Understanding string distances

INTERMEDIATE REGULAR EXPRESSIONS IN R

Angelo Zehr
Data Journalist

What is a string distance?

0	r a i n
1	r a n
2	r u n

- insertion
- deletion
- substitution

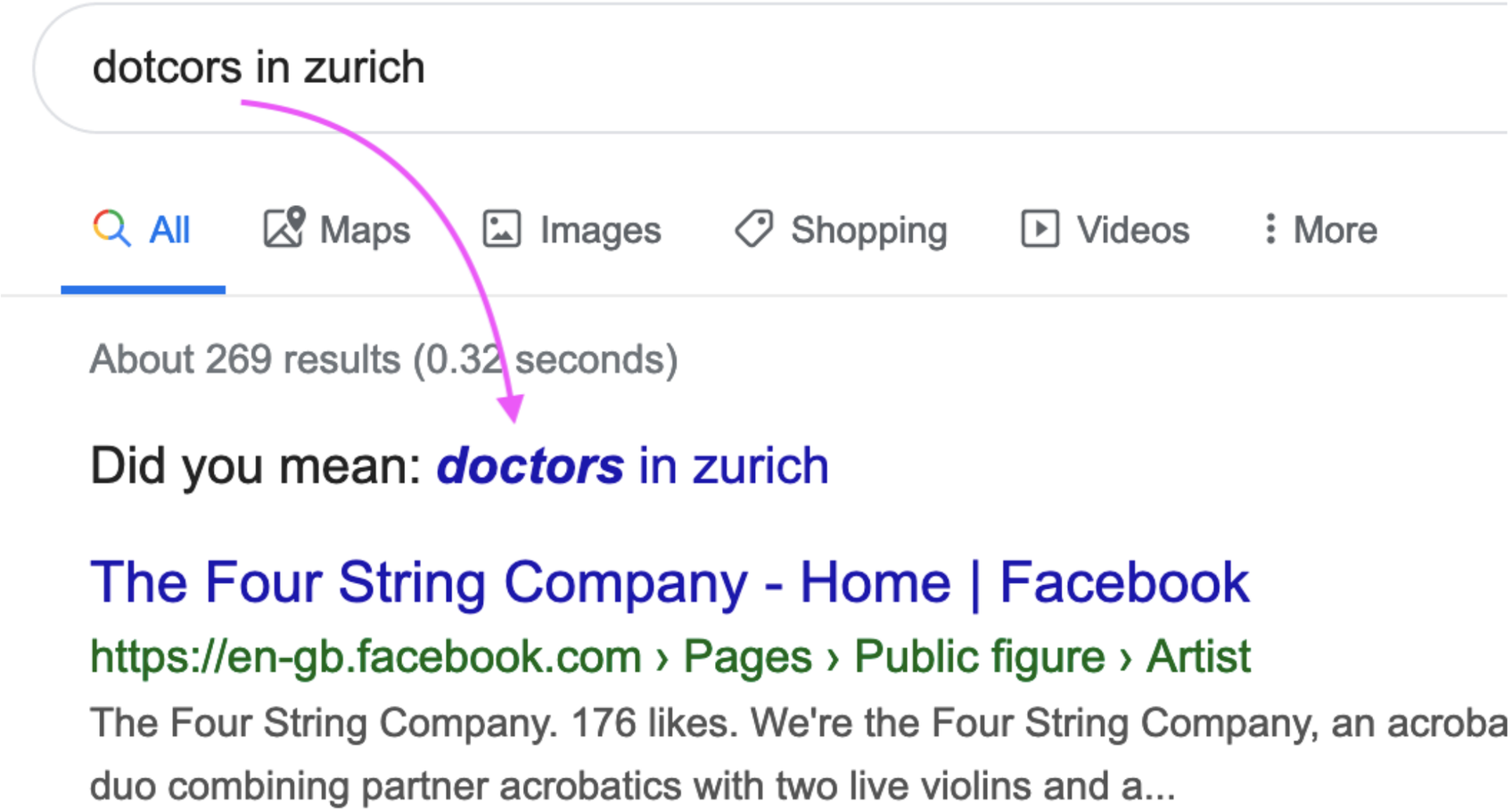
What is a string distance?

0	s u n d a y
1	s u r d a y
2	s a u r d a y
3	s a t u r d a y

- insertion
- deletion
- substitution

Real world applications

```
0   Emile Brown
1   Emilie Brown
```


dotcors in zurich
All
Maps
Images
Shopping
Videos
More
About 269 results (0.32 seconds)
Did you mean: doctors in zurich
The Four String Company - Home | Facebook
https://en-gb.facebook.com › Pages › Public figure › Artist
The Four String Company. 176 likes. We're the Four String Company, an acroba
duo combining partner acrobatics with two live violins and a...

