

Spleling Corecssion

Alfan Farizki Wicaksono

Fakultas Ilmu Komputer, Universitas Indonesia

853.000.000 Results

Date ▾

Open links in new tab



Indonesia

Country

Climate

Official language

Economy

Culture

Map

Geography

History

Political system

On this page:



Travel



Related Search



Booking.com

Including results for **Indonesia**.

Do you want results only for indnonsai?

Indonesia

<https://www.bing.com/travel/place-information>

Popular destinations



Bali

Mount Agung, Tanah Lot,
Mount Batur

Jakarta

National Monument, Istiqlal
Mosque, Taman Mini Indon...

Surabaya

Mount Bromo, Surabaya Zoo,
Heroes Monument

Country in southeast asia

The world's largest island country and the 14th-largest
country by land area.Capital
JakartaOfficial language
Indonesian



indnonsai



Login

[Semua](#)

[Maps](#)

[Berita](#)

[Video](#)

[Gambar](#)

[Lainnya](#)

Alat

SafeSearch a

Sekitar 5.700.000.000 hasil (0,77 detik)

Menampilkan hasil untuk **Indonesia**

Atau telusuri [indnonsai](#)

<https://id.wikipedia.org/wiki/Indonesia>

Wikipedia bahasa Indonesia, ensiklopedia bebas

Indonesia merupakan negara terluas ke-14 sekaligus negara kepulauan terbesar di dunia dengan luas wilayah sebesar 1.904.569 km², serta negara dengan pulau ...

Bahasa daerah: Lebih dari 700 bahasa

Agama (2018): 86,70% Islam; 10,72% ...

Kota terbesar: [Jakarta](#); 6°10'S 106°49'E / ...

Format tanggal: DD/MM/YYYY

[Indonesia Raya](#) · [Bahasa Indonesia](#) · [Bangsa Indonesia](#) · [Presiden Indonesia](#)



Indonesia

Negara di Asia

Indonesia, dengan nama resmi Republik Indonesia, atau lengkapnya Negara Kesatuan Republik Indonesia, adalah sebuah negara kepulauan di Asia Tenggara yang dilintasi garis khatulistiwa dan berada di antara ... [Wikipedia](#)

Ibu kota: [Jakarta](#)

Luas: 1,905 juta km²

Presiden: [Joko Widodo](#)

Populasi: 273,5 juta (2020) [Bank Dunia](#)

Berita utama

ANTARAKASEL

[Anggota DPRD Kalsel Karlie: Pancasila aset negara Indonesia](#)

2 jam lalu



ANTARANEWS.com

[Borussia Dortmund akan jalani tur ke Indonesia](#)

2 jam lalu



Tolerant Retrieval

Bagaimana membuat search engine yang "mempunyai toleransi" terhadap kesalahan syntax pada query?

Rates of Spelling Errors -> **26% for Web queries** (Wang et al., 2003)

Types of Spelling Errors

- Non-word Errors

- graffe -> giraffe

Kata-kata yang tidak ditemukan di kamus baku.
Biasanya **context insensitive**

- Real-word Errors

- Typographical errors

- three -> there

- Cognitive errors

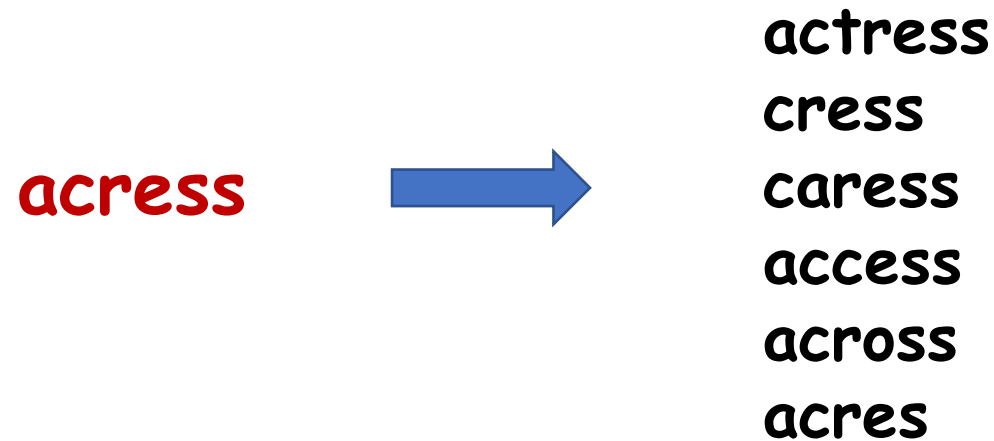
- piece -> peace
- too -> two
- your -> you're

