Analisis Perbandingan Metode Spelling Corrector Peter Norvig dan Spelling Checker BK-Trees pada Kata Berbahasa Indonesia

Mutammimah[#1], Herry Sujaini[#2], Rudy Dwi Nyoto[#3]
[#]*Program Studi Teknik Informatika Fakultas Teknik Universitas Tanjungpura*
Jl. Prof Dr H. Hadari Nawawi, Kota Pontianak, 78115
[1]iema.mutammimah@gmail.com,[2]herrysujaini@gmail.com,[3]rudy_dn@yahoo.com

***Abstrak*- Bahasa menjadi faktor penting dalam penyampaian pengetahuan dan acuan dalam penulisan dokumen, komunikasi dan pencarian informasi. Apabila dalam penulisan dokumen terdapat kesalahan maksud penulisan tersebut menjadi berbeda. Oleh karena itu, dibutuhkan sebuah program yang dapat mendeteksi kesalahan penulisan dan memberikan sugesti kata yang benar. Salah satu fitur yang dapat digunakan untuk mendeteksi kesalahan dan memberikan sugesti kata yang benar adalah fitur *spelling corrector* atau *spelling checker* atau *spelling suggestion*. Fitur ini berfungsi sebagai pendeteksi kesalahan dan memberikan panduan bagi penggunanya dengan menandai kata-kata yang tidak terdaftar dalam kamus suatu bahasa tertentu. Fitur ini juga disertai dengan sugesti kata yang berfungsi menyediakan rekomendasi kata-kata yang mendekati kata yang dimaksud. Hal ini yang mendasari untuk membandingkan metode *spelling corrector* Peter Norvig dan *spelling checker* BK-Trees dalam hal memberikan sugesti kata menggunakan bahasa Indonesia sehingga dapat dijadikan acuan sebagai pilihan pengguna membuat aplikasi yang membutuhkan fitur pengoreksian kata. *Spelling Corrector* Peter Norvig menggunakan cara dengan mengubah jarak kata yang salah atau dengan mengubah kata yang salah menjadi dua kata dan sejumlah suntingan yang dibutuhkan untuk mengubah satu ke yang lain. BK-Trees atau Burkhard-Keller Tree adalah struktur data berupa pohon yang dibuat oleh Burkhard dan Keller pada tahun 1973 untuk mencari satu atau beberapa string yang mirip atau mendekati *string* yang menjadi *input*-an dan memanfaatkan metode *Levesthein Distance* untuk mendapatkan nilai sebagai pembanding kata yang salah dengan yang benar. Kamus yang dibuat menggunakan data korpus data berita online. Hasil analisis perbandingan metode *Spelling Corrector* Peter Norvig dan *Spelling Cherker* BK-Trees pada kata berbahasa Indonesia ini dapat diketahui bahwa metode Peter Norvig dapat memberikan 52,8% tingkat ketepatan yang lebih baik daripada metode BK-Trees yang menghasilkan 9%. Namun, metode *Spelling Checker* BK-Trees lebih unggul dalam hal tingkat keberhasilan 100% memberikan sugesti kata dan kecepatan rata-rata pemberian sugesti kata yang lebih rendah dari Peter Norvig.**

Kata kunci : analisis, *spelling corrector* Peter Norvig, *spelling checker* BK-Trees, korpus

I. PENDAHULUAN

Bahasa adalah salah satu komponen penting dalam kehidupan bermasyarakat. Dalam bentuk tulisan, bahasa menjadi faktor penting dalam penyampaian pengetahuan dari generasi ke generasi. Bahasa juga menjadi acuan dalam penulisan dokumen, komunikasi dan pencarian informasi. Apabila dalam penulisan dokumen terdapat kesalahan maksud penulisan, pencarian atau komunikasi tersebut menjadi berbeda. Oleh karena itu, dibutuhkan sebuah program yang dapat mendeteksi kesalahan penulisan dan memberikan sugesti kata yang benar.

