# *Aplikasi String Matching Dalam Spell Checker*

*Yusak Yuwono Awondatu13514005*
*Program Studi Teknik Informatika*
*Sekolah Teknik Elektro dan Informatika*
*Institut Teknologi Bandung, Jl. Ganesha 10 Bandung 40132, Indonesia*
*13514005@std.stei.itb.ac.id*

***Abstrak* — makalah ini berisi pembahasan mengenai aplikasi string matching dalam Spell Checker, sebuah fitur komputer yang berfungsi untuk menganalisis teks yang diberikan oleh pengguna sehingga membantu pengguna dalam melakukan pengejaan kata yang tepat. Aplikasi ini mungkin terlihat sepele, namun dapat menjadi krusial dalam dunia nyata.**

***Keywords — String Matching, Spell Checker, brute force***

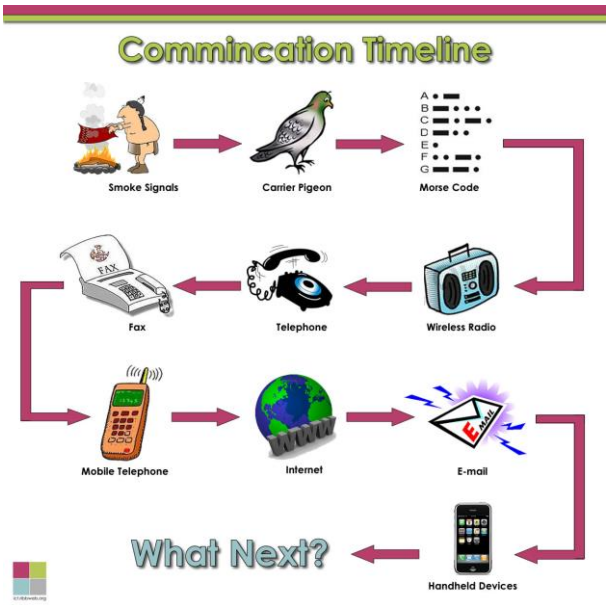
## I. PENDAHULUAN

[1]Komunikasi adalah salah satu kegiatan utama yang dilakukan manusia sehari-harinya. Bahkan sejarah dan kebudayaan membuktikan bahwa setiap manusia di berbagai tempat memiliki caranya untuk berkomunikasi. Namun penghambat utama dalam komunikasi adalah jarak karena komunikasi umumnya dilakukan dengan berbicara dan berbicara hanya dapat dilakukan dengan bertemu. Pesan tidak akan dapat sampai jika lawan bicara tidak berada dihadapan kita. Seiring dengan perkembangan jaman, manusia mulai memikirkan berbagai macam cara untuk mengilangkan kelemahan ini, yaitu komunikasi jarak jauh. banyak cara yang dilakukan untuk berkomunikasi jarak jauh, dimulai dari sinyal asap, kurir hewan ataupun manusia untuk mengirimkan pesan tertulis, telegraf untuk komunikasi teks jarak yang lebih jauh yang lebih cepat dan terenkripsi, lalu berlanjut ke telepon untuk komunikasi suara jarak jauh, mesin fax yang menyertai pesan teks untuk telepon, radio untuk *broadcast* pesan, mobil phone yang merupakan telepon nirkabel, hingga saat ini yang termutakhir adalah komputer dan Smartphone yang menggunakan internet.

Smartphone kini menjadi alat yang paling banyak digunakan diantara computer masa kini. Hampir setiap orang, tua maupun muda mengenalnya dan memilikinya. Tidak hanya nirkabel dan pocket size seperti mobile phone biasa, smartphone memiliki puluhan fitur lainnya sehingga tidak hanya berfungsi sebagai alat komunikasi.

Tentunya berbeda dengan komunikasi jaman dulu yang rumit. Smartphone memudahkan penggunanya berkomunikasi satu sama lain. Baik dengan teks maupun dengan suara.