String distances in R

```
library(stringdist)
stringdist("saturday", "sunday", method = "lv")
```

Returns:

```
3
```

Is identical:

```
stringdist("sunday", "saturday", method = "lv")
```

Finding a match

```
amatch(
  x = "Sonday",
  table = c("Friday", "Saturday", "Sunday"),
  maxDist = 1,
  method = "lv"
)
```

Returns:

```
3
```

Let's practice!

INTERMEDIATE REGULAR EXPRESSIONS IN R

Methods of string distances

INTERMEDIATE REGULAR EXPRESSIONS IN R

Angelo Zehr
Data Journalist

Damerau-Levenshtein

Regular Levenshtein distance

0	Rcik Caplan
1	Rik Caplan
2	Rick Caplan

Damerau-Levenshtein distance

0	Rcik Caplan
1	Rick Caplan

transposition

Method abbreviations

Regular Levenshtein distance:

```
stringdist(a, b, method = "lv")
```

Damerau-Levenshtein distance:

```
stringdist(a, b, method = "dl")
```

Optimal String Alignment distance:

```
stringdist(a, b, method = "osa")
```

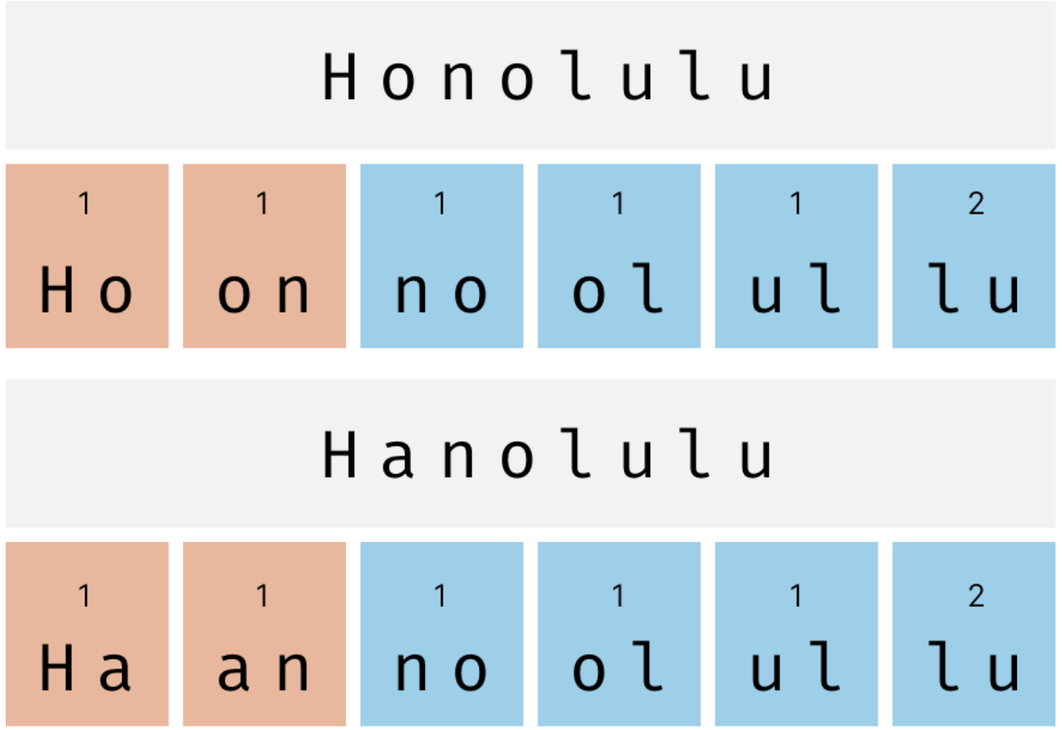
Q-Grams (or n-grams)

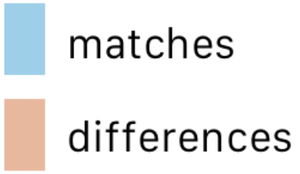
Q-Grams `(q = 2)`

Honolulu					
1	1	1	1	1	2
Ho	on	no	ol	ul	lu

Q-Grams (or n-grams)

Q-Grams `(q = 2)`





Inspecting q-grams

```
qgrams("Honolulu", "Hanolulu", q = 2)
```

Returns:

```
   Ho on ul no ol lu la
V1  1  1  1  1  1  2  0
V2  1  1  1  1  1  1  1
```

Method abbreviations

Sum of qgrams that are not shared

```
stringdist(a, b, method = "qgram") # equals 4
```

Not shared qgrams divided by total number of qgrams

```
stringdist(a, b, method = "jaccard") # equals 0.5
```

Optimal String Alignment distance

```
stringdist(a, b, method = "cosine") # equals 0.22
```

Let's practice!