Salah satu fitur yang dapat digunakan untuk mendeteksi kesalahan dan memberikan sugesti kata yang benar adalah fitur *spelling corrector* atau *spelling checker* atau *spelling suggestion*. Fitur ini berfungsi sebagai pendeteksi kesalahan dan memberikan panduan bagi penggunanya dengan menandai kata-kata yang tidak terdaftar dalam kamus suatu bahasa tertentu. Fitur ini juga disertai dengan sugesti kata yang berfungsi menyediakan rekomendasi kata-kata yang mendekati kata yang dimaksud. Hal ini dapat meminimalkan kesalahan eja atau salah ketik pada penulisan dokumen, pencarian informasi dan komunikasi.

Penerapan *spelling corrector* atau *spelling checker* yang sudah banyak digunakan pada bahasa Inggris juga dapat diimplementasikan pada bahasa Indonesia karena adanya kesamaan alfabet yang digunakan.

Spelling corrector Peter Norvig menjadi salah satu kandidat yang menarik untuk dibahas, karena pembuatnya adalah seorang *Director of Research* Google yang merupakan perusahaan *search engine* dan algoritma unik yang digunakan Peter dengan mengkombinasikan proses menghapus, menambah, mengubah dan mengganti huruf pada kata yang salah. Proses ini memungkinkan kombinasi kata dengan huruf alfabet menjadi lebih bervariasi untuk dapat di cek kembali pada kamus.

Sedangkan *spelling checker* BK-Trees yang mengusung kamus kata seperti akar pohon ini menggunakan nilai dari metode Levesthein Distance untuk menemukan kata yang benar dengan mencocokan nilai satu kata dengan kata lainnya.

Hal ini yang mendasari untuk membandingkan metode *spelling corrector* Peter Norvig dan *spelling checker* BK-Trees dalam hal memberikan sugesti kata menggunakan bahasa Indonesia sehingga dapat dijadikan acuan sebagai pilihan pengguna membuat aplikasi yang membutuhkan fitur pengoreksian kata.

Submitted 09-12-2016; Revised 09-12-2016; Accepted 09-12-2016

II. URAIAN PENELITIAN

A. Pengertian Analisis

Analisis adalah sekumpulan aktivitas dalam proses. Salah satu bentuk analisis adalah merakum sejumlah besar data yang masih mentah menjadi informasi yang dapat diinterpresikan. Semua bentuk analisis berusaha menggambarkan pola-pola secara konsisten dalam data sehingga dapat dipelajari dan diterjemahkan dengan cara yang singkat dan penuh arti.

B. Spelling Corrector Peter Norvig

Algoritma utama *Spelling Corrector* Peter Norvig adalah dengan mengubah jarak kata yang salah atau dengan mengubah kata yang salah menjadi dua kata dan sejumlah suntingan yang dibutuhkan untuk mengubah satu ke yang lain.

Split (memisahkan kata yang salah menjadi 2 kata dengan memisahkan antara 1 huruf dengan huruf selanjutnya), *deletion* (Menghapus satu huruf), *transposition* (pertukaran huruf yang berdekatan), *an alteration* (Mengubah satu huruf ke yang lain) atau *an insertion* (menambahkan sebuah huruf).

C. Spelling Checker BK-Trees

BK-Trees atau Burkhard-Keller Tree adalah struktur data berupa pohon yang dibuat oleh Burkhard dan Keller pada tahun 1973 untuk mencari satu atau beberapa string yang mirip atau mendekati *string* yang menjadi *input*-an. BK-Trees digunakan untuk melakukan '*fuzzy*' search dan *spelling checking*. Contongnya ketika kita mencari "*thab*" maka spell-checker akan menampilkan kata "*than*" dan "*that*".

Untuk mencari satu atau beberapa string yang mirip atau mendekati string yang menjadi *input*-an, dibutuhkan suatu metode yaitu dengan mancari *Leveshtein Distance* (Jarak Levenshtein). Jarak *Levenshtein* antara dua *string* adalah jumlah minimum penyisipan, penghapusan, dan penggantian yang dibutuhkan agar string sama dengan string lainnya.