Pada makalah ini akan dibahas komunikasi teks yang digunakan dalam komputer. Lebih spesifiknya didalam Spell Checker. Makalah ini akan membahas apa itu Spell Checker, cara kerja dan kegunaannya dalam komunikasi masa kini.

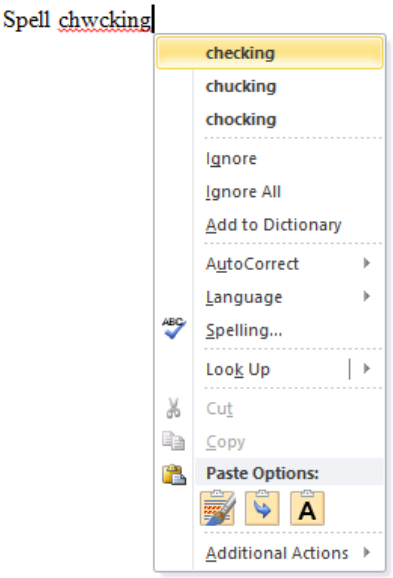



Gambar 1. Sejarah Singkat Komunikasi

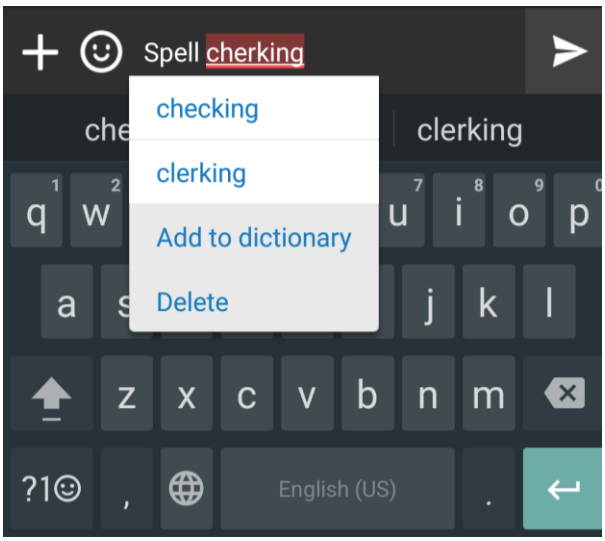
## II. DASAR TEORI

### A. *Spell Checker*

[5]Spell Checker adalah salah satu fitur yang terdapat didalam computer, khususnya didalam aplikasi yang berhubungan dengan pemroses kata seperti *Office Application*, selain itu Search Engine juga umumnya menggunakan Spell Checker untuk hasil pencarian yang lebih optimal. Spell Checker akan memberi tanda pada teks yang tidak memenuhi standard bahasa yang benar umumnya berupa garis merah dibagian bawah teks tersebut, pada umumnya spell checker juga akan memberikan sugesti kata yang memiliki ejaan yang lebih tepat untuk kata yang dinyatakan salah.




Gambar 2. Spell Checker pada Microsoft Word




Gambar 3. Spell Checker Google Keyboard Android

Spell Checker juga terdapat dalam Shartphone, biasanya lebih dikenal dengan istilah Google Keyboard dan Autocorrect. Fungsinya juga sama dengan yang terdapat dalam computer biasa, yakni memperbaiki kesalahan dalam pengejaan kata.

Pada pengembangannya, Spell Checker juga dapat secara otomatis mengubah teks tersebut menjadi teks yang terdapat dalam sugesti sehingga pengguna dapat mengetik lebih cepat tanpa menghiraukan adanya kesalahan ketik karena akan langsung diganti. Fitur ini umumnya menggunakan Bahasa Inggris sebaga Bahasa yang digunakan internasional.

## III. Cara Kerja

Pada proses Spell Checking, akan dibagi menjadi 2 macam kasus kesalahan yang biasa ditemui.

