /ayat	Perkataan	Perkataan	Perkataan
Berasaskan peraturan	Ya	Ya	Ya	Ya	Ya
Bilangan peraturan	121	561	204	33	Tidak pasti
Pemeriksaan kamus	Ya, selepas pencantasan imbuhan	Ya, sebelum aplikasi setiap peraturan	Ya, selepas pemeriksaan kata ganda	Ya, selepas pencantasan imbuhan	Ya, selepas pencantasan imbuhan
Menerima istilah bahasa asing yang diadaptasikan ke Bahasa Melayu	Tidak	Ya	Tidak pasti	Tidak	Tidak
Menerima kata ganda nerimbuhan	Ya	Ya	Ya	Tidak	Tidak
Antara muka	Tiada	Tiada	Tiada	Tiada	Ada
Kajian perluasan	Baiki kamus	Baiki susunan peraturan	Hapus pengunaan kamus	Mengkaji skema untuk mencari kata nama	Baiki algoritma untuk menerima ayat

Jadual 1 Perbandingan algoritma yang dikaji

4.3 **Fasa Reka Bentuk**

Bahagian ini akan menerangkan tentang reka bentuk Algoritma Lemmatizer Bahasa Melayu dan cara pelaksanaannya. Tujuan reka bentuk dirancang supaya apabila proses pengaturcaraan dijalankan nanti masalah dapat diselesai dengan merujuk kembali struktur-struktur yang telah dibina. Ini juga mampu menjimatkan masa perancangan dan pelaksanaan keseluruhan projek ini.

4.3.1 Keperluan Fungsian

Bahagian ini menerangkan keperluan fungsi bagi *Lemmatizer* Bahasa Melayu.

1. Program yang dibina boleh menerima entri pengguna.
   Program menyediakan ruangan kosong untuk pengguna isi dengan perkataan atau ayat Bahasa Melayu.
2. Entri program boleh ditetapkan semula.
   Program akan mengosongkan ruangan entri sekiranya pengguna ingin mengubah ayat atau perkataan.
3. Program mampu mencari kata dasar daripada entri pengguna.
   Entri pengguna akan memproses entri pengguna dan mencari kata dasar bagi entri pengguna.
4. Program akan memaparkan kata dasar.
   Kata dasar yang berjaya dicari akan dipaparkan kepada pengguna.

4.3.2 Spesifikasi Fungsian

Bahagian ini menerangkan fungsi perisian bagi *Lemmatizer* Bahasa Melayu.

1. Fungsi tetapan semula entri
   Tetapan semula boleh dilakukan diruangan teks yang disediakan pada anatara muka program.
2. Fungsi menerima entri teks
   Program boleh menerima sebarang perkataan atau ayat berbahasa Melayu diruangan entri setelah menekan butang ‘semak kata dasar’ pada antara muka.
3. Fungsi memproses entri

Program akan melakukan rujukan dan pencantasan bagi mencari kata dasar bagi entri pengguna.

4. Fungsi papar kata dasar

   Program akan memaparkan kata dasar bagi entri pengguna.

4.3.3 Spesifikasi Bukan Fungsi

Spesifikasi bukan fungsi merupakan kualiti seluruh sistem dan kekangannya.

1. Kebolehgunaan

   Para pengguna yang bukan untuk tujuan penyelidikan algoritma boleh menggunakan program tanpa sebarang bimbingan.

   Pengguna yang bertujuan penyelidik juga boleh menggunakan program melalui mana-mana perisian pembangunan Python.