Biasanya **context sensitive**

Non-Word Errors: Bagaimana?

acress

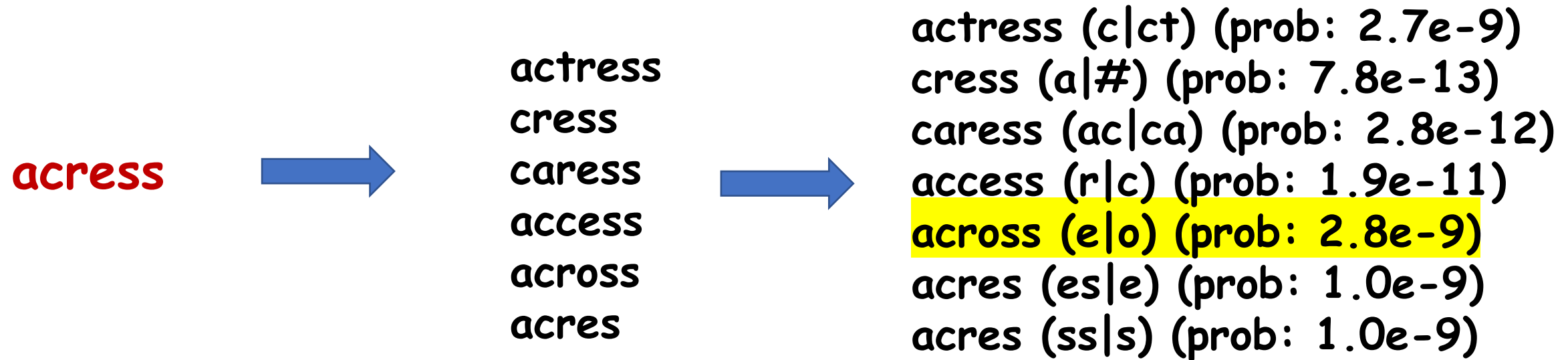
Non-Word Errors: Bagaimana?



Candidate Generation

- Words with similar spelling
- Words with similar pronunciation

Non-Word Errors: Bagaimana?



Choose the best candidate!

Scoring with a probability model,
such as Noisy Channel Model.

Candidate Testing: Minimal Edit Distance

- Kita perlu mekanisme atau metrik untuk mengukur “kedekatan ejaan” antara dua buah string.
- Salah satu realisasi “kedekatan ejaan” -> berapa banyak langkah minimal yang perlu saya lakukan untuk edit string X ke string Y .
- Metrik tersebut perlu efisien untuk dihitung.

4 Kemungkinan Edit Actions

- Insertions

car → car^r

- Deletions

train^e → train

- Substitutions

^cart → ^dart

- Transpositions

^{ac}t → ^{ca}t

Words within 1 of **acress**

Introduction to Information Retrieval				
Words within 1 of acress				
Error	Candidate Correction	Correct Letter	Error Letter	Type
acress	actress	t	–	deletion
acress	cress	–	a	insertion
acress	caress	ca	ac	transposition
acress	access	c	r	substitution
acress	across	o	e	substitution
acress	acres	–	s	insertion

a cash -> an act

#Edit actions = 5

- a cash -> an cash -> an aash -> an acsh -> an acth -> an act

insertion

substitution

substitution

substitution

deletion

- a cash -> an cash -> an acsh -> an ach -> ac ac -> an act

insertion

transposition

deletion

deletion

insertion

#Edit actions = 5

- a cash -> an cash -> an acsh -> an acth -> an act

insertion

transposition

substitution

deletion

#Edit actions = 4

Dan yang lainnya. Ada banyak kemungkinan ...

Damerau-Levenshtein Edit Distance

Given two strings x and y , **DL Edit Distance** between them is the shortest or cheapest possible sequence of edit actions from x to y .

a cash -> an act

Menurut Anda, paling sedikit, berapa banyak edit actions?

Menurut Damerau, 80% spelling error di Information Retrieval system adalah pada edit distance 1.