### A. *Kasus 1*

Pada Kasus pertama, kesalahan yang terjadi adalah adanya kesalahan pengetikan, contohnya pada kata CHECKING salah eja menjadi CHWCKING. Kesalahan pengejaan ini hanya berupa huruf yang salah ketik, dan panjang huruf tetaplah sama. Biasanya terjadi karena menekan huruf yang bersebelahan dengan keyboard dan sangat mudah untuk terjadi salah pengetikan.

Untuk memeriksa kesalahan seperti ini, dapat dilakukan secara sederhana dengan menggunakan *brute force string matching*, yakni dengan membandingkan setiap huruf yang ada didalam kata yang salah ketik dengan huruf dari kata yang ada dalam kamus Bahasa, lalu memcari kata dengan persentase ketepatan tertinggi.

Sebagai contoh sederhana pada gambar 2 dan 3

kata yang tepat :

CHECKING

Kesalahan :

CHWCKING (kesalahan pada huruf E menjadi W)

CHERKING (kesalahan pada huruf C menjadi R)

CHECKIGN (kesalahan pada huruf N dan G yang terbalik)

Dengan algoritma brute force, CHWCKING memiliki kemiripan 87.5% dengan CHECKING, CHUCKING, dan CHOCKING

Sedangkan CKERKING memiliki 87.5% kemiripan dengan CHECKING dan CLERKING.

Untuk CHECKIGN memiliki 75% kemiripan dengan CHECKING.

### B. *Kasus 2*

Pada kasus kedua, kesalahan pengetikan berupa huruf yang tidak tertekan, atau huruf yang kelebihan ditekan. Hal ini biasanya terjadi karena pengguna mengetik huruf terlalu cepat sehingga tanpa sadar luput atau kelebihan mengetikkan suatu huruf tertentu. Kesalahan ini akan mengubah panjang karakter sehingga ejaan menjadi tidak tepat.

Sebagai contoh :

Kata yang tepat :

SPELLING

Kesalahan :

SPELING (kurang satu huruf L di tengah)

ASPELLING (kelebihan satu huruf A di awal)

SPELLINGF (kelebihan satu huruf F di akhir)

Semua kesalahan tersebut memiliki kesalahan kecil berupa kekurangan atau kelebihan satu huruf. Namun jika menggunakan brute force, pada kasus SPELING, akan terjadi kesalahan pengecekan dimulai dari ING sehingga hanya 4/7 huruf yang dinyatakan benar. Pada kasus SSPELLING, kesalahan di awal tidak akan terlalu berpengaruh pada string matching sehingga mendapat 8/9 huruf yang benar. Untuk kasus SPELLINGF, kesalahan terdapat pada akhir saja sehingga 8/9 huruf benar dan mendapat suggesti SPELLINGS dan SPELLING.

Brute force memang masih dapat digunakna untuk sebagian kasus 2, namun hanya efektif untuk kesalahan pengejaan di awal dan di akhir. Kesalahan pengejaan di tengah kata akan menyebabkan kekacauan dalam pencarian sehinggan memberikan ketepatan hasil yang rendah. Oleh karena itu diperlukan pengembangan, salah satunya dengan memotong huruf pada posisi tertentu, lalu membandingkannya dengan kata yang terdapat pada kamus

Sebagai salah satu contoh :

SPELING dapat dipotong di tengah menjadi

| | | |
|---|---|---|
| S + PELING | SP + ELING | SPE + LING |
| SPEL + ING | SPELI + NG | SPELIN + G |

SPELLING dapat dipotong menjadi

| | | |
|---|---|---|
| S + PELLING | SP + ELLING | SPE + LLING |
| SPEL + LING | SPELL + ING | SPELLI + NG |
| SPELLIN + G | | |

Jika SPELING dibandingkan dengan SPELLING, maka akan terdapat kecocokan pada S, SP, SPE, SPEL, ING, NG, dan G. sehingga dapat ditemukan bahwa ada karakter yang hilang diantara L dan I. jika dilakukan pemotongan lebih banyak maka dapat mengatasi kesalahan pengetikan lebih dari satu buah.