2. Kebolehpercayaan

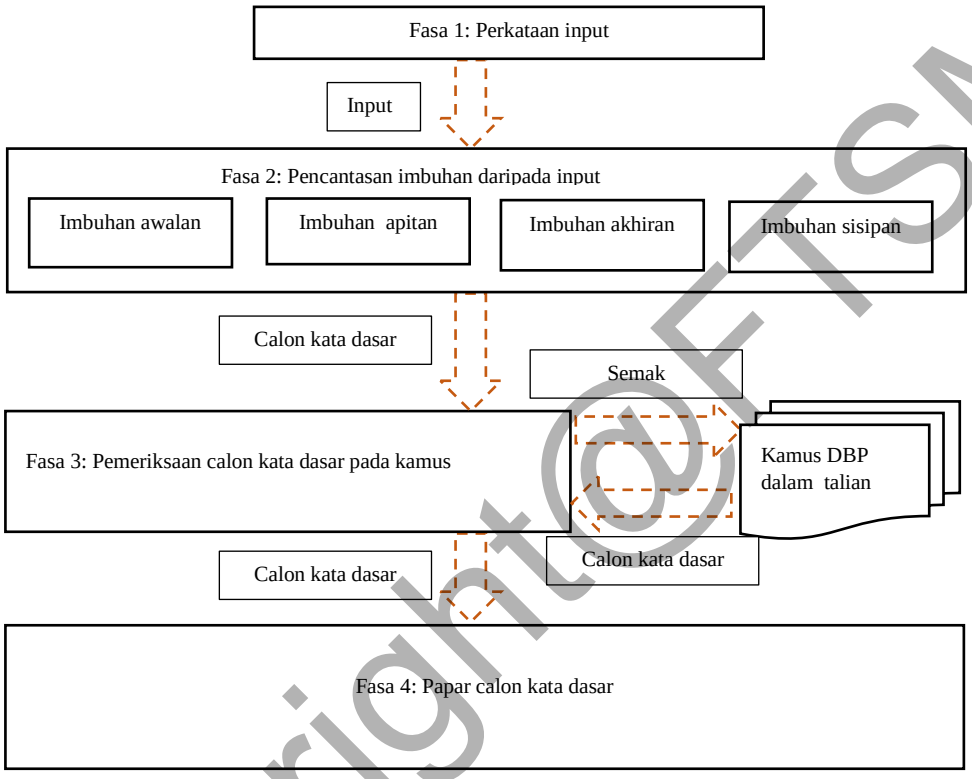
   Program boleh mengeluarkan kata dasar selagi perkataan atau ayat yang dicari adalah wujud dalam kamus.

3. Kebolehsediaan

   Program sedia digunakan pada bila-bila masa.

4.3.4 Rangka Kerja Pembangunan Algoritma *Lemmatizer* Bahasa Melayu

Bahagian ini akan menerangkan rangka kerja bagi Algoritma *Lemmatizer* Bahasa Melayu menggunakan rajah yang disertakan.

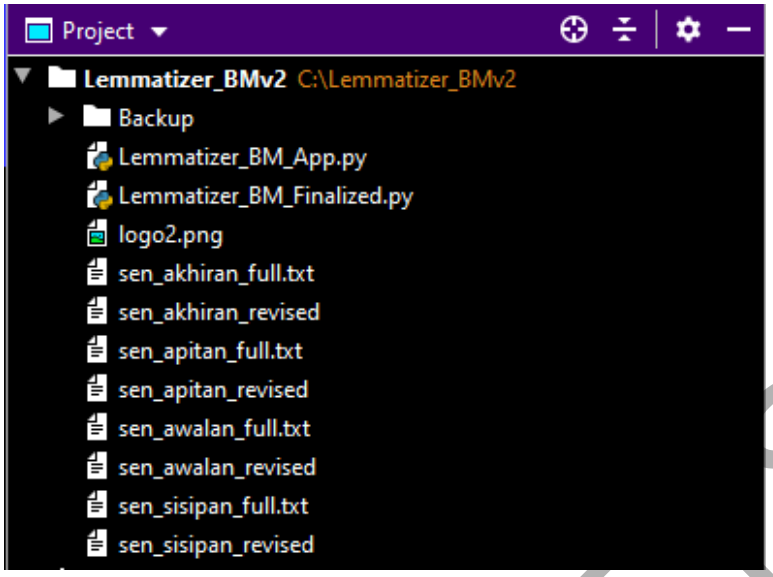


Rajah 2 Gerak kerja Algoritma *Lemmatizer* Bahasa Melayu

4.4 **Fasa Implementasi Reka Bentuk**

Pengaturcaraan bagi bahagian ini adalah baikpulih berdasarkan hasil calon kata dasar yang dilakukan pada algoritma di fasa latihan. Terdapat beberapa perubahan telah dilakukan pada algoritma dan dianggarkan dapat memberikan keputusan yang lebih baik daripada algoritma yang sebelumnya. Kali ini algoritma akan mengeluarkan jumlah kata dasar yang lebih kecil berbanding yang sebelumnya. Perbandingan hasil output antara set akan diterangkan di bahagian pengujian.

4.4.1 Struktur Fail



Rajah 3 Struktur fail pengaturcaraan terbaharu

4.4.2 Fail Pengaturcaraan

Perubahan telah dilakukan bagi menjadikan pengaturcaraan lebih bersifat bermandiri dan mengurangkan proses jumlah perkataan yang akan diperiksa pada kamus.