D. Data Korpus

Istilah korpus berasal dari bahasa latin yang berarti *body* dan jamak dari korpus adalah *corpora* (Knowles dan Zuraidah Mohd. Don, 2006:9). Istilah ini diambil dari kata "korpus" yang terdapat pada kalimat *"...any body of texts, however small nad homogeneous, which the linguist assembled for the purpose of analysis...*" (Butler, 2004:150). Bagaimanapun, dalam kajian ini istilah yang digunakan adalah korpus atau data korpus sama dalam bentuk tunggal atau jamak. Dalam bidang linguistik, istilah korpus linguistik merupakan istilah modern yang muncul pada tahun 1980-an dan telah dijadikan metodelogi secara meluas dalam kajian linguistik. Pada umumnya istilah ini digunakan untuk merujuk data bahasa yang terbentuk yang dijadikan asas dalam penyelidikan linguistik (Leech, 1997:1).

E. Precision dan Recall

Precision adalah proposi jumlah dokumen yang ditemukan dan dianggap relevan untuk kebutuhan pencari informasi. *Recall* adalah proposi jumlah dokumen yang dapat ditemukan kembali oleh sebuah proses pencarian sistem IR (Putubuku:2008).

Recall sebenarnya sulit diukur karena jumlah seluruh dokumen yang relevan dalam database sangat besar. Oleh karena itu presisi-lah (*precision*) yang biasanya menjadi salah satu ukuran yang digunakan untuk menilai keefektivan suatu sistem temu balik informasi. Perolehan (*recall*) berhubungan dengan kemampuan sistem untuk memanggil dokumen yang relevan, sedangkan ketepatan (*precision*) berkaitan dengan kemampuan sistem untuk tidak memanggil dokumen yang tidak relevan (Hasugian, 206:5). Presisi juga merupakan cara ukur tingkat efektivitas sistem temu balik informasi. Pengukuran tingkat ketepatan (*precision*) dalam kegiatan penelusuran menurut Hasugian (2006:5):

Perolehan (*recall*) berhubungan dengan kemampuan sistem untuk memanggil dokumen yang relevan, sedangkan ketepatan (*precision*) berkaitan dengan kemampuan sistem untuk tidak memanggil dokumen yang tidak relevan (Hasugian, 2006 : 5). Berikut rasio tingkat ketepatan (*precision*) dalam kegiatan penelusuran menurut Hasugian (2006:5):

$$Precision = \frac{TP}{TP + TN}\ x\ 100\%$$

$$Recall = \frac{TP}{TP + FN}\ x\ 100\%$$

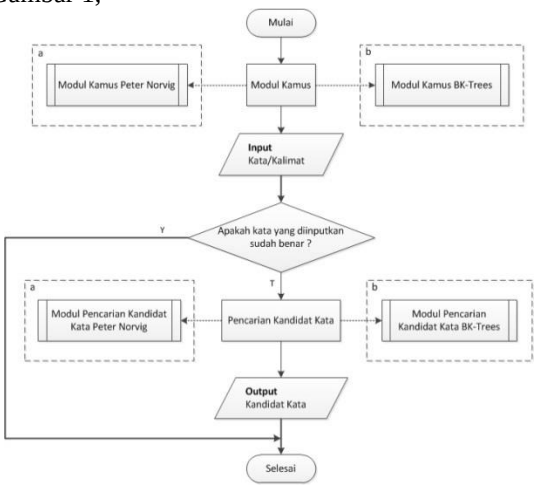
Keterangan :

Recall	=	presentasi tingkat keberhasilan metode menemukan sugesti yang tepat
TP (True Positive)	=	kata sugesti relevan yang sesuai dengan skenario pengujian
TN (True Negative)	=	kata yang tidak relevan dengan skenario pengujian, tapi tersugesti oleh metode
FN (False Negative)	=	kata yang relevan dengan skenario pengujian di dalam kamus, tapi tidak tersugesti oleh metode