Cara lain yang dapat digunakan adalah dengan brute force menghitung banyak karakter yang digunakan oleh kata tersebut, memasukkannya kedalam array jumlah karakter, lalu membandingkannya dengan karakter yang terdapat dalam kamus.

Contoh :

SPELING terdiri dari

| | | | | | |
|---|---|---|---|---|---|
| A = 0 | B = 0 | C = 0 | D = 0 | E = 1 | F = 0 |
| G = 1 | H = 0 | I = 1 | J = 0 | K = 0 | L = 1 |
| M = 0 | N = 1 | O = 0 | P = 1 | Q = 0 | R = 0 |
| S = 1 | T = 0 | U = 0 | V = 0 | W = 0 | X = 0 |
| Y = 0 | Z = 0 | | | | |

Dibandingkan dengan SPELLING yang merupakan kata tujuan, maka hanya terdapat satu perbedaan huruf sehingga mendapat kecocokan 87.5%

## IV. Analisis Metode

### A. *Brute Force String Matching*

Menggunakan cara brute force berarti membandingkan kata dengan seluruh kata yang ada didalam kamus Bahasa. [6]Berdasarkan kamus Oxford English Dictionary, terdapat 171,476 kata dalam Bahasa inggris. Dengan panjang karakter rata-rata 7.6 karakter[3]. Sehingga pencarian dengan brute force akan medapatkan worst case 1,303,217.6 kali pencocokan. Jika digabung dengan increment karakter, maka akan menghasilkan rata-rata 9,904,453.76 operasi.

Namun tidak hanya sampai disana, kamus kata Inggris selalu mengalami *update* karena melalui kehidupan sosial muncul kata-kata baru yang diciptakan. Contohnya seperti ‘Selfie’, ‘Tweet’, ‘Facebooking’, dan bahasa slang internet seperti ‘OMG’(Oh My God), ‘TL;DR’(Too Long; Didn’t Read), ‘THX’ (Thank You), dan sebagainya.[8][9]

[4]Dengan pencarian menggunakan WolframAlpha pada 8 Mei 2016, sekarang sudah terdapat 183.351 kata dalam kamus Bahasa Inggris. Bahkan jika melihat Bahasa nonformal dan istilah yang digunakan sehari-hari yang tidak tercatat dalam kamus Oxford, terdapat 1.025.110 kata[10], jumlah ini 10 kali lebih banyak dibandingkan yang tercatat dalam kamus dan akan terus bertambah. Belum lagi jika hal ini diterapkan kedalam Bahasa lain yang memiliki kosa kata yang lebih banyak dibandingkan Bahasa Inggris.

### B. *Optimalisasi Brute Force*

[4]Terdapat sejumlah cara untuk mengoptimalkan pencarian, salah satunya dengan menggunakan indeks huruf pertama sehingga pencarian brute force dibatasi dengan huruf pertama dari kata tersebut.

Berdasarkan 183.351 kata yang ditemukan melalui pencarian dengan WolframAlpha, distribusi kata dengan awalan huruf tertentu adalah sebagai berikut

| letter | Count | Letter | count |
|---|---|---|---|
| a | 13067 | n | 5084 |
| b | 10893 | o | 4820 |
| c | 17004 | p | 15083 |
| d | 10165 | q | 924 |
| e | 6897 | r | 8671 |
| f | 6784 | s | 19734 |
| g | 5836 | t | 9360 |
| h | 7586 | u | 5620 |
| i | 6316 | v | 2657 |
| j | 1663 | w | 4268 |
| k | 2163 | x | 311 |

| | | | |
|---|---|---|---|
| l | 5849 | y | 708 |
| m | 11212 | z | 676 |
| **Total** | | | **183351** |
| **Average** | | | **7051.96** |