4.4.2.1 Fungsi Utama

Fungsi ini telah diubahsuai dan kini pencantasan imbuhan dilakukan berturut-turut mengikut susunan yang telah diaturkan. Berbeza dengan yang sebelumnya, pencantasan tetap akan dilakukan mengikut setiap jenis imbuhan. Kini pencatasan jenis imbuhan seterusnya hanya akan berlaku jika pencantasan jenis imbuhan sebelumnya tidak berjaya. Oleh kerana, pencastasan telah dikurangan maka, jumlah calon kata dasar pun berkurang, ini membantu mengecilkan skop pemilihan kata dasar tersebut. Penerangan langkah-langkah diambil fungsi ini diterangkan dibawah.

Langkah 1: Dapatkan input “perkataan”

Langkah 2: Panggil Fungsi “check_apitan(perkataan)”

a) Jika saiz senarai calon kata dasar yang diterima bersamaan kosong maka pergi ke Langkah 3

Langkah 3: Panggil “check_awalan(perkataan)”

a) Jika saiz senarai calon kata dasar yang diterima bersamaan kosong maka pergi ke Langkah 4
b) Jika saiz senarai bukan kosong maka keluarkan perkataan. Tamat program.

Langkah 4: Panggil "check_akhiran(perkataan)"

a) Jika saiz senarai calon kata dasar yang diterima bersamaan kosong maka pergi ke Langkah 5
b) Jika saiz senarai bukan kosong maka keluarkan perkataan. Tamat program.

Langkah 5: Panggil "check_sisipan(perkataan)"

a) Jika saiz senarai calon kata dasar yang diterima bersamaan kosong maka tiada pencantasan dilakukan. Periksa perkataan input di kamus. Tamat program.
b) Jika saiz senarai bukan kosong maka keluarkan perkataan. Tamat program.

4.4.2.2 Fungsi Periksa Imbuhan

Fungsi-fungsi periksaan imbuhan adalah sama dengan pengaturcaraan fasa latihan secara dasarnya, namun kini tugas memeriksa perkataan calon kata dasar telah diserahkan kepada fungsi-fungsi periksa imbuhan. Ini bermaksud, main() tidak lagi akan memeriksa senarai perkataan pada kamus, sebaliknya ia hanya akan meneruskan algoritma mengikut susun atus pencatasan imbuhan sekiranya pencantasan sebelumnya gagal menenemui sebarang kata dasar.

Langkah 1: Menerima perkataan yang di proses di 4.4.1.

Langkah 2: Mengambil data senarai imbuhan awalan daripada sen_awalan_revised.txt dan akan disimpan didalam objek Python List.

Langkah 3: Membina objek Python List bagi menyimpan perkataan yang telah dicantas imbuhan.

Langkah 4: Bagi setiap imbuhan dari senarai Langkah 2

a) Jika ianya sepadan dengan perkataan yang diterima maka imbuhan itu akan dibuang dari perkataan tersebut. Kemudian, simpan perkataan tersebut dalam senarai di Langkah 3 dan terus ke Langkah 5.

b) Jika tiada imbuhan yang sepadan maka senarai adalah kosong .

Langkah 5: Bagi setiap perkataan yang disimpan dalam senarai akan diperiksa dalam kamus dan bina sebuah Python List bagi calon kata dasar yang sah dalam kamus.

a) Jika perkataan wujud dalam kamus, simpan dalam senarai di atas.

b) Jika tiada perkataan yang wujud dalam kamus, maka senarai tersebut adalah kosong. main() akan memanggil periksa imbuhan yang seterusnya.

Tamat fungsi.

Contoh bagi pengaturcaraan adalah seperti berikut:

```
            calon_dasar, calon_dasarOri = tukar_huruf(imb, perkataan_noImb)
            calon_awalan_list.append(calon_dasar)
            calon_awalan_list.append(calon_dasarOri)
    if len(calon_awalan_list) > 0:
        ulasan_kamus, calon_dasarSah = carian_kamus_list(calon_dasarOri, calon_awalan_list)
        if len(calon_dasarSah) > 0:
            return calon_dasarSah
        else:
            check_akhiran(perkataan)
            return []
    else:
        check_akhiran(perkataan)
        return []
```

Rajah 4 Contoh pengaturcaraan

Dalam fasa percubaan, algoritma telah diberi satu susunan yang dianggap paling berkesan bagi menghadapi kes perkataan imbuhan apitan. Algoritma pada asalnya ketika difasa latihan telah membazirkan masa dan ruangan memori dengan menyimpan hasil cantasan bagi imbuhan awalan dan akhiran yang melalui dua proses yang berbeza (kes imbuhan apitan), kini algoritma boleh terus memcantas imbuhan apitan dan memeriksa kewujudan kata dasar dalam kamus. Periksa imbuhan sisipan diletakkan

diurutan terakhir kerana pada zahirnya perkataan dengan imbuhan sisipan adalah bukan sesuatu yang sering digunakan jika dibandingkan dengan tiga lagi jenis imbuhan. Oleh itu, susunan telah ditetapkan berdasarkan kekerapan penggunaan jenis penggunaan dalam teks berbahasa Melayu.