III. PERANCANGAN DAN IMPLEMENTASI

A. Flowchart Aplikasi

Flowchart aplikasi secara keseluruhan ditunjukkan pada Gambar 1,



Gambar 1 *Flowchart* Aplikasi

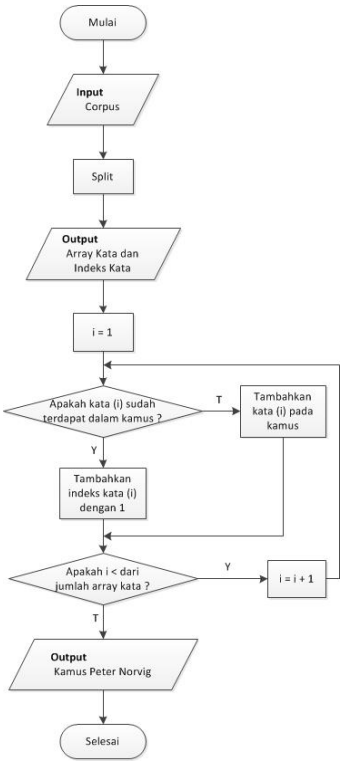
Aplikasi menerima *input*-an kata/kalimat yang salah untuk dapat diproses oleh metode untuk menghasilkan kata-kata suggestion yang mendekati kata yang dimaksud.

B. Flowchart Kamus Peter Norvig

Pada f*lowchart* Metode *Spelling Corrector* Peter Norvig memiliki 2 modul. Modul Membuat Kamus pada

Korespondensi : Mutammimah

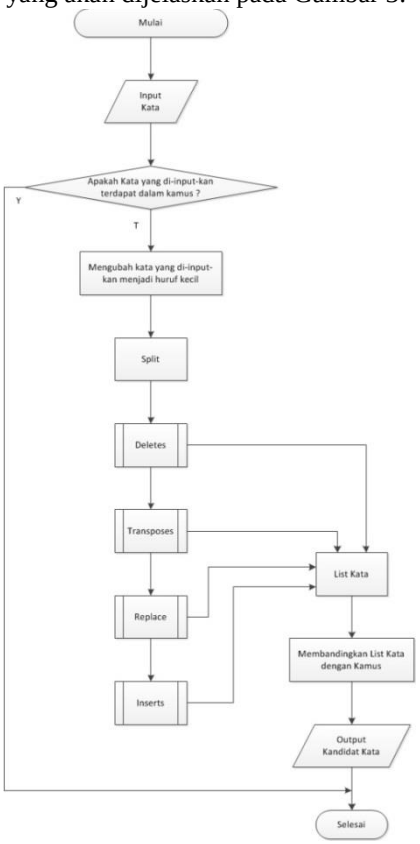
Metode *Spelling Corrector* Peter Norvig akan dijelaskan pada Gambar 2.



Gambar 2 *Flowchart* Kamus Peter Norvig

C. Flowchart Kandidat Kata Peter Norvig

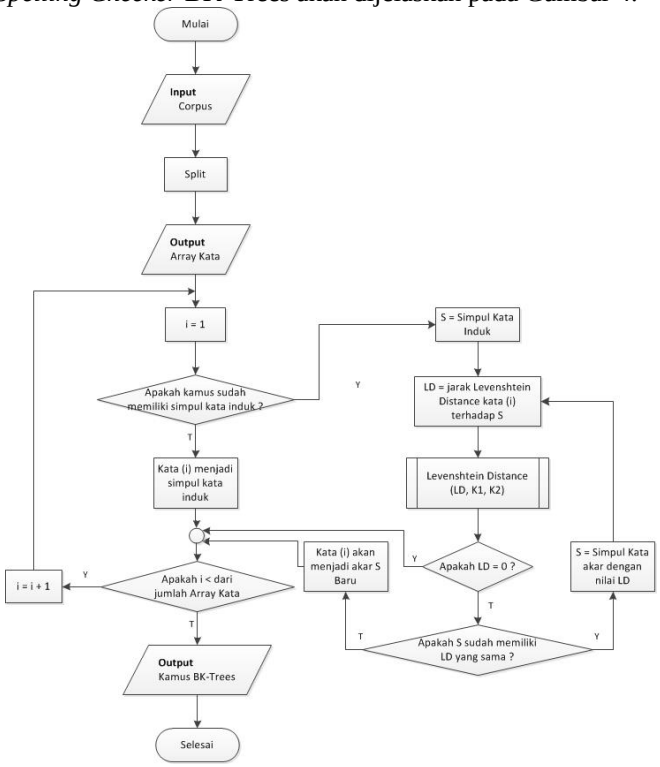
Selanjutnya Metode dapat menerima *input*-an kata/kalimat untuk dapat dicek apakah kata yang di-*input*kan salah atau benar yang di proses pada modul Pencarian Kandidat Kata yang akan dijelaskan pada Gambar 3.