Tabel 1. distribusi kata dengan awalan huruf tertentu

Dengan menggunakan indeks, pencarian, selama tidak ada kesalahan pada awal huruf, pencarian bisa dibatasi hingga 1/26 dari keseluruhan kata yang ada

Salah satu cara lainnya juga dengan membandingkan panjang kata yang ada. Berdasarkan data dari http://norvig.com/mayzner.html yang melakukan analisis terhadap 97.565 kata, distribusi kata adalah sebagai berikut

| length | count | length | count |
|---|---|---|---|
| 1 | 26 | 13 | 2,272 |
| 2 | 662 | 14 | 1,202 |
| 3 | 4,615 | 15 | 668 |
| 4 | 6,977 | 16 | 283 |
| 5 | 10,541 | 17 | 158 |
| 6 | 13,341 | 18 | 64 |
| 7 | 14,392 | 19 | 40 |
| 8 | 13,284 | 20 | 16 |
| 9 | 11,079 | 21 | 1 |
| 10 | 8,468 | 22 | 5 |
| 11 | 5,769 | 23 | 2 |
| 12 | 3,700 | | |
| **Total** | | | **97565** |
| **Average Length** | | | **7.6** |
| **Average Count** | | | **4241.96** |

Table 2. distribusi kata dengan panjang karakter per kata

Dengan menggunakan indeks huruf pertama atau panjang kata, atau pula keduanya pencarian dapat dikecilkan menjadi ribuan kata saja, jauh lebih kecil dibandingkan dengan brute force biasa yang membandingkan dengan lebih dari 180.000 kata.

Namun pencarian dengan menggunakan index maupun panjang kata dapat menemukan kebuntuan. Misalnya saja jika pada kasus 1, jika terjadi kesalahan ketik pada awal karakter akan menyebabkan index menuju index yang salah sehingga malah dilakukan pencarian pada tempat yang sama sekali tidak sesuai. Untuk kasus 2 juga tidak bisa karena panjang karakter yang berbeda juga akan menyebabkan proses pencarian di tempat yang tidak sesuai. Karena itu brute force tetaplah kurang efisien untuk melakukan Spell Checking.

### *C. Potongan kata*

Menggunakan potongan dari kata akan memeriksa bagian dari kata yang diberikan, akan terdapat n-1 potongan dan membuat pencocokan akan hamper sama lamanya dengan brute force, namun akan menemukan hasil yang lebih presisi karena kata yang dipenggal akan sampai pada lokasi yang tepat.

Selain itu, dengan melakukan potongan kata yang lebih banyak, seperti memotong per huruf, lalu melakukan pencocokan first fit, maka pencocokan karakter akan dapat dilakukan untuk lebih dari satu macam kesalahan pengetikan.

### *D. Menghitung Karakter*

Menghitung banyak karakter yang digunakan kata akan menghasilkan jawaban yang lebih presisi, baik untuk kasus 1 maupun kasus 2, tapi hal ini berarti melakukan secara fix 26 kali perbandingan terhadap ratusan ribu kata dalam dalam kamus.

Metode ini akan menghasilkan hasil yang lebih tepat dibandingkan dengan brute force string matching, dan juga hasil yang relative lebih cepat untuk kata yang lebih panjang dari 5 huruf karena brute force string matching akan memiliki waktu rata-rata $n^2$ dikalikan banyak kata dalam kamus karena kesalahan pengetikan umumnya memiliki panjang karakter yang tidak terlalu jauh. Cara ini juga akan lebih cepat dibandingkan dengan potongan kata-kata karena hanya menghitung banyak karakter yang digunakan tanpa memperhitungkan posisinya

Kelemahan dari cara ini adalah jika terdapat anagram atau kata yang memiliki komposisi huruf yang sama, contohnya mengetikkan kata PRESNETS mengharapkan kata yang dimaksud adalah PRESENTS. Namun PRESENTS juga punya anagram SERPENTS dan PERTNESS sehingga akan masuk kedalam daftar sugesti.