4.5 **Fasa Pengujian**

Ini merupakan bahagian penerangan untuk pengujian algoritma Data diambil daripada keratan akhbar dan merupakan senarai 50 perkataan berimbuhan. Berikut merupakan jadual bagi membandingkan hasil program dan Jawapan Cadangannya.

Senarai Perkataan Berimbuhan	Output *Lemmatizer* Bahasa Melayu Set A	Output *Lemmatizer* Bahasa Melayu Set B	Jawapan Cadangan
perwakilan	wakil	wakil	wakil
kehadiran	hadir	hadir	hadir
mengelakkan	gelak, kelak, lak, elak	lak, gelak, kelak, elak	elak
menyifatkan	sifat	sifat	sifat
memandangkan	pandang,andang, mandang	mandang, andang, pandang	pandang
merupakan	rupa	rupa	rupa
pemeriksaan	periksa	periksa	periksa
menerusi	terus	terus	terus
tindakan	tindak	tidak	tindak
pembersihan	sih, bersih	bersih	bersih
mencatatkan	catat	catat	catat
dilaporkan	lapor	lapor	lapor

melayari	layar	layar	layar
persoalan	soal	soal	soal
berkomunikasi	komunikasi	komukasi	komunikasi
mengisytiharkan	isytihar	isytihar	isytihar
dipermudahkan	mudah	mudah	mudah
kesesuaian	suai, sesuai	sesuai	sesuai
persediaan	sedia	sedia	sedia
pergigian	gigi	gigi	gigi
kemahiran	mahir	mahir	mahir
terjebak	jebak	jebak	jebak
berpunca	punca	punca	punca
mengawal	gawal, awal, kawal	kawal, gawal, awal	kawal
berentap	rentap	rentap	rentap
berkomunikasi	komunikasi	komunikasi	komunikasi
tersebut	sebut	sebut	sebut
perosak	rosak	rosak	rosak
semalam	malam, alam	malam	malam
sebuah	buah	buah	buah
dijual	jual	jual	jual
berkawan	kawan	kawan	kawan
petanda	tanda	tanda	tanda
bertanya	tanya	tanya	tanya
dipendam	pendam	pendam	pendam

menanggung	tanggung, nanggung, anggung	taggung, nanggung, anggung	tanggung
bernafas	nafas	nafas	nafas
susulan	susul	susul	susul
rakaman	raka, rakam	rakam	rakam
tindakan	tindak	tidak	tindak
utusan	utus	utus	utus
rawatan	rawat	rawat	rawat
sedangkan	dang	sedang	sedang
sokongan	sokong	sokong	sokong
saringan	saring	saring	saring
sumbangan	sumbang	sumbang	sumbang
hiasan	hias, hia	hias	hias
rundingan	runding	runding	runding
tinjauan	tinjau	tinjau	tinjau
tayangan	tayang	tayang	tayang
serbuan	serbu	serbu	serbu
kemuning	kemung, kuning	kemung, kuning	kuning
gementar	gentar	gentar	gentar

Jadual 2 Perbandingan Hasil Lemmatizer Bahasa Melayu Fasa Perubaan dan *Jawapan Cadangan*

Hasil output menunjukkan bahawa algoritma telah mengurangkan jumlah calon kata dasar apabila mencantas imbuhan apitan. Cuma pada sesetengah imbuhan yang mempunyai konflik padanan imbuhan awalan lebih daripada satu seperti “memandangkan” akan padan dengan imbuhan “mem..kan”, dan “me..kan ”. Ini akan

menghasilkan perkataan dasar “pandang”, “andang”, dan “mandang” dan kesemua kata ini wujud dalam pangkalan data Dewan Bahasa dan Pustaka. Algoritma turut mencalonkan perkataan yang masih berbaki satu imbuhan awalan sekirannya mempunyai konflik imbuhan awalan yang lebih daripada satu. Contoh boleh dilihat pada “diperhatikan” yang akan padan dengan imbuhan “di..kan” dan “diper..kan” seterusnya menghasilkan perkataan tanpa imbuhan padanan seperti “perhati” dan “hati” dan kedua-dua perkataan ini juga wujud dipangkalan data DBP.

