PTA-FTSM-2018-101

PEMBANGUNAN TEKNIK DALAM MERINGKASKAN TEKS DALAM BAHASA MELAYU.

Mohamed Razin bin Mohd Firoz

Prof Dr. Shahrul Azman bin Mohd Noah

Fakulti Teknologi & Sains Maklumat, Universiti Kebangsaan Malaysia

ABSTRAK

Perkataan "snippet" ialah kepingan kecil atau guntingan ayat atau ringkasan teks. Ringkasan teks memudahkan proses memahami teks yang ditulis. Pemahaman teks boleh dikenalpasti melalui pengecaman kata kunci atau perkataan yang paling banyak digunakan di dalam sebuah teks. Ringkasan teks Bahasa Melayu yang dibangunkan merupakan peringkasan jenis ekstraktif yang akan membantu meringkaskan sesebuah petikan atau teks. Penentuan kata kunci bagi dokumen sangat penting untuk menghasilkan ringkasan. Kata kunci mungkin berada di bahagian awalan, pertengahan dan akhiran perenggan. Dengan adanya ringkasan teks ini, teks atau petikan yang panjang dapat diringkaskan dengan pantas dan tepat. Algoritma yang digunakan untuk membangunkan sistem ini ialah algoritma frekuensi perkataan. Projek ini berhasrat untuk membangunkan sistem yang boleh meringkaskan sesebuah teks atau petikan yang panjang, menjadi sebuah teks yang pendek dan ringkas untuk memudahkan pengguna untuk membaca dan memahami.

1 PENGENALAN

Ringkasan teks memudahkan proses teks yang hendak ditulis. Pemahaman teks boleh dikenalpasti melalui pengecaman kata kunci atau perkataan yang paling banyak digunakan di dalam sesebuah teks atau petikan tidak termasuk perkataan fungsi khusus sekali kata hubung. Ringkasan teks Bahasa Melayu ini digunakan dengan peringkasan jenis ekstraktif dimana ia akan membantu meringkaskan sesebuah teks atau petikan Bahasa Melayu.

2 PENYATAAN MASALAH

Memandangkan tujuan peringkasan artikel dan teks adalah untuk menghasilkan satu dokumen yang ringkas dan berinformasi, maka penentuan kata kunci yang utama amat penting. Kata kunci bagi sesebuah teks mungkin berada di awalan, pertengahan atau akhiran perenggan.

Selain itu, pengenalpastian fungsi ayat bukanlah satu tugas yang mudah. Sebaris ayat mungkin mempunyai fungsi seperti ayat penyata dan lain-lain. Dengan permasalahan yang dinyatakan, sedikit sebanyak dapat membantu untuk membangunkan sistem peringkasan teks yang mudah.

3 OBJEKTIF KAJIAN

Projek ini dijalankan adalah untuk membangunkan sistem peringkasan teks dalam Bahasa Melayu. Objektif kajian ini adalah untuk mengenalpasti kata kunci dengan cara mengesan perkataan yang mempunyai frekuensi yang tinggi dalam teks atau artikel. Selain itu, ia juga bertujuan untuk membangunkan sistem yang boleh membantu pengguna untuk meringkas ayat dalam sesebuah teks atau artikel dalam Bahasa Melayu.

4 METOD KAJIAN

Penggunaan model pembangunan yang sesuai amat penting bagi memastikan perjalanan projek berjalan dengan lancar dan menjamin kualiti yang baik dan memuaskan. Kaedah yang digunakan untuk pembangunan model projek ini boleh dibahagikan kepada beberapa fasa. Fasa-fasa yang terlibat ialah fasa perancangan dan analisis, fasa reka bentuk, fasa implementasi dan fasa pengujian. Model ini penting bagi memastikan perjalanan projek lancar dan teratur. Rajah 1 menunjukkan model pembangunan yang digunakan untuk membangunkan sistem peringkasan teks dalam Bahasa Melayu.

4.1 Fasa Perancangan dan Analisis

Dalam fasa ini, penyataan masalah, skop kajian, objektif kajian serta jadual perancangan dikenalpasti. Pencarian bahan bacaan untuk dijadikan sebagai rujukan juga dilakukan dalam fasa ini. Hal ini kerana, ia dapat memberi kefahaman yang lebih terperinci mengenai topik kajian dan dapat mengenalpasti cara atau kaedah untuk peringkasan teks.