Namun hal ini dapat diantisipasi dengan melakukan pengecekan baik secara brute force biasa maupun dengan potongan kata.

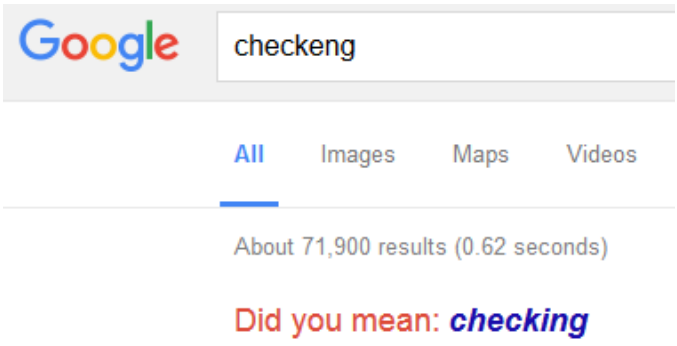
### *E. Machine Learning*

Terdapat pula cara yang lebih sederhana, namun membutuhkan tenaga dan waktu yang panjang, yaitu Machine Learning. Machine learning akan mempelajari sendiri tingkah laku penggunanya yang menyadari adanya kesalahan dalam penulisan, lalu memasukkannya kedalam database sehingga dari ribuan atau jutaan kali kesalahan, mesin dapat membuat keesimpulan mengenai kata yang tepat dari kesalahan tersebut. Umumnya cara ini digunakan oleh search engine yang menganalisis seluruh penggunanya

Sebagai contoh : seseorang mencari sebuah kata ALGORITM (yang memiliki kesalahan) dengan search engine. Karena kesalahan pengejaan, search engine tidak akan memberikan hasil apapun. Karena tidak memberikan hasil apapun, pennggguna akan mengecek ulang kata kunci yang

dicari dan menyadari sendiri kesalahannya dan memperbaikinya menjadi ALGORITHM. Kesalahan ini tentunya tidak hanya satu diantara 7 miliar manusia yang melakukan pencarian setiap saat. Dari ratusan-ribuan-jutaan kesalahan, mesin pencari akan menyimpulkan pencarian kata yang tepat dari kesalahan yang dilakukan oleh penggunanya.

Satu-satunya kelemahan cara ini adalah jika mayoritas pengguna melakukan kesalahan tanpa memperbaikinya, maka mesin akan menyetujui bahwa kesalahan tersebut adalah kebenaran sehingga selalu menampilkan informasi yang salah karena berdasarkan mayoritas pemakai search engine adalah benar.




Gambar 4. Search Engine Google

## APLIKASI

Aplikasi dari Spell Checking terdapat dalam kehidupan sehari-hari, misalkan saja dalam menulis makalah ini spell checking juga bekerja untuk memperbaiki istilah-istilah Bahasa Inggris yang tidak tepat pengejaannya. Dalam dunia internasional, spell checking berguna untuk menuliskan pesan yang formal sehingga lawan bicara dapat membaca pesan yang diberikan tanpa adanya kesalahan pengetikan yang dapat mengakibatkan miskomunikasi.

Pada dunia internet yang mayoritas menggunakan Bahasa Inggris, Bahasa inggris yang formal perlu digunakan agar tidak terjadi misinformasi karena kesalahan pengetikan, karena itu kebanyakan search engine yang merupakan penyokong utama dari dunia internet memiliki fungsi spell checker, contohnya seperti Google, Bing, Yahoo, dan search engine lainnya menggunakan Spell Checker, spesifiknya dengan Machine Learning.