5 HASIL KAJIAN

Bahagian ini akan mengulas beberapa isu yang telah dikenalpasti daripada output program. Bagi mendapatkan keralatan algoritma, dua jenis ujian tambahan dilakukan. Satu ujian adalah dengan menginput beberapa perkataan yang mungkin akan menghasilkan output yang bermasalah atau tidak tepat dan satu lagi ujian adalah input ayat daripada petikan rawak yang telah dikumpul daripada teks keratan akhbar dalam talian di awal projek. Calon kata dasar daripada program yang akan dibincangkan adalah berikut:

5.1 Ujian Menggunakan Perkataan Rawak

Perkataan	Output *Lemmatizer* Bahasa Melayu Set A	Output *Lemmatizer* Bahasa Melayu Set B	Cadangan Jawapan
rakan	rak	rak	rakan
perasaan	asa, rasa	rasa, asa	rasa
taman	tam	tam	taman
dari	dar	dar	dari
badan	bad	bad	badan
kaki	kak	kak	kaki
makan	mak	mak	makan
tekan	tek	tek	tekan
mari	mar	mar	mari

tari	tar	tar	tari
tuli	tul	tul	tuli
kari	kar	kar	kari
bukan	buk	buk	bukan
peran	ran	ran	peran
pelik	lik	lik	pelik
bersih	sih	sih	bersih
tarpemenang	tang, ang	ang, tang	menang
perang	rang,ang	ang, rang	perang
merah	mer	mah	merah
terang	rang,ang	ang, rang	terang
pening	ning,ting	ning, ting	pening

Jadual 3 Hasil Output Bagi Analisa Kesalahan Pencatasan Bagi Perkataan

5.2 Ujian Menggunakan Ayat Petikan Akhbar

> Kedua aktiviti ini diharap dapat mengeratkan lagi hubungan antara penduduk tempatan dan anggota agensi penyelamatan bagi melicinkan lagi operasi penyelamatan pada musim tengkujuh nanti katanya kepada pemberita di Dataran Kampung Dusun, Kuala Berang di sini hari ini.

Sumber: Keratan akhbar dalam talian, *400 penduduk Hulu Terengganu berakit sempena Hari Malaysia*, Kosmo Online.

Hasil calon kata dasar bagi ayat di atas adalah seperti berikut:

Perkataan	Output *Lemmatizer* Bahasa Melayu Set A	Output *Lemmatizer* Bahasa Melayu Set B	Jawapan Cadangan
Kedua	dua	dua	dua
aktiviti	aktiviti	aktiviti	aktiviti
ini	ini	ini	ini
diharap	harap	harap	harap

dapat	dap	dapat	dapat
mengeratkan	gerat, erat, rat, kerat	gerat, erat, rat, kerat	erat
lagi	lagi	lagi	lagi
hubungan	hubung	hubung	hubung
antara	antar	antara	antara
penduduk	duduk	duduk	duduk
tempatan	tempat	tempat	tempat
dan	dan	dan	dan
anggota	anggota	anggota	anggota
agensi	agen	agensi	agensi
penyelamatan	selamat	selamat	selamat
bagi	bagi	bagi	bagi
melicinkan	licin	licin	licin
lagi	lagi	lagi	lagi
operasi	oper, opera	operasi	operasi
penyelamatan	selamat	selamat	selamat
pada	pad	pada	pada
musim	musim	musim	musim
tengkujuh	tengkujuh	tengkujuh	tengkujuh
nanti	nanti	nanti	nanti
katanya	tanya	katanya	katanya
kepada	pada	pada	pada
pemberita	berita	berita	berita
di	di	di	di
Dataran	datar	datar	Dataran
Kampung	kampung	kampung	Kampung
Dusun,	dusun	dusun	Dusun
Kuala	ala	kuala	Kuala
Berang	rang, ang	rang, ang	Berang
di	di	di	di
sini	sin	sin	sini
hari	hari	hari	hari

ini	ini	ini	ini

Jadual 4 Hasil Output Bagi Analisa Kesalahan Pencatasan Bagi Ayat

5.3 Analisis Kesalahan Pencatasan

Setelah melakukan beberapa ujian pada program, telah dikenalpasti bahawa algoritma mengalami isu pencantasan imbuhan pada beberapa perkataan yang mempunyai konflik imbuhan dan pada perkataan yang tidak sepatutnya dicantas kerana ia sudah pun di dalam bnetuk yang paling asas.