Gambar 3 *Flowchart* Kandidat Kata Peter Norvig

D. *Flowchart* Kamus BK-Tees

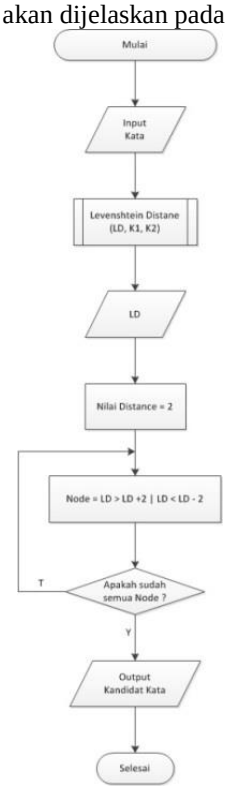
Pada *flowchart* Metode *Spelling Checker* BK-Trees terdapat 2 modul. Modul Membuat Kamus pada Metode *Spelling Checker* BK-Trees akan dijelaskan pada Gambar 4.



Gambar 4 *Flowchart* Kamus BK-Trees

E. *Flowchart* Kandidat Kata BK Trees

Selanjutnya Metode dapat menerima *input*-an kata/kalimat untuk dapat dicek apakah kata yang di-*input*kan salah atau benar yang di proses pada modul Pencarian Kandidat Kata yang akan dijelaskan pada Gambar 5.



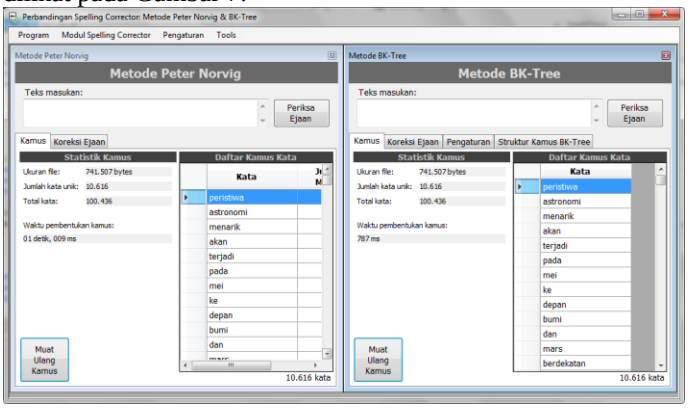
Gambar 5 *Flowchart* Kandidat Kata BK-Trees

Korespondensi : Mutammimah

IV. HASIL IMPLEMENTASI DAN PENGUJIAN APLIKASI

A. Hasil Perancangan

Form Utama Kamus merupakan Form yang digunakan untuk mengecek dan melihat korpus yang telah menjadi kamus di kedua metode. Pada form ini terdapat perhitungan ukuran file, jumlah kata unik dalam kamus, total kata, serta waktu yang dibutuhkan metode membuat kamus. Antarmuka Hasil perancangan form utama Kamus dapat dilihat pada Gambar 7.



Gambar 7 Form Utama Kamus

Form Utama Koreksi ini merupakan form yang digunakan untuk menampilkan hasil suggestion dari kata input-an yang salah.

B. Hasil Pengujian Keberhasilan Suggestion Kata

Pengujian ini dilakukan dengan 160 kata salah dengan 9 skenario kesalahan yang ditemukan pada korpus.