Di wilayah sosial yang mayoritas menggunakan Bahasa Inggris, Spell Checking sangatlah berguna karena dapat mempercepat waktu yang dipakai untuk menuliskan pesan teks. Namun untuk wilayah yang tidak menggunakan Bahasa Inggris sebagai *native language*, Spell Checking akan sangat mengganggu karena akan memberikan sugesti perubahan Bahasa yang tidak sesuai dengan Bahasa lokal yang digunakan, contohnya dalam Bahasa Indonesia, kata-kata seperti Teh, tabel, komputer, fokus, akan langsung mengalami pengubahan dari Spell Checker sehingga membuat makalah dalam Bahasa Indonesia juga menjadi lebih sulit.. mengatasi hal ini dapat dilakukan dengan mematikan spell checker, selain itu Spell Checker sendiri juga mulai dibuat dalam Bahasa masing-masing dengan memasukkan kamus Bahasa lokal kedalam database dari Spell Checker, meskipun akan menggunakan memory lebih, spell checking akan memudahkan penggunanya mengetik dengan lebih formal.

Namun Spell Checking juga kurang tepat penggunaannya untuk Bahasa *chat* yang digunakan sehari-hari karena kebanyakan pengguna teks lebih suka menggunakan Bahasa singkatan. Contohnya dalam Bahasa Inggris seperty thx, srsly, drving, gr8, 4ever, dan singkatan lain yang telah diberikan pada pembahasan karena dapat dituliskan lebih singkat dan lebih cepat. Selain itu Spell Checker yang menggunakan machine learning umumnya juga akan memeriksa kata-kata yang dipakai pengguna dalam kehidupan sehari-harinya sehingga jika melakukan chat dengan orang diluar kegiatan sehari-hari, seringkali akan terjadi mismatch kata dan menyebabkan keambiguan ataupun miskomunikasi.


Gambar 5. Contoh terjadinya mismatch pada Spell Checker

## KESIMPULAN

Spell Checker adalah salah satu fitur dari computer yang berfungsi untuk memperbaiki kesalahan dalam pengejaan kata dan menjadikannya kata yang lebih tepat.

Spell Checker bekerja baik dengan menjalankan algoritma yang mencocokkan kata yang dituliskan dengan kata yang terdapat dalam kamus Bahasa, maupun dengan machine learning yang mempelajari pengguna yang memperbaiki kesalahannya sendiri.

Spell Checker dipakai dalam pesan tertulis dan umumnya menggunakan Bahasa formal dalam dunia internasional, namun kurang begitu efektif dalam dunia sehari-hari yang lebih suka menggunakan Bahasa singkatan dalam *chatting* atau *texting*.

## REFERENCES


[1] http://www.historyworld.net/wrldhis/PlainTextHistories.asp?historyid=aa93
diakses pada 7 Mei 2016

[2] http://norvig.com/spell-correct.html
diakses pada 7 Mei 2016

[3] http://norvig.com/mayzner.html
diakses pada 7 Mei 2016

[4] http://www.wolframalpha.com
diakses pada 7 Mei 2016

[5] https://www.google.com/patents/US20020194229
diakses pada 7 Mei 2016

[6] http://www.oxforddictionaries.com/words/how-many-words-are-there-in-the-english-language
diakses pada 7 Mei 2016

[7] http://www.internetmarketingninjas.com/blog/search-engine-optimization/google-spell-check/
diakses pada 8 Mei 2016

[8] http://www.hongkiat.com/blog/dictionary-words-from-internet/
diakses pada 8 Mei 2016

[9] http://www.oxforddictionaries.com/words/what-s-new
diakses pada 8 Mei 2016

[10] http://www.languagemonitor.com/number-of-words/number-of-words-in-the-english-language-1008879/
diakses pada 8 Mei 2016


## PERNYATAAN

Dengan ini saya menyatakan bahwa makalah yang saya tulis ini adalah tulisan saya sendiri, bukan saduran, atau terjemahan dari makalah orang lain, dan bukan plagiasi.

Bandung, 9 Meil 2016

ttd

Yusak Yuwono Awondatu

13514005