5.3.1 Terlebih Cantas (*Overstemming*)

Isu ini merupakan isu terbesar yang telah dikenalpasti. Hal ini di sebabkan oleh dua faktor iaitu kamus dalam talian (DBP Online) dan kekurangan kecekapan algoritma untuk menentukan sejauh mana pencatasan perlu dilakukan. Jika dilihat daripada contoh perkataan yang berakhir dengan huruf i seperti ‘mari’, ‘kaki’, ‘tari’, akan menghasilkan calon kata dasar ‘mar’, ‘kak’, dan ‘tar’. Ini merupakan kesalahan terlebih cantasan imbuhan keran perkataan-perkataan ini sudah pun kata dasar dan tidak perlu dicantas mana-mana imbuhan.

Seperti contoh di atas, pencantasan yang dilakukan adalah mencancantas imbuhan akhiran i. hal ini kerana algoritma tidak tahu bahawa huruf akhir i itu adalah sebahagian daripada kata dasar atau imbuhan akhiran sebelum mencantasnya. Isu ini juga hanya terjadi sekiranya apabila imbuhan akhiran i dicantas, calon kata dasar yang dihasilkan wujud dalam kamus. Dalam kes seperti ‘mar’, ‘kak’, dan ‘tar’, perkataan-perkataan ini wujud di dalam Kamus Dewan Bahasa dan Pustaka dalam talian. Oleh itu pada proses pemeriksaan pada kamus, program menganggap ini merupakan calon kata dasar dan akan mengoutputkan hasil ini.

Algoritma juga mencantas kata nama khas yang sepatutnya tidak perlu dicantas seperti ‘Berang’. Hal ini kerana algoritma tiada pengetahuan tentang kata nama dan tidak mempunyai sebarang pengetahuan morfologi Bahasa Melayu yang memastikan jenis perkataan yang diinput.

5.3.2 Terkurang Cantas (*Understemming*)

Isu terkurang cantas boleh dilihat apabila berlaku konflik imbuhan. Contoh seperti perkataan ‘mengeratkan’ kan menghasilkan calon kata dasar iaitu ‘gerat’, ‘erat’, ‘rat’, ‘kerat’. Terkurang cantas algoritma telah menyebabkan penghasilan ‘gerat’ dan ‘kerat’. Tetapi, algoritma juga akan meneruskan pencantasan sehinggan jumlah padanan imbuhan yang maksimum. Maka terhasilah pertambahan calon kata dasar yang lain iaitu ‘erat’dan‘rat’.

6 KESIMPULAN

Bahagian ini, Pembangunan Algoritma *Lemmatizer* Bahasa Melayu akan diringkaskan dari segi keseluruhan pengalaman pembangunan algoritma, kelebihan algoritma yang dibangun, kekurangannya dan cadangan untuk penambahbaikan.

Selama setahun pembangunannya, algoritma ini telah banyak melalui proses penambahbaikan bahkan sebelum fasa latihan bermula penambahbaikan telah mula dilakukan pada perancangan pengaturcaraannya. Pelbagai teknik telah diterokai bagi memenuhi keperluan supaya program boleh dijalankan dengan bahasa pengaturcaraan Python. Python dipilih kerana sifatnya yang mudah difahami dan dipelajari. Kod-kod yang ditulis juga adalah sangat mudah untuk diolah dan telus.

Ditambah dengan aplikasi mudah yang dibina agar program lebih interaktif bersama pengguna menunjukkan bahawa projek Pembangunan Algoritma *Lemmatizer* Bahasa Melayu telah berjaya mencapai objektif-objektif yang ditetapkan. Walaupun projek ini masih memerlukan penambahbaikan, namun projek ini adalah permulaan yang baik bagi pembinaan algoritma *lemmatizer* yang lebih baik untuk Pemprosesan Bahasa Tabii Bahasa Melayu.