Dynamic Programming

Distance("xxazzzb", "xxbuua") =

minimum

Distance("xxazzz", "xxbuu") + 1

Distance("xxazzzb", "xxbuu") + 1

Distance("xxazzz", "xxbuua") + 1

Distance("xx", "xx") + 1 + 3 + 2

Dynamic Programming

Distance("xxazzzb", "xxbuua") =

Cost untuk
Substitution b
dengan a

minimum

Distance("xxazzz", "xxbuu") + 1

Distance("xxazzzb", "xxbuu") + 1

Distance("xxazzz", "xxbuua") + 1

Distance("xx", "xx") + 1 + 3 + 2

Dynamic Programming

Distance("xxazzzb", "xxbuua") =

minimum

Distance("xxazzz", "xxbuua") + 1

Distance("xxazzzb", "xxbuu") + 1

Distance("xxazzz", "xxbuua") + 1

Distance("xx", "xx") + 1 + 3 + 2

Cost untuk
Insertion a ke
string pertama

Dynamic Programming

Distance("xxazzzb", "xxbuua") =

minimum

Distance("xxazzz", "xxbuu") + 1

Distance("xxazzzb", "xxbuu") + 1

Distance("xxazzz", "xxbuua") + 1

Distance("xx", "xx") + 1 + 3 + 2

Cost untuk
Deletion b pada
string pertama

Dynamic Programming

Distance("xxazzzzb", "xxbuua") =

Cost untuk menyisipkan 3 karakter di antara a dan b di string pertama

Distance("xxaz

Cost untuk transposisi a dan b di string pertama

, "xxbuu") + 1

minimum

Distance("xxazzz", "xxbuua") + 1

Banyaknya karakter antara b dan a di string kedua

Distance("xx", "xx") + 1 + 3 + 2

Assumption for Transposition Cost

Hanya ada dua cara:

- Transposisi huruf dan sisipkan karakter diantara mereka
- Hapus semua karakter diantara dua karakter yang mau di-transposisi; baru lakukan transposisi

$ab \rightarrow bxxa$

$ab \rightarrow ba \rightarrow bxa \rightarrow bxxa$

$aSb \rightarrow bTa$

$axxb \rightarrow ba$

$axxb \rightarrow axb \rightarrow ab \rightarrow ba$

$Cost = 1 + |S| + |T|$

$axxb \rightarrow bya$

$axxb \rightarrow axb \rightarrow ab \rightarrow ba \rightarrow bya$

Dynamic Programming

Distance("xxxab", "xxxcb") =

Cost untuk
Substitution b
dengan b

minimum

Distance("xxxa", "xxxc") + 0

Distance("xxxab", "xxxc") + 1

Distance("xxxa", "xxxcb") + 1

Distance(-, -) + 1 + 4 + 4

Wagner-Fischer Table

			A	N		A	C	T
	12	12	12	12	12	12	12	12
	12	0	1	2	3	4	5	6
A	12	1	0	1	2	3	4	5
	12	2	1	1	1	2	3	4
C	12	3	2	2	2	2	2	3
A	12	4	3	3	3	2	2	3
S	12	5	4	4	4	3	3	3
H	12	6	5	5	5	4	4	4

12 adalah Panjang(AN ACT) + Panjang
(A CASH) = 6 + 6 = 12

Apa maksudnya?

Jawaban: edit distance

Wagner-Fischer Table

			A	N		A	C	T
	12	12	12	12	12	12	12	12
	12	0	1	2	3	4	5	6
A	12	1	0	1	2	3	4	5
	12	2	1	1	1	2	3	?
C	12	3						
A	12	4						
S	12	5						
H	12	6						

$\text{Dist}(\text{"A "}, \text{"AN AC"}) = 3$

$\text{Dist}(\text{"A"}, \text{"AN AC"}) = 4$

$\text{Dist}(\text{"A"}, \text{"AN ACT"}) = 5$

Berapakah $\text{Dist}(\text{"A "}, \text{"AN ACT"})$?

Siapa yang paling minimal total cost-nya diantara 4 jenis aksi?

Misal, baru terisi Sebagian, dan yang akan diisi berikutnya adalah baris 2 kolom 6

Wagner-Fischer Table

			A	N		A	C	T
	12	12	12	12	12	12	12	12
	12	0	1	2	3	4	5	6
A	12	1	0	1	2	3	4	5
	12	2	1	1	1	2	3	?
C	12	3						
A	12	4						
S	12	5						
H	12	6						

Jika substitution:

$$= \text{Dist}(\text{"A"}, \text{"AN AC"}) + 1$$

$$= 4 + 1 = 5$$

4 adalah total cost edit ("A", "AN AC") setelah memilih opsi substitution

Wagner-Fischer Table

			A	N		A	C	T
	12	12	12	12	12	12	12	12
	12	0	1	2	3	4	5	6
A	12	1	0	1	2	3	4	5
	12	2	1	1	1	2	3	?
C	12	3						
A	12	4						
S	12	5						
H	12	6						

Jika insertion:

$$= \text{Dist}(\text{"A "}, \text{"AN AC"}) + 1$$

$$= 3 + 1 = 4$$

3 adalah total cost Edit("A ", " AN AC") setelah memilih opsi insert

Wagner-Fischer Table

			A	N		A	C	T
	12	12	12	12	12	12	12	12
	12	0	1	2	3	4	5	6
A	12	1	0	1	2	3	4	5
	12	2	1	1	1	2	3	?
C	12	3						
A	12	4						
S	12	5						
H	12	6						

