Distance Between Words for spellchecking

Distance between strings

STUDENTS-HUB.com

Errors are close to correct words

Need a measure of distance between strings (words/phrases):

- Jackard coefficient/similarity (done)
- Edit distance: effort to convert one into another
- Soundex: phonetic closeness (sound: تشابه النطق)
- More likely to mistake with adjacent keyboard characters, or without language switch: Really ثشممغ really in Arabic keyboard!
- Confusion letters:{ء ئ ؤ أ آ إ}, {ى ي}, {ة ه} less distance within
- **Final distance: A weighted sum!**

STUDENTS-HUB.com

How similar are two strings?

- Spell correction
  - The user typed “graffe”

  Which is closest?
    - graf
    - graft
    - grail
    - giraffe

- Computational Biology
  - Align two sequences of nucleotides

```
AGGCTATCACCTGACCTCCAGGCCGATGCCC
TAGCTATCACGACCGCGGTCGATTTGCCCGAC
```

  - Resulting alignment:

```
-AGGCTATCACCTGACCTCCAGGCCGA--TGCCC---
TAG-CTATCAC--GACCGC--GGTCGATTTGCCCGAC
```

- Also for Machine Translation, Information Extraction, Speech Recognition

STUDENTS-HUB.com

Edit Distance

- The minimum edit distance between two strings
- Is the minimum number of editing operations
  - Insertion
  - Deletion
  - Substitution
- Needed to transform one into the other

STUDENTS-HUB.com

Minimum Edit Distance

- Two strings and their **alignment**:

```
I N T E * N T I O N
| | | | | | | | | |
* E X E C U T I O N
```

STUDENTS-HUB.com

Minimum Edit Distance

```
I N T E * N T I O N
| | | | | | | | | |
* E X E C U T I O N
d s s   i s
```

- If each operation has cost of 1
  - Distance between these is 5
- If substitutions cost 2 (Levenshtein)
  - Distance between them is 8

STUDENTS-HUB.com

Alignment in Computational Biology

- Given a sequence of bases

  AGGCTATCACCTGACCTCCAGGCCGATGCCC
  TAGCTATCACGACCGCGGTCGATTTGCCCGAC

- An alignment:

  –AGG**CTATCAC**C**TGACC**T**C**C**A**GG**C**CGA**––**TGCCC**–––
  TT**AG**–**CTATCAC**––**GACC**G**C**––**GG**T**CGA**T**T**T**GCCC**G**AC**

- Given two sequences, align each letter to a letter or gap

STUDENTS-HUB.com

Other uses of Edit Distance in NLP

- Evaluating Machine Translation and speech recognition

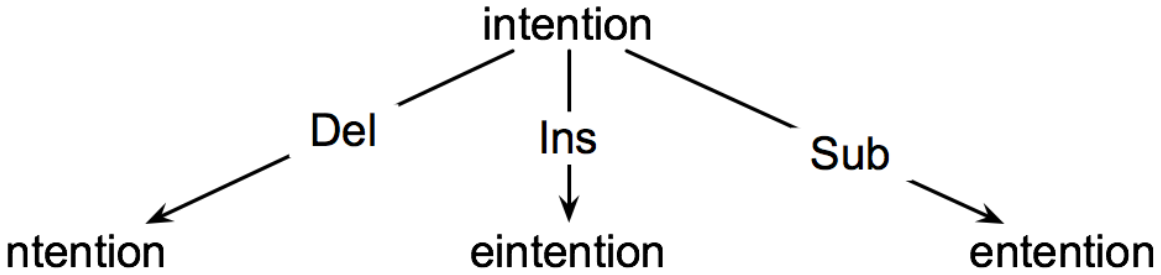
```
R Spokesman confirms    senior government adviser was shot
H Spokesman said    the senior            adviser was shot dead
              S      I               D                      I
```

- Named Entity Extraction and Entity Coreference
  - IBM Inc. announced today
  - IBM profits
  - Stanford President John Hennessy announced yesterday
  - for Stanford University President John Hennessy

STUDENTS-HUB.com

How to find the Min Edit Distance?

- Searching for a path (sequence of edits) from the start string to the final string:
  - **Initial state**: the word we're transforming
  - **Operators**: insert, delete, substitute [Levenshtein] exchange[Demerau]
  - **Goal state**: the word we're trying to get to
  - **Path cost**: what we want to minimize: the number of edits



STUDENTS-HUB.com

Minimum Edit as Search

- But the space of all edit sequences is huge!
  - We can't afford to navigate naïvely
  - Lots of distinct paths wind up at the same state.
    - We don't have to keep track of all of them
    - Just the shortest path to each of those revisted states.