6.1 RINGKASAN PROJEK

Algoritma ini dibangunkan bagi tujuan membantu para pengkaji Pemprosesan Bahasa Tabii (NLP) Bahasa Melayu untuk membina suatu pengakar kata ataupun *lemmatizer* yang mampu memberi cadangan calon kata dasar yang baik bagi teks berbahasa

Melayu. *Lemmatizer* amat penting bagi mengekstrak informasi dalam teks atau artikel yang berbahasa Melayu, kerana kebanyakan perkataan yang ditaip adalah menggunakan imbuhan. Pada peringkat permulaan ini, algoritma berjaya mencapai objektifnya iaitu dengan mencantas imbuhan dan memastikan perkataan yang terhasil daripada pencantasan adalah perkataan yang wujud dalam Kamus Dewan Bahasa dan Pustaka. Oleh itu boleh disimpulkan bahawa projek ini merupakan suatu kejayaan, namun penambahbaikan oleh pengkaji baharu yang akan meneruskan projek inilah yang akan menjadikan algoritma ini setanding dengan algoritma *lemmatizer* berbahasa Imggeris .

6.2 KELEBIHAN ALGORITMA

Terdapat beberapa kelebihan yang dapat disenaraikan daripada penggunaan algoritma *lemmatizer* yang telah dibangunkan, antaranya adalah:

a) Pencantasan imbuhan

   Perkataan atau ayat yang mempunyai perkataan imbuhan boleh diringkaskan dengan membuang imbuhan yang ada pada perkataan tersebut. Ini membantu, penganalisa teks memberi penekanan utama pada ini pati teks berbanding terpaksa membaca perkataan yang panjang.

b) Mudah diolah

   Algoritma yang dibangunkan menggunakan bahasa pengaturcaraan Python yang amat mudah untuk dipelajari dan difahami. Pengaturcaraan juga telah dilengkapi dengan petunjuk-petunjuk yang dikomenkan dalam baris kod yang ditulis.

c) Interaktif

   Bagi memudahkan pengguna atau pengkaji memahami tujuan algoritma *lemmatizer*, suatu antara muka telah dibangunkan dan output bagi setiap kali algoritma dilancarkan juga dalam bentuk ringkasan proses yang telah dilakukan oleh program.

6.3 KEKURANGAN ALGORITMA

Oleh kerana algoritma *lemmatizer* ini masih lagi dalam fasa permulaan, terdapat beberapa kekurangan yang boleh dilihat dari segi penyenaraian calon kata dasarnya dan isu masa pengkomputeran.

a) Senarai perkataan dalam Dewan Bahasa Pustaka versi Dalam Talian.
   Salah satu kekurangan *lemmatizer* ini adalah ia menggunakan perkhidatan Kamus Dalam Talian (DBP Online) yang bergantung kepada seberapa kemas kininya pangkalan data dalam talian DBP. Berlaku juga isu di mana pencantasan dilakukan pada perkataan dan menghasilkan perkataan yang bukan kata dasar perkataan tersebut tetapi masih dianggap sebagai calon kata dasar. Contohnya, perkataan ‘mengelakkan’ yang menghasilkan ‘lak, gelak, kelak, elak’ sedangkan kata dasarnya hanyalah ‘elak’. Hal ini terjadi terdapat lebih daripada satu imbuhan apitan yang sepadan dengan perkataan input dan apabila dicari di kamus DBP pula, kesemua perkataan tersebut wujud. Oleh itu, algoritma hanya akan menganggap kesemua calon tersebut adalah output bagi input tadi.

b) Isu pencantasan terlebih cantas dan terkurang cantas
   Isu ini terjadi apabila algoritma terlebih cantas imbuhan atau terkurang cantas imbuhan pada perkataan dan menghasilkan kcalon kata dasar yang kurang tepat bagi perkataan yang diinput. Isu ini telah dibincangkan di bab empat bahagian analisis output.

c) Sambungan Internet
   Isu juga dilihat dari segi sambungan Internet kerana program memerlukan sambungan ke DBP *Online* yang stabil untuk mencari calon kata dasar. Sekirannya sambungan Internet tiada maka program tidak boleh mengesahkan kewujudan calon kata dasar di dalam Kamus Bahasa Melayu.

d) Masa pengkomputeran
   Masa juga merupakan isu dalam program ini. Program dilihat memerlukan masa yang akan lam berbanding *lemmatizer* lain. Hal ini mungkin disebabkan oleh

program perlu mendapatkan respons daripada DBP dalam talian sebelum boleh mengeluarkan output kepada pengguna.