Metode	TP	TN	FN	Precision	Recall
Peter Norvig	186	162	56	53,4%	76,9%
BK-Trees	242	2445	0	9%	100%

C. Analisis Pengujian

Berikut ini adalah analisis dari hasil pengujian yang telah dilakukan :

1. Dari keseluruhan pengujian yang dilakukan didapatkan hasil 52,8% tingkat ketepatan metode Peter Norvig lebih baik dari BK-Trees dengan 9% tingkat ketepatan dalam menemukan kata sugesti yang sesuai dengan skenario pengujian pada setiap kata yang salah. Rendahnya tingkat ketepatan BK-Trees ini diakibatkan nilai *distance* yang di atur, semakin tinggi nilai *distance* yang di atur pada aplikasi, maka semakin banyak pula sugesti kata yang diberikan.
2. Dari keseluruhan pengujian juga didapatkan hasil 100% tingkat keberhasilan metode BK-Trees dalam menemukan kata yang benar dan sesuai dengan skenario pengujian daripada metode Peter Norvig yang menghasilkan 77,3% tingkat keberhasilan. Hal ini menjelaskan bahwa metode BK-Trees dapat mengambil semua sugesti yang sesuai dengan skenario pengujian dari korpus yang memiliki nilai *distance* 2.
3. Dari pengujian yang dilakukan ketahui bahwa dengan tingkat ketepatan yang tinggi, metode Peter Norvig masih memiliki kekurangan karena beberapa pengujian yang telah dilakukan ditemukan bahwa metode Peter Norvig tidak dapat menemukan keseluruhan sugesti yang ada dalam korpus, bahkan tidak dapat memberikan sugesti terhadap kata yang salah. Dapat dilihat pula Peter Norvig mulai kesulitan dengan kesalahan dengan 2 huruf, hal ini dimungkinkan karena algoritma Peter Norvig yang melakukan perubahan dengan rentan 1 huruf.
4. Dari pengujian yang telah dilakukan juga diketahui bahwa rendahnya tingkat ketepatan metode BK-Trees dikarenakan banyaknya kata sugesti yang diberikan oleh metode karena korpus masih terdapat kata bahasa Inggris dan singkatan atau kode yang menjelaskan suatu hal dalam sebuah artikel berita online.
5. Dari hasil waktu yang tercatat, Metode *Spelling Corrector* Peter Norvig memiliki keunggulan kecepatan pembentukan kamus daripada Metode *Spelling Checker* BK-Trees. Hal ini dimungkinkan karena data korpus yang digunakan masih tergolong kecil dan tidak sesuai dengan sistematika penggunaan *database* untuk *Spelling Cherker* BK-Trees.
6. Dari hasil waktu pencarian sugesti pada setiap kata yang di ujikan, bahwa rata-rata kecepatan waktu untuk pencarian kata pada Spelling Checker BK-Trees lebih stabil di kecepatan 3,611 ms, dibandingkan Spelling Corrector Peter Norvig. Hal ini dikarenakan sistem penyusunan kamus BK-Trees yang menggunakan kamus pohon yang meminimalisir pencarian yang terlalu jauh karena dibatasi dengan nilai *distance* yang telah di atur. Sedangkan pada Peter Norvig yang memanfaatkan sistem kamus model *list* kecepatan waktu yang lebih lama dalam pencarian kata dipengaruhi oleh model kamusnya, sehingga pada *list* kata yang paling bawah memerlukan waktu yang lebih banyak karena harus melakukan pengecekan satu persatu. Selain dikarenakan kamus Peter Norvig, kemungkinan kecepatan waktu yang lama juga karena sistem pencarian kandidat yang memerlukan modul *deletes*, *transposes*, *replace*, dan *insert* pada 1 kata yang berulang-ulang, pengulangan akan semakin panjang jika jumlah huruf pada kata yang dicari makin banyak.

V. KESIMPULAN

Kesimpulan yang didapat diambil dalam penelitian ini antara lain :

1. Hasil Analisis Perbandingan Metode *Spelling Corrector* Peter Norvig dan *Spelling Cherker* BK-Trees pada Kata Berbahasa Indonesia ini dapat diketahui bahwa metode Peter Norvig dapat memberikan 52,8% tingkat ketepatan yang lebih baik daripada metode BK-Trees yang menghasilkan 9%.
2. Hasil Analisis Perbandingan Metode *Spelling Corrector* Peter Norvig dan *Spelling Cherker* BK-Trees pada Kata Berbahasa Indonesia ini dapat diketahui bahwa Metode *Spelling Checker* BK-Trees lebih unggul dalam hal keberhasilan dengan 100% memberikan sugesti kata dan kecepatan rata-rata pemberian sugesti kata.