INTERMEDIATE REGULAR EXPRESSIONS IN R

Fuzzy joins

INTERMEDIATE REGULAR EXPRESSIONS IN R

Angelo Zehr
Instructor

A regular join

Table a

user_name	user_id
Bryan	1
Barbara	2
Tom	3

Table b

user_id	email
1	bryan@example.com
3	tom@example.com
2	barbara@example.com

Joined

user_name	user_id	email
Bryan	1	bryan@example.com
Barbara	2	barbara@example.com
Tom	3	tom@example.com

A fuzzy join

Table a

user_input
Brian
Barbra
Thom

Table b

name	email
Bryan	bryan@example.com
Tom	tom@example.com
Barbara	barbara@example.com

Joined

user_input	name	email
Brian	Bryan	bryan@example.com
Thom	Tom	tom@example.com
Barbra	Barbara	barbara@example.com

The fuzzyjoin package

```r
library(fuzzyjoin)
```

```r
stringdist_join(
  user_input,
  database,
  by = c("user_input" = "name"),
  method = "lv",
  max_dist = 1,
  distance_col = "distance"
)
```

stringdist_join: Result

user_input	name	email	distance
Brian	Bryan	bryan@example.com	1
Thom	Tom	tom@example.com	1
Barbra	Barbara	barbara@example.com	1

Let's practice!

INTERMEDIATE REGULAR EXPRESSIONS IN R

Custom Fuzzy Matching

INTERMEDIATE REGULAR EXPRESSIONS IN R

Angelo Zehr
Data Journalist

Combining two fuzzy matches

Table a

title	year
Star Wars: Episode III	2005
The Children Of Men	2006
The Pursuit Of Happyness	2007
Twilight Saga, The: Breaking Dawn, Teil 1	2011
Wild Tales - Relatos Salvajes	2015
X-Men 3	2006

Table b

prod_title	prod_year	external_id
Star Wars 3 – Episode III	2004	Q42051
Children of Men	2006	Q221090
The Prusuit of Happiness	2005	Q19608917
The Pursuit Of Happyness	2015	Q220515
The Twilight Saga: Breaking Dawn	2010	Q60506
Wild Tales	2015	Q16672466
X-Men 3	2006	Q221168
X-Men 2	2001	Q12578

Combining two fuzzy matches

Table a

title	year
Star Wars: Episode III	2005
The Children Of Men	2006
The Pursuit Of Happyness	2007
Twilight Saga, The: Breaking Dawn, Teil 1	2011
Wild Tales - Relatos Salvajes	2015
X-Men 3	2006

Table b

prod_title	prod_year	external_id
Star Wars 3 – Episode III	2004	Q42051
Children of Men	2006	Q221090
The Prusuit of Happiness	2005	Q19608917
The Pursuit Of Happyness	2015	Q220515
The Twilight Saga: Breaking Dawn	2010	Q60506
Wild Tales	2015	Q16672466
X-Men 3	2006	Q221168
X-Men 2	2001	Q12578

Fuzzy matches: Helper functions

For the string comparison:

```
small_str_distance <- function(left, right) {
  stringdist(left, right) <= 5
}
```

For the number comparison:

```
close_to_each_other <- function(left, right) {
  abs(left - right) <= 3
}
```

The fuzzy join

```
fuzzy_left_join(
  a, b,
  by = c(
    "title" = "prod_title",
    "year" = "prod_year"
  ),
  match_fun = c(
    "title" = small_str_distance,
    "year" = close_to_each_other
  )
)
```

The fuzzy join: The result

Joined

title	year	prod_title	prod_year	external_id
Star Wars: Episode III	2005	Star Wars 3 – Episode III	2004	Q42051
The Children Of Men	2006	Children of Men	2006	Q221090
The Pursuit Of Happyness	2007	The Prusuit of Happiness	2005	Q19608917
Twilight Saga, The: Breaking Dawn, Teil 1	2011			
Wild Tales - Relatos Salvajes