6.4 CADANGAN PEMBAIKAN

Daripada senarai kelemahan yang telah dinyatakan di atas terdapat beberapa cadangan penambahbaikan yang ingin dicadangkan kepda pengkaji akan datang jika ingin meneruskan projek ini.

a) Menambah peraturan imbuhan

   Ianya amat penting untuk projek ini terus berkembang menjadi sebuah algoritma yang sempurna bagi *lemmatizer* Bahasa Melayu. Oleh itu lebih banyak peraturan perlu ditambah dalam algoritma bagi menampung jenis perkataan yang mengandungi imbuhan kata nama, klinikal, islamik dan perkataan imbuhan berganda.

b) Membina kamus bahasa

   Program seharusnya boleh bergerak dengan lebih pantas jika tidak perlu merujuk pangkalan data secara dalam talian. Oleh itu, pembangunan suatu kamus Bahasa Melayu amatlah dicadangkan bagi meneruskan projek ini.

c) Penggunaan kaedah pembelajaran mesin

   Pengkaji akan datang dicadangkan juga untuk meneroka kaedah baru untuk memperbaiki prestasi dan kecekapan algoritma dengan menggunakan pendekatan pembelajaran mesin.

d) Penggunaan *Part of Speech Tagging* (POS).

   Teknik POS merupakan suatu teknik di mana perkataan di dalam ayat akan ditanda dengan *part of speech*nya. Selalunya perkataan boleh dikenalpasti sama ada ia adalah kata nama, kata adjektif atau lain-lain dengan menggunakan teknik ini. Oleh itu, sekirannya teknik diaplikasi dalam algoritma *lemmatizer*, algoritma akan berupaya untuk memilih perkataan mana yang perlu dicantas, akan membantu algoritma menentukan jenis imbuhan yang sesuai bagi

perkataan tersebut. Ini juga mampu manyelesaikan isu terlebih dan terkurang cantasan.

7 RUJUKAN

Ahmad, F., Yusoff, M. & Sembok, T. M. T. 1996. Experiments with a *stemming* algorithm for Malay words. *Journal of the American Society for Information Science* 47(12): 909–918.

Asian, J., Williams, H. E. & Tahaghoghi, S. M. M. 2005. *Stemming Indonesia*. *Conferences in Research and Practice in Information Technology Series* 38(September 2018): 307–314.

Berita Harian. New Straits Times Press (M) Bhd. BH Online. https://www.bharian.com.my/

Bond, F. n.d.. Wordnet Bahasa. http://wn-msa.sourceforge.net/index.eng.html [23 December 2019].

Choudhary, S. n.d.. Reading selected webpage content using Python Web Scraping. https://www.geeksforgeeks.org/reading-selected-webpage-content-using-python-web-scraping/ [15 January 2020].

Fadzli, S., Norsalehen, A. & Syarilla, I. 2012. Simple Rules Malay *Stemmer*. *The International Conferenceon Informatics and Applications*(January 2012): 28–35.

Hanif Amal Robbani. Sastrawi 1.0.1.. https://pypi.org/project/Sastrawi/ [18 Januari 2016].

Harian Metro. New Straits Times Press (M) Bhd. myMetro. https://www.hmetro.com.my/

Huges, B., Cotterell, M., 1999. *Software Project Management*, Ed. Ke-2. Berkshire: . Hatter. D.

Husein Zolkepli. Malaya Stem GitHub. https://github.com/huseinzol05/Malaya/blob/master/malaya/stem.py . [2020]

Kamus Dewan Bahasa dan Pustaka. 2005. Ed. Ke-4. Kuala Lumpur: Dewan Bahasa dan Pustaka.

Kamus Pelajar. 2016. Ed. Ke-2. Kuala Lumpur: Dewan Bahasa dan Pustaka.

Othman. A. 1993. *Pengakar Perkataan Melayu untuk* Sistem *Capaian Dokumen*. Unpublished master' s thesis, Universiti Kebangsaan Malaysia, Bangi, Malaysia.

Refsnes Data. n.d.. Python Strings.https://www.w3schools.com/python/python_strings.asp [17 November 2019].

Salinah Ja'afar. 1995. Proses Pembentukan Kata dalam Leksikografi Bahasa Melayu: Tumpuan kepada Kamus Dewan. Kuala Lumpur. Disertasi S. Sa, Jabatan Pengajian Melayu, Universiti Malaya.

Sarkar, D. 2016. *Text Analytics with Python: A Practical and Real-World Approach and to and Gaining Actionable and Insights from and Your Data*. Banglore, Karnataka: Apress Media, LLC.

Suhartono, D., Christiandy, D. & Rolando, R. 2014. *Lemmatization Technique in Bahasa: Indonesian Language. Journal of Software* 9(5):1202-1209.

Zahid Bakar. 2012. Senarai Imbuhan. http://bahasaku-zahid.blogspot.com/2012/02/senarai-imbuhan.html [3 October 2019].