STUDENTS-HUB.com

Defining Min Edit Distance

- For two strings
  - X of length *n*
  - Y of length *m*
- We define D(*i*,*j*)
  - the edit distance between X[1..*i*] and Y[1..*j*]
    - i.e., the first *i* characters of X and the first *j* characters of Y
  - The edit distance between X and Y is thus D(*n*,*m*)

STUDENTS-HUB.com

Minimum Edit Distance

Definition of Minimum Edit Distance

STUDENTS-HUB.com

Minimum Edit Distance

Computing Minimum Edit Distance

STUDENTS-HUB.com

Dynamic Programming for Minimum Edit Distance

- **Dynamic programming**: A tabular computation of D(*n*,*m*)
- Solving problems by combining solutions to subproblems.
- Bottom-up
  - We compute D(i,j) for small *i*,*j*
  - And compute larger D(i,j) based on previously computed smaller values
  - i.e., compute D(*i*,*j*) for all *i* (0 < *i* < n) and *j* (0 < j < m)

STUDENTS-HUB.com

Defining Min Edit Distance (Levenshtein)

- Initialization

  $D(i,0) = i$

  $D(0,j) = j$

- Recurrence Relation:

For each $i = 1 \ldots M$

  For each $j = 1 \ldots N$

$$D(i,j) = \min \begin{cases} D(i-1,j) + 1 \\ D(i,j-1) + 1 \\ D(i-1,j-1) + \begin{cases} 2; & \text{if } X(i) \neq Y(j) \\ 0; & \text{if } X(i) = Y(j) \end{cases} \end{cases}$$

- Termination:

  $D(N,M)$ is distance

STUDENTS-HUB.com

The Edit Distance Table

N	9									
O	8									
I	7									
T	6									
N	5									
E	4									
T	3									
N	2									
I	1									
#	**0**	1	2	3	4	5	6	7	8	9
	#	E	X	E	C	U	T	I	O	N

STUDENTS-HUB.com

The Edit Distance Table

N	9									
O	8									
I	7									
T	6									
N	5									
E	4									
T	3									
N	2									
I	1									
#	**0**	1	2	3	4	5	6	7	8	9
	#	E	X	E	C	U	T	I	O	N

$$D(i,j) = \min \begin{cases} D(i-1,j) + 1 \\ D(i,j-1) + 1 \\ D(i-1,j-1) + \begin{cases} 2; & \text{if } S_1(i) \neq S_2(j) \\ 0; & \text{if } S_1(i) = S_2(j) \end{cases} \end{cases}$$

STUDENTS-HUB.com

Edit Distance

$$D(i,j) = \min \begin{cases} D(i-1,j) + 1 \\ D(i,j-1) + 1 \\ D(i-1,j-1) + \begin{cases} 2; & \text{if } S_1(i) \neq S_2(j) \\ 0; & \text{if } S_1(i) = S_2(j) \end{cases} \end{cases}$$

N	9									
O	8									
I	7									
T	6									
N	5									
E	4									
T	3									
N	2									
I	1									
#	0	1	2	3	4	5	6	7	8	9
	#	E	X	E	C	U	T	I	O	N

STUDENTS-HUB.com

The Edit Distance Table

N	9	8	9	10	11	12	11	10	9	**8**
O	8	7	8	9	10	11	10	9	8	9
I	7	6	7	8	9	10	9	8	9	10
T	6	5	6	7	8	9	8	9	10	11
N	5	4	5	6	7	8	9	10	11	10
E	4	3	4	5	6	7	8	9	10	9
T	3	4	5	6	7	8	7	8	9	8
N	2	3	4	5	6	7	8	7	8	7
I	1	2	3	4	5	6	7	6	7	8
#	**0**	1	2	3	4	5	6	7	8	9
	#	E	X	E	C	U	T	I	O	N

STUDENTS-HUB.com

Minimum Edit Distance

Weighted Minimum Edit Distance

STUDENTS-HUB.com

Levenshtein vs Demerau

- Levenshtein: three ops: add, delete, substitute.
- Demerau: Add, delete, substitute, exchange:
  - Resaerch → research 1 Demerau, 2 or more Levenshtein