4.2 Fasa Reka Bentuk

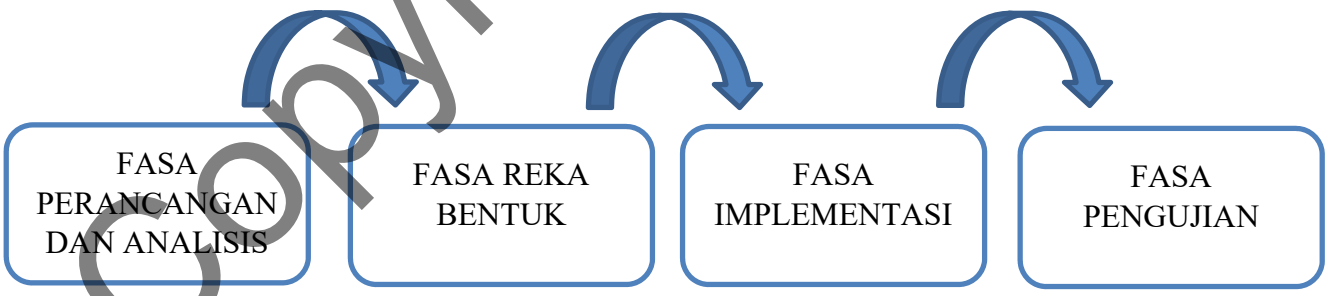
Dalam fasa ini pula, proses reka bentuk algorithm dilaksanakan. Pemilihan bahasa pengaturcaraan turut membantu bagi menjalankan kajian ini. Penghasilan pengaturcaraan menggunakan bahasa yang dipilih. Jika terdapat kecelaruan teknik, fasa perancangan dan analisi perlu dilakukan semula.

4.3 Fasa Implementasi

Setelah fasa reka bentuk dijalankan, proses implementasi akan dilakukan. Jika terdapat ralat pelaksanaan, fasa reka bentuk harus dilakukan bagi menyelesaikan masalah yang dihadapi.

4.4 Fasa Pengujian

Fasa ini dilaksanakan terhadap sistem yang dibangunkan agar dapat memenuhi skop dan objektif kajian. Jika terdapat ralat, fasa implementasi perlu dilakukan semula supaya dapat meningkatkan lagi kecekapan sistem



Rajah 1: Model Pembangunan Sistem Peringkasan Teks

5 HASIL KAJIAN

Bahagian ini membincangkan hasil daripada proses pembangunan sistem peringkasan teks dalam Bahasa Melayu. Dalam bahagian ini juga, fungsi yang terdapat di dalam algoritma

diterangkan fungsinya dari segi input yang diterima atau hasil yang dipulangkan. Pengujian alatan ringkasan teks yang dibincangkan dalam bahagian ini akan menerangkan tentang kebolehan alatan menghasilkan ringkasan.

Perisian yang digunakan untuk membangunkan algoritma ringkasan petikan Bahasa Melayu ialah Python 2.7. Algoritma yang dibangunkan untuk menghasilkan ringkasan teks Bahasa Melayu ialah dengan cara menggunakan perkataan-perkataan yang mempunyai frekuensi tinggi dalam sesebuah teks.

Algoritma yang digunakan untuk membangunkan sistem ini adalah dengan menggunakan algoritma frekuensi perkataan atau "*word frequency*". Dalam pembinaan algoritma ini, senarai yang mengandungi perkataan yang mempunyai frekuensi dan skor tinggi digunakan dalam menentukan ayat yang akan menghasilkan ringkasan. Penentuan perkataan yang mempunyai frekuensi tinggi ditentukan melalui kekerapan perkataan tersebut muncul di dalam petikan teks. Semakin kerap sesuatu perkataan itu muncul dalam sesuatu dokumen maka semakin tinggi nilai perkataan tersebut. Nilai perkataan dikira menggunakan persamaan berikut:

```
wordlist = st.stopword(content)
wordfreq = []
for w in wordlist:
if (wordlist.count(w) > 6):
wordfreq.append(wordlist.count(w))
```

f (w) = Jumlah perkataan terkandung di dalam petikan

Dimana w, merupakan perkataan terkandung di dalam ayat.

Akan tetapi perkataan yang berfungsi sebagai kata nama, atau kata hubung tidak diambil kira sebagai perkataan yang mempunyai frekuensi tinggi. Hal ini kerana perkataan-perkataan seperti perkataan "yang", "dia", "ialah", "kita" merupakan perkataan yang kerap muncul di dalam ayat dan tidak mempunyai nilai dalam peringkasan. Ayat yang mengandungi perkataan yang nilainya tertinggi akan diambil untuk dijadikan sebagai hasil ringkasan.

Alatan untuk pengujian merupakan peranan yang penting bagi memeriksa sebarang ralat yang terdapat di dalam pembangunan ringkasan teks Bahasa Melayu. Ralat yang ditemui perlu

dibetulkan untuk memastikan alatan ringkasan dapat menghasilkan ringkasan yang dikehendaki. Pengujian dilakukan ke atas algoritma yang dibangunkan iaitu algoritma frekuensi perkataan atau "*word frequency*". Pengujian dilakukan dengan menggunakan tujuh petikan Bahasa Melayu yang di ambil dari laman sesawang.