DAFTAR PUSTAKA

Ahli, Pengertian. Januari 18, 2016. *"Pengertian Analisis: Apa itu Analisis?"*. http://www.pengertianahli.com/2014/08/pengertian-analisis-apa-itu-analisis.html

Al-zuoud, Khalid M. Dan Mohammad K. Kabilan. 2013. *Investigating Jordanian EFL Student's Spelling*

Korespondensi : Mutammimah

Errors at Tertiary Level. Jurnal. Malaysia : University Science Malaysia

Bassil, Youssef dan Mohammad Alwani. 2012. *OCR Post-Processing Error Correction Algoritm Using Google's Online Spelling Sugesti*. Jurnal. Lebanon : Lebanese Association for Computational Sciences

Dagdag, Rowena dan Kimberly P. Weber. 2002. *The Use and Evaluation of a Sound Out or Error Only Sound Out Procedure on the Spelling Performance of a Third Grade Student*. Jurnal. Washington : Gonzaga University

Dataq. November 03, 2015. "*Perbedaan: precision, recall & accuracy"*. https://dataq.wordpress.com/2013/06/16/perbedaan-precision-recall-accuracy/

Foundation, Inc. Wikimedia. November 13, 2015. *"Spell checker"*. https://en.wikipedia.org/wiki/Spell_checker

Ganesan, Kavitan. Oktober 08, 2016. "*Computing Precision and Recall form Multi Class Classification Problems"*. http://text-analytics101.rxnlp.com/2014/10/computing-precision-and-recall-for.html

Gunaidi, Aang. Oktober 07, 2016. "*Recall and Precision Efektifitas Sistem Temu Kembali Informasi Melalui Penilaian Recall dan Precision di Perpustakaan Kementrian Sosial dan Tinjauan Menurut Islam*". http://aanggunaidi.blogspot.co.id/2014/04/efektivitas-sistem-temu-kembali_7.html

Hamberg, Erlend. November 03, 2015. *"BK-Trees"*. https://hamberg.no/erlend/posts/2012-01-17-BK-trees.html'

Johnson, Nick. Desember 17, 2015. *"Damn Cool Algorithms, Part 1:BK-Trees"*. http://blog.notdot.net/2007/4/Damn-Cool-Algorithms-Part-1-BK-Trees

Jurafsky, Daniel dan James H. Martin. 2000. *Speech and Language Processing*. United States of America : Prentice-Hall, Inc.

Klanting. Oktober 07, 2016. "*Pengertian Recall, Precision, F-Measure*". http://ladangbelajar.blogspot.co.id/2013/11/pengertian-recall-precision-f-measure.html

Mitton, Roger. Desember 20, 2015. *"Spellchecking by computer"*. http://www.dcs.bbk.ac.uk/~roger/spellchecking.html

Norvig, Peter. November 01, 2015. *"How to Write a Spelling Corrector"*. http://www.norvig.com/spell-correct.html

Norvig, Peter. Mei 01, 2016. *"Natural Language Corpus Data: Beautiful Data"*. http://norvig.com/ngrams/

Olimpiyanto Oktpriadi Limbong, Saul. November 22, 2016. "*Levensthein Distance*". https://saulimbong.wordpress.com/2011/05/17/levenshtein-distance/

Powers, David MW. 2007. *Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation*. Jurnal. South Australia :Flinders University

Small, Chris. Mei 25, 2016. *"How to Wrie a Spelling Corrector in C#"*. http://www.anotherchris.net/csharp/how-to-write-a-spelling-corrector-in-csharp/

Xenopax. Mei 25, 2016. *"The BK-Tree – A Data Structure for Spelling Checking"*. https://nullwords.wordpress.com/2013/03/13/the-bk-tree-a-data-structure-for-spell-checking/

Korespondensi : Mutammimah