Jika deletion:

= $\text{Dist}(\text{"A"}, \text{"AN ACT"})$

= $5 + 1 = 6$

5 adalah total cost $\text{Edit}(\text{"A"}, \text{"AN ACT"})$ setelah memilih opsi deletion

Wagner-Fischer Table

			A	N		A	C	T
	12	12	12	12	12	12	12	12
	12	0	1	2	3	4	5	6
A	12	1	0	1	2	3	4	5
	12	2	1	1	1	2	3	?
C	12	3						
A	12	4						
S	12	5						
H	12	6						

Jika transposition:

$$\begin{aligned}
 &= \text{Dist}(-, \text{"AN"}) + 1 + 1 + 2 \\
 &= 12 + 1 + (2 - 0 - 1) + (6 - 3 - 1) = 15
 \end{aligned}$$

12 adalah total cost Edit(-, "AN") setelah memilih opsi transposition

Wagner-Fischer Table

			A	N		A	C	T
	12	12	12	12	12	12	12	12
	12	0	1	2	3	4	5	6
A	12	1	0	1	2	3	4	5
	12	2	1	1	1	2	3	4
C	12	3						
A	12	4						
S	12	5						
H	12	6						

Cost insertion yang totalnya menjadi paling kecil

$$= \text{Dist}(\text{"A "}, \text{"AN AC"}) + 1$$

$$= 3 + 1 = 4$$

Wagner-Fischer Table

			A	N		A	C	T
	12	12	12	12	12	12	12	12
	12	0	1	2	3	4	5	6
A	12	1	0	1	2	3	4	5
	12	2	1	1	1	2	3	4
C	12	3	2	2	2	2	2	3
A	12	4	3	3	3	2	?	
S	12	5						
H	12	6						

Misal, baru terisi Sebagian, dan yang akan diisi berikutnya adalah baris 4 kolom 5

Wagner-Fischer Table

			A	N		A	C	T
	12	12	12	12	12	12	12	12
	12	0	1	2	3	4	5	6
A	12	1	0	1	2	3	4	5
	12	2	1	1	1	2	3	4
C	12	3	2	2	2	2	2	3
A	12	4	3	3	3	2	?	
S	12	5						
H	12	6						

Jika substitution:

$$2 + 1 = 3$$

Jika insertion:

$$2 + 1 = 3$$

Jika deletion:

$$2 + 1 = 3$$

Wagner-Fischer Table

			A	N		A	C	T
	12	12	12	12	12	12	12	12
	12	0	1	2	3	4	5	6
A	12	1	0	1	2	3	4	5
	12	2	1	1	1	2	3	4
C	12	3	2	2	2	2	2	3
A	12	4	3	3	3	2	?	
S	12	5						
H	12	6						

Jika substitution:

$$2 + 1 = 3$$

Jika insertion:

$$2 + 1 = 3$$

Jika deletion:

$$2 + 1 = 3$$

Jika transposisi:

$$1 + 1 + (4-3-1) + (5-4-1) = 2$$

Wagner-Fischer Table

			A	N		A	C	T
	12	12	12	12	12	12	12	12
	12	0	1	2	3	4	5	6
A	12	1	0	1	2	3	4	5
	12	2	1	1	1	2	3	4
C	12	3	2	2	2	2	2	3
A	12	4	3	3	3	2	2	
S	12	5						
H	12	6						

Jika substitution:

$$2 + 1 = 3$$

Jika insertion:

$$2 + 1 = 3$$

Jika deletion:

$$2 + 1 = 3$$

Jika transposisi:

$$1 + 1 + (4-3-1) + (5-4-1) = 2$$

Cost paling kecil



Wagner-Fischer Table

			A	N		A	C	T
	12	12	12	12	12	12	12	12
	12	0	1	2	3	4	5	6
A	12	1	0	1	2	3	4	5
	12	2	1	1	?			
C	12	3						
A	12	4						
S	12	5						
H	12	6						

Misal, baru terisi Sebagian, dan yang akan diisi berikutnya adalah baris 2 kolom 3

Wagner-Fischer Table

			A	N		A	C	T
	12	12	12	12	12	12	12	12
	12	0	1	2	3	4	5	6
A	12	1	0	1	2	3	4	5
	12	2	1	1	?			
C	12	3						
A	12	4						
S	12	5						
H	12	6						

Jika substitution:

$$1 + 0 = 1$$

Jika insertion:

$$1 + 1 = 2$$

Jika deletion:

$$2 + 1 = 3$$

Jika transposisi:

$$12 + 1 + (2 - 0 - 1) + (3 - 0 - 1) = 16$$

Mengapa 0?
Bukan 1?

Wagner-Fischer Table

			A	N		A	C	T
	12	12	12	12	12	12	12	12
	12	0	1	2	3	4	5	6
A	12	1	0	1	2	3	4	5
	12	2	1	1	1			
C	12	3						
A	12	4						
S	12	5						
H	12	6						

Jika substitution:

$$1 + 0 = 1$$

Jika insertion:

$$1 + 1 = 2$$

Jika deletion:

$$2 + 1 = 3$$

Jika transposisi:

$$12 + 1 + (2 - 0 - 1) + (3 - 0 - 1) = 16$$

	-1	0	1	2	3	4	5	6
			A	N		A	C	T
-1								
0								
1	A							
2								
3	C							
4	A							
5	S							
6	H							

Solusi yang terinspirasi Lowrance-Wagner Algorithm untuk String-to-String Correction

		-1	0	1	2	3	4	5	6
				A	N		A	C	T
-1		12	12	12	12	12	12	12	12
0		12	0	1	2	3	4	5	6
1	A	12	1						
2		12	2						
3	C	12	3						
4	A	12	4						
5	S	12	5						
6	H	12	6						

```
DL-dist(a[1 .. len(a)],
        b[1 .. len(b)]):
```

```
// wagner-fischer matrix
```

```
// ukuran: (len(a) + 2) x (len(b) + 2)
```

```
init dist = array[-1 .. len(a), -1 .. len(b)]
```

```
maxdist = len(a) + len(b)
```

```
dist[-1, -1] = maxdist
```

```
for i = 0 to len(a):
```

```
    dist[i, -1] = maxdist
```

```
    dist[i, 0] = i
```

```
for j = 0 to len(b):
```

```
    dist[-1, j] = maxdist
```

```
    dist[0, j] = j
```

Continued ...

Solusi yang terinspirasi Lowrance-Wagner Algorithm untuk String-to-String Correction

	-1	0	1	2	3	4	5	6
			A	N		A	C	T
-1		12	12	12	12	12	12	12
0		12	0	1	2	3	4	5
1	A	12	1	0	1	2	3	4
2		12	2	1	1	1	2	3
3	C	12	3	2	2	2	2	3
4	A	12	4	3	3	3	2	3
5	S	12	5	4	4	4	3	3
6	H	12	6	5	5	5	4	4

```
init lastrow = {} // a map or dictionary
```

```
for i = 1 to len(a):
    lastcol = 0
    for j = 1 to len(b):
        lmr = lastrow[b[j]] // return 0 if not found
        lmc = lastcol
        if a[i] == b[j]:
            cost = 0
            lastcol = j
        else:
            cost = 1

        dist[i, j]
            = min( dist[i-1, j-1] + cost, //substitution
                  dist[i, j-1] + 1,      //insert
                  dist[i-1, j] + 1,      //delete

                  dist[lmr - 1, lmc - 1] + 1
                  + (i - lmr - 1) //transposition
                  + (j - lmc - 1) )

    lastrow[a[i]] = i

return dist[len(a), len(b)]
```

Menyimpan informasi posisi di string a terakhir (last row) yang match dengan karakter di string b yang sedang diinspeksi

6

T

-1		12	12	12	12	12	12	12	12
0		12	0	1	2	3	4	5	6
1	A	12	1	0					
2		12	2	1					
3	C	12	3	2	2	2	2	2	3
4	A	12	4	3	3	3	2	2	3
5	S	12	5	4	4	4	3	3	3
6	H	12	6	5	5	5	4	4	4

lmr: last match row
lmc: last match column

```
init lastrow = {} // a map or dictionary
```

```
for i = 1 to len(a):
    lastcol = 0
    for j = 1 to len(b):
        lmr = lastrow[b[j]] // return 0 if not found
        lmc = lastcol
        if a[i] == b[j]:
            cost = 0
            lastcol = j
        else:
            cost = 1

        dist[i, j]
            = min( dist[i-1, j-1] + cost, //substitution
                  dist[i, j-1] + 1,      //insert
                  dist[i-1, j] + 1,      //delete

                  dist[lmr - 1, lmc - 1] + 1
                  + (i - lmr - 1) //transposition
                  + (j - lmc - 1) )

        lastrow[a[i]] = i

return dist[len(a), len(b)]
```

Latihan

- Buat tabel Wagner-Fischer untuk DL-distance("BKAO", "KACO")

How to generate candidates?

- Periksa ke setiap kata di kamus baku, lalu pilih kata-kata dengan DL edit distance $< k$. Misal $k = 2$.
- Ada solusi yang lebih cepat dengan **Levenshtein automaton**, yaitu $O(N)$ dengan N adalah panjang string input.
 - <http://blog.notdot.net/2010/07/Damn-Cool-Algorithms-Levenshtein-Automata>

Jika ada yang mau coba eksplorasi dan implementasikan kode pada blog di atas, akan diberikan nilai **400** untuk week 4.

Finding The Best Candidate

Noisy Channel Model = Bayes' Rule

Yang diprediksi merupakan kata yang benar

$$\hat{w} = \operatorname{argmax}_{w \in V} P(w|x)$$

x : observation, kata yang salah eja

$$= \operatorname{argmax}_{w \in V} \frac{P(x|w)P(w)}{P(x)}$$

$$\propto \operatorname{argmax}_{w \in V} P(x|w)P(w)$$

Prior Probability

Sebuah "Likelihood":
Seberapa mungkin kata w "rusak" menjadi kata x karena melewati noisy channel?

Noisy Channel Model

Prior Probability $P(w)$?

- Merupakan "language model", atau lebih tepatnya "word model"
- Seberapa besar kemungkinan kita observasi w di corpus yang sangat besar.
- Dengan Maximum Likelihood Estimation (MLE), dapat diestimasi dengan

Estimasi dari $P(w)$

$$\hat{P}(w) = \frac{C(w)}{T}$$

Berapa kali kata w muncul di koleksi yang besar

Banyaknya token di sebuah koleksi yang besar

Prior Probability $P(w)$?

Koleksi

- D1: hujan sejuk pagi hari
- D2: udara sejuk dan pagi penuh semangat
- D3: tiada hujan tanpa kebaikan

$$\hat{P}(hujan) = \dots$$

Prior Probability $P(w)$

404.253.213 kata pada *Corpus of Contemporary English (COCA)*

word	word frequency $C(w)$	Estimated $P(w)$
actress	9.231	.0000230573
cress	220	.0000005442
caress	686	.0000016969
access	37.038	.0000916207
across	120.844	.0002989314
acres	12.847	.0000318463

Noisy Channel Model

Salah satunya adalah dengan **Edit Probability**
(Kernighan, Church & Gale, 1990)

Edit Probability - koreksi hanya 1 step dari 4 kemungkinan:
insertion, deletion, substitution, transposition


$$P(x|w) = P(x_1, x_2, x_3, \dots, x_n | w_1, w_2, w_3, \dots, w_n)$$

Untaian karakter dari kata
salah eja x

Untaian karakter dari kata
yang benar

Single Step Correction

Aksi yang menyebabkan salah eja.



Typo	Correction	Transformation			
acress	actress	@	t	2	deletion
acress	cress	a	#	0	insertion
acress	caress	ac	ca	0	reversal
acress	access	r	c	2	substitution
acress	across	e	o	3	substitution
acress	acres	s	#	4	insertion
acress	acres	s	#	5	insertion

@ and # represents nulls in the typo and correction, respectively.

Noisy Channel Model

Aksi yang menyebabkan salah eja.



Jika deletion

$$\frac{del[w_{i-1}, w_i]}{count[w_{i-1}, w_i]}$$

Jika insertion

$$\frac{ins[w_{i-1}, x_i]}{count[w_{i-1}]}$$

Jika substitution

$$\frac{sub[w_i, x_i]}{count[w_i]}$$

Jika transposition

$$\frac{trans[w_i, w_{i+1}]}{count[w_i, w_{i+1}]}$$

$$\hat{P}(x|w) =$$

del[x,y]: berapa kali karakter berurutan **xy** (pada kata benar) diketik sebagai **x** pada training dataset.

ins[x,y]: berapa kali karakter **x** diketik sebagai **xy**

sub[x,y]: berapa kali karakter **x** diketik sebagai **y**

trans[x,y]: berapa kali karakter **xy** diketik sebagai **yx**

count[x,y]: berapa kali karakter berurutan **xy** muncul di training dataset

count[x]: berapa kali karakter **x** muncul di training dataset

Kernighan's Confusion Matrix

X	Substitution of X (incorrect) for Y (correct)																									
	Y (correct)																									
	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z
a	0	0	7	1	342	0	0	2	118	0	1	0	0	3	76	0	0	1	35	9	9	0	1	0	5	0
b	0	0	9	9	2	2	3	1	0	0	0	5	11	5	0	10	0	0	2	1	0	0	8	0	0	0
c	6	5	0	16	0	9	5	0	0	0	1	0	7	9	1	10	2	5	39	40	1	3	7	1	1	0
d	1	10	13	0	12	0	5	5	0	0	2	3	7	3	0	1	0	43	30	22	0	0	4	0	2	0
e	388	0	3	11	0	2	2	0	89	0	0	3	0	5	93	0	0	14	12	6	15	0	1	0	18	0
f	0	15	0	3	1	0	5	2	0	0	0	3	4	1	0	0	0	6	4	12	0	0	2	0	0	0
g	4	1	11	11	9	2	0	0	0	1	1	3	0	0	2	1	3	5	13	21	0	0	1	0	3	0
h	1	8	0	3	0	0	0	0	0	0	2	0	12	14	2	3	0	3	1	11	0	0	2	0	0	0
i	103	0	0	0	146	0	1	0	0	0	0	6	0	0	49	0	0	0	2	1	47	0	2	1	15	0
j	0	1	1	9	0	0	1	0	0	0	0	2	1	0	0	0	0	0	5	0	0	0	0	0	0	0
k	1	2	8	4	1	1	2	5	0	0	0	0	5	0	2	0	0	0	6	0	0	0	4	0	0	3
l	2	10	1	4	0	4	5	6	13	0	1	0	0	14	2	5	0	11	10	2	0	0	0	0	0	0
m	1	3	7	8	0	2	0	6	0	0	4	4	0	180	0	6	0	0	9	15	13	3	2	2	3	0
n	2	7	6	5	3	0	1	19	1	0	4	35	78	0	0	7	0	28	5	7	0	0	1	2	0	2
o	91	1	1	3	116	0	0	0	25	0	2	0	0	0	0	14	0	2	4	14	39	0	0	0	18	0
p	0	11	1	2	0	6	5	0	2	9	0	2	7	6	15	0	0	1	3	6	0	4	1	0	0	0
q	0	0	1	0	0	0	27	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
r	0	14	0	30	12	2	2	8	2	0	5	8	4	20	1	14	0	0	12	22	4	0	0	1	0	0
s	11	8	27	33	35	4	0	1	0	1	0	27	0	6	1	7	0	14	0	15	0	0	5	3	20	1
t	3	4	9	42	7	5	19	5	0	1	0	14	9	5	5	6	0	11	37	0	0	2	19	0	7	6
u	20	0	0	0	44	0	0	0	64	0	0	0	0	2	43	0	0	4	0	0	0	0	2	0	8	0
v	0	0	7	0	0	3	0	0	0	0	0	1	0	0	1	0	0	0	8	3	0	0	0	0	0	0
w	2	2	1	0	1	0	0	2	0	0	1	0	0	0	0	7	0	6	3	3	1	0	0	0	0	0
x	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	9	0	0	0	0	0	0	0
y	0	0	2	0	15	0	1	7	15	0	0	0	2	0	6	1	0	7	36	8	5	0	0	1	0	0
z	0	0	0	7	0	0	0	0	0	0	0	7	5	0	0	0	0	2	21	3	0	0	0	0	3	0

Kernighan's Confusion Matrix

X	Insertion of Y after X Y (Inserted Letter)																									
	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z
a	15	1	14	7	10	0	1	1	33	1	4	31	2	39	12	4	3	28	134	7	28	0	1	1	4	1
b	3	11	0	0	7	0	1	0	50	0	0	15	0	1	1	0	0	5	16	0	0	3	0	0	0	0
c	19	0	54	1	13	0	0	18	50	0	3	1	1	1	7	1	0	7	25	7	8	4	0	1	0	0
d	18	0	3	17	14	2	0	0	9	0	0	6	1	9	13	0	0	6	119	0	0	0	0	0	5	0
e	39	2	8	76	147	2	0	1	4	0	3	4	6	27	5	1	0	83	417	6	4	1	10	2	8	0
f	1	0	0	0	2	27	1	0	12	0	0	10	0	0	0	0	0	5	23	0	1	0	0	0	1	0
g	8	0	0	0	5	1	5	12	8	0	0	2	0	1	1	0	1	5	69	2	3	0	1	0	0	0
h	4	1	0	1	24	0	10	18	17	2	0	1	0	1	4	0	0	16	24	22	1	0	5	0	3	0
i	10	3	13	13	25	0	1	1	69	2	1	17	11	33	27	1	0	9	30	29	11	0	0	1	0	1
j	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
k	2	4	0	1	9	0	0	1	1	0	1	1	0	0	2	1	0	0	95	0	1	0	0	0	4	0
l	3	1	0	1	38	0	0	0	79	0	2	128	1	0	7	0	0	0	97	7	3	1	0	0	2	0
m	11	1	1	0	17	0	0	1	6	0	1	0	102	44	7	2	0	0	47	1	2	0	1	0	0	0
n	15	5	7	13	52	4	17	0	34	0	1	1	26	99	12	0	0	2	156	53	1	1	0	0	1	0
o	14	1	1	3	7	2	1	0	28	1	0	6	3	13	64	30	0	16	59	4	19	1	0	0	1	1
p	23	0	1	1	10	0	0	20	3	0	0	2	0	0	26	70	0	29	52	9	1	1	1	0	0	0
q	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
r	15	2	1	0	89	1	1	2	64	0	0	5	9	7	10	0	0	132	273	29	7	0	1	0	10	0
s	13	1	7	20	41	0	1	50	101	0	2	2	10	7	3	1	0	1	205	49	7	0	1	0	7	0
t	39	0	0	3	65	1	10	24	59	1	0	6	3	1	23	1	0	54	264	183	11	0	5	0	6	0
u	15	0	3	0	9	0	0	1	24	1	1	3	3	9	1	3	0	49	19	27	26	0	0	2	3	0
v	0	2	0	0	36	0	0	0	10	0	0	1	0	1	0	1	0	0	0	0	1	5	1	0	0	0
w	0	0	0	1	10	0	0	1	1	0	1	1	0	2	0	0	1	1	8	0	2	0	4	0	0	0
x	0	0	18	0	1	0	0	6	1	0	0	0	1	0	3	0	0	0	2	0	0	0	0	1	0	0
y	5	1	2	0	3	0	0	0	2	0	0	1	1	6	0	0	0	1	33	1	13	0	1	0	2	0
z	2	0	0	0	5	1	0	0	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	4
@	46	8	9	8	26	11	14	3	5	1	17	5	6	2	2	10	0	6	23	2	11	1	2	1	1	2

Wikipedia

https://en.wikipedia.org/wiki/Wikipedia:Lists_of_common_misspellings/C

- [carmel](#) (caramel, Carmel-by-the-Sea)
- [carniverous](#) (carnivorous)
- [carraige](#) ([carriage](#))
- [carrear](#) (career)
- [carred](#) (cared, carried)
- [carreer](#) (career)
- [carrer](#) (career)
- [Carribbean](#) (Caribbean)
- [Carribean](#) (Caribbean)

- [carring](#) (caring, carrying)
- [carryng](#) (carrying)
- [Carthagian](#) (Carthaginian)
- [carthographer](#) (cartographer)
- [cartilege](#) (cartilage)
- [cartilidge](#) (cartilage)
- [cartrige](#) (cartridge)
- [casette](#) (cassette)
- [casion](#) (caisson)
- [cassawory](#) (cassowary)

Peter Norvig's Single-Edit Corrections List

https://norvig.com/ngrams/count_1edit.txt

e i	917	er re	189	B b	9
a e	856	i is	133	A E	9
i e	771	u o	130	-	9
e a	749	h he	129	y ya	8
a i	559	s se	128	x s	8
t te	478	o or	127	w e	8
r re	392	u a	126	wo ow	8
s c	383	y i	125	cr c	6
e ea	354	a u	123	ag a	6
a o	353	is i	122	I It	6
o a	352	ei ie	122	>p >	6
a al	352	al a	122	e	6
i a	313	el le	121	z x	5
re r	299	s st	120	z c	5
e o	295	u ur	119	y t	5
				yl ly	5

Smoothing

Bagaimana jika ada **unseen errors**? Bisa menghasilkan nilai probabilitas 0. Ini tidak kita harapkan karena terlalu berlebihan.

Contoh: di Kernighan's substitution confusion matrix, substitusi dari **q** ke **a** dan **a** ke **q** bernilai 0. Padahal huruf **q** dan **a** bertetangga di keyboard.

$$\hat{P}(w) = \frac{C(w) + 0.5}{T}$$

$$\hat{P}(x|w) = \frac{sub[w_i, x_i] + 1}{count[w_i] + |A|}$$



Banyaknya alphabet

acress?

Candidate Correction	Correct Letter	Error Letter	x/w	$P(x/w)$	$P(w)$	$10^9 * P(x/w)P(w)$
actress	t	-	c c t	.000117	.0000231	2.7
cress	-	a	a #	.00000144	.000000544	.00078
caress	ca	ac	ac ca	.00000164	.00000170	.0028
access	c	r	r c	.000000209	.0000916	.019
across	o	e	e o	.0000093	.000299	2.8
acres	-	s	es e	.0000321	.0000318	1.0
acres	-	s	ss	.0000342	.0000318	1.0 ⁴³

Gambar diambil tanpa malu dari slide Chris Manning & Pandu Nayak, IR & Web Search, Stanford U.