Perkataan-perkataan yang mempunyai frekuensi tinggi dalam sesebuah petikan dapat dikenalpasti dengan menggunakan algoritma frekuensi perkataan. Ia juga akan dijadikan sebagai kata kunci untuk menghasilkan sebuah ringkasan. Kekerapan yang muncul dalam petikan ditunjukkan di dalam rajah.

Perkataan	Kekerapan perkataan muncul di dalam petikan
derita	8
gambaran	5
banjir	3
keadaan	4
kampung	2

Rajah 2 Frekuensi perkataan dalam petikan pertama

Perkataan	Kekerapan perkataan muncul di dalam petikan
jerebu	26
kesihatan	6
udara	5
pihak	3
kesan	3

Rajah 3 Frekuensi perkataan dalam petikan kedua

Perkataan	Kekerapan perkataan muncul di dalam petikan
banjir	10
kawasan	10
operasi	10
menyelamat	10
bencana	8
mangsa	8
kritikal	7
negeri	7
bantuan	7

Rajah 4 Frekuensi perkataan dalam petikan ketiga

Setelah berjaya menghasilkan ringkasan dari algoritma tersebut, perbandingan antara hasil ringkasan dijalankan. Ciri-ciri atau kriteria yang diambil kira ialah panjang ringkasan yang dihasilkan serta kualiti ringkasan yang dihasilkan. Perbandingan hasil ringkasan dari algoritma tersebut ditunjukkan dalam bentuk jadual.

	Petikan Asal	Algoritma Frekuensi Perkataan
Jumlah perkataan	2967	2163
Nisbah hasil ringkasan dengan petikan asal (%)		27

Jadual 1 Perbandingan hasil ringkasan dari teks pertama

	Petikan Asal	Algoritma Frekuensi Perkataan
Jumlah perkataan	5598	4351
Nisbah hasil ringkasan dengan petikan asal (%)		22

Jadual 2 Perbandingan hasil ringkasan dari teks kedua

	Petikan Asal	Algoritma Frekuensi Perkataan
Jumlah perkataan	2931	2572
Nisbah hasil ringkasan dengan petikan asal (%)		12

Jadual 3: Perbandingan hasil ringkasan dari teks ketiga

Berdasarkan perbandingan yang telah dilakukan, didapati bahawa terdapat pengurangan di antara hasil ringkasan yang dihasilkan oleh algoritma frekuensi perkataan dan petikan asal. Hasil ringkasan yang dihasilkan oleh algoritma frekuensi perkataan difahamkan lebih bertetapan dan informasi dapat dikekalkan sebanyak yang mungkin. Hal ini kerana, kekerapan perkataan yang muncul dalam sesebuah petikan memainkan peranan yang penting dalam menghasilkan ringkasan.

6 KESIMPULAN

Secara kesimpulannya, sistem peringkasan teks dalam Bahasa Melayu sedikit sebanyak membantu pengguna untuk memahami sesebuah teks dengan lebih mudah dan berinformasi.

Pengujian ini dilakukan untuk memastikan sesebuah ringkasan itu dapat dihasilkan bagi memenuhi objektif kajian yang dinyatakan dalam fasa perancangan dan analisis.

Secara tuntasnya, sistem peringkasan teks dalam Bahasa Melayu ini dapat membantu meringkaskan Bahasa Melayu untuk mendapatkan informasi daripada satu petikan yang panjang. Dengan adanya sistem ringkasan teks ini, masa yang diambil untuk memahami suatu teks dapat disingkatkan.

7 RUJUKAN

Andreas Gohr. Open Text Summarizer

Edmudson, H. P. 1969. New method in automatic extracting.

Jones. Automatic Summarising: the state of the art.

Luhm, H. P. 1958. The automatic creation of literature abstract.

Jacob Perkins. Python Text Processing with NLTK 2.0 Cookbook

Kamal Sarkar, 2009. Using Domain Knowledge for Text Summarization in Medical Domain

Mohamed Abdel Fatah, Fuji Ren. 2008. Automatic Text Summarization,

https://scholar.google.com/citations?user=7HD9rlEAAAAJ&hl=en

Mohamed Abdel Fatah. World Academy of Science, Engineering and Technology.

Mohd Sabri Hassan. Penjanaan Ringkasan Isi Utama Berdasarkan Ciri Kata Bagi Dokumen Berita Bahasa Melayu.

Shubhamajera, 2015. Automatic Text Summarization.

Suraya Alias, 2017. MYTextSum: A Malay Text Summarizer Model Using a Constrained Pattern-Growth Sentence Compression Technique

Word2vec. https://code.google.com/archive/p/word2vec/