STUDENTS-HUB.com

Weighted Edit Distance

- Why would we add weights to the computation?
  - Spell Correction: some letters are more likely to be mistyped than others
  - Biology: certain kinds of deletions or insertions are more likely than others

STUDENTS-HUB.com

Confusion matrix for spelling errors

sub[X, Y] = Substitution of X (incorrect) for Y (correct)

X	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

STUDENTS-HUB.com

Weighted Min Edit Distance

- Initialization:

$$D(0,0) = 0$$
$$D(i,0) = D(i-1,0) + del[x(i)]; \quad 1 < i \leq N$$
$$D(0,j) = D(0,j-1) + ins[y(j)]; \quad 1 < j \leq M$$

- Recurrence Relation:

$$D(i,j) = \min \begin{cases} D(i-1,j) + del[x(i)] \\ D(i,j-1) + ins[y(j)] \\ D(i-1,j-1) + sub[x(i),y(j)] \end{cases}$$

- Termination:

$D(N,M)$ is distance

STUDENTS-HUB.com

Keyboarding Errors: more for mobiles

STUDENTS-HUB.com

Where did the name, dynamic programming, come from?

...The 1950s were not good years for mathematical research. [the] Secretary of Defense …had a pathological fear and hatred of the word, research…

I decided therefore to use the word, "**programming**".

I wanted to get across the idea that this was dynamic, this was multistage… I thought, let's … take a word that has an absolutely precise meaning, namely **dynamic**… it's impossible to use the word, **dynamic**, in a pejorative sense. Try thinking of some combination that will possibly give it a pejorative meaning. It's impossible.

Thus, I thought dynamic programming was a good name. It was something not even a Congressman could object to."

STUDENTS-HUB.com Richard Bellman, "Eye of the Hurricane: an autobiography" 1984.

Soundex Algorithm

Idea

- –Vowels are viewed as interchangeable in transcribing names
- Consonants with similar sounds (e.g., D and T) are put in equivalence classes
- related names often have the same soundex codes

Algorithm

- –Turn every term to be indexed into a four-character reduced form,
- build an inverted index from these reduced forms to the original terms called the soundex index
- Do the same with the query terms
- Search the soundex index

STUDENTS-HUB.com

Soundex Algorithm: Four-Character Code

- The first character is a letter of the alphbet and the other three are digits between 0 and 9

- Algorithm

–Retain the first letter of the term

–Change all occurrences of the following letters to ‘0’

A, E, I, O U, H, W, and Y

–Change letters to digits as follows

B, F, P, V →1

STUDENTS-HUB.com

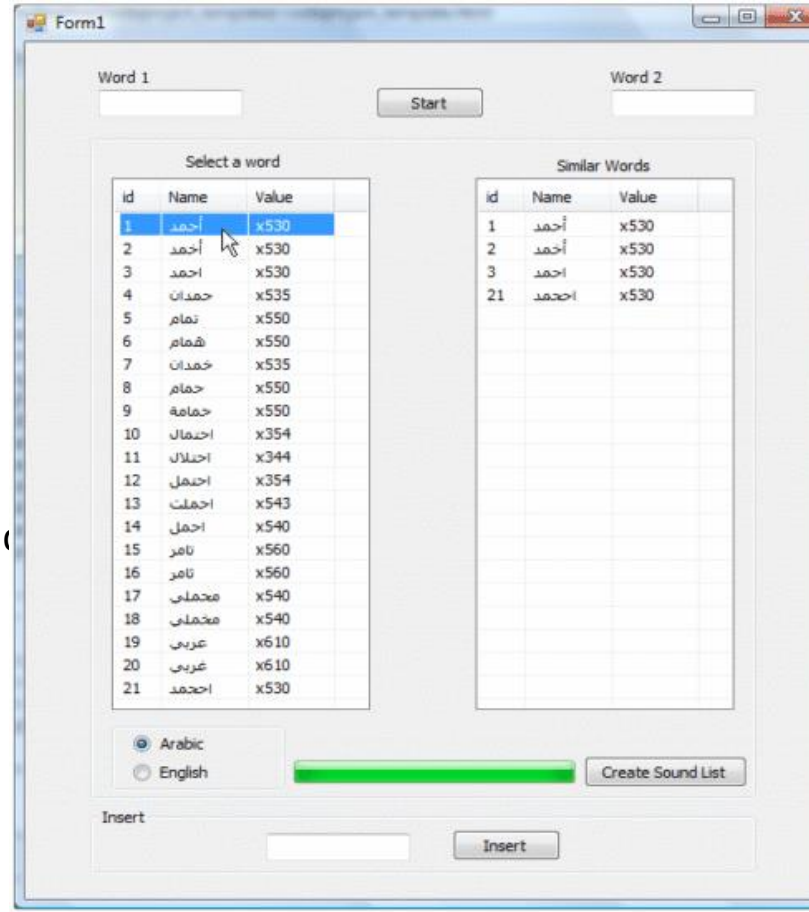
Soundex Algorithm: Four-Character Code

- •D, T → 3
- •L →4
- •M, N → 5
- •R → 6
- –Repeatedly remove one out of each pair of consecutive identical digits
- –Remove all 0's from the resulting string, pad the resulting string with trailing zeros and return the first four positions: a letter followed by three digits
- •**Example:** Hermann →H655, Herman →H655, matched!

STUDENTS-HUB.com

Arabic Soundex : Four-Character Code

- Hold the first letter.
- Replace the characters (ا, أ, إ, آ, ح, ع, غ, ش,و,ي) with the value 0.
- Replace the characters (ف, ب) with the value 1.
- Replace the characters (خ, ج, ز, س, ص, ظ, ق, ك) with the value 2.
- Replace the characters (ت, ث,د,ذ,ض,ط) with the value 3.
- Replace the character (ل) with the value 4.
- Replace the characters (م, ن) with the value 5.
- Replace the character (ر)with the value 6.
- https://www.codeproject.com/Articles/26880/Arabic-Soun



Soundex Code	المحارف	المجموعة
1	b, f, p, v	شَفَهِيّ
2	c, g, j, k, q, s, x, z	حروفٌ حَلْقِيّة و حرْوفُ صَفير
3	d, t	حرف نطعي (ملفوظ بوضع اللسان على مؤخر الاسنان الامامية العليا)
4	l	حرف صامت ملفوظ بلطف طويل
5	m, n	حرف يلفظ من الانف
6	r	حرف صامت ملفوظ بلطف قصير

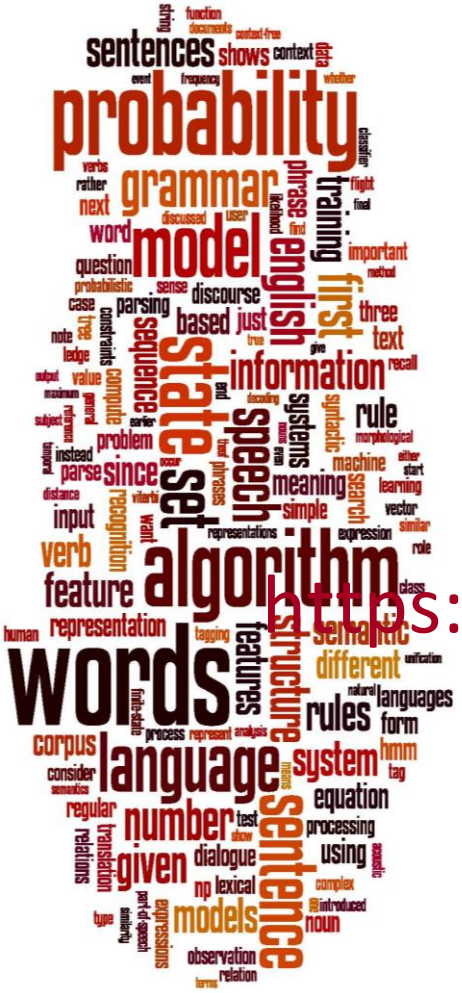
STUDENTS-HUB.com

Errors are close to correct words

Distance between error and correct words: a combination of:

- Jackard coefficient/similarity (distance)
- Edit distance (Levenshtein, Demerau): *not all equal*
- Soundex: phonetic closeness (sound: تشابه النطق): *less distance here*
- More likely to mistake with adjacent keyboard characters, or without language switch: Really ثشممغ really in Arabic keyboard!
- Confusion letters:{ء ئ ؤ أ آ إ}, {ى ي}, {ة ه} less distance within
- **Final distance: A weighted sum!**

STUDENTS-HUB.com



Minimum Edit Distance

https://www.youtube.com/watch?v=Xxx0b7djCrs

STUDENTS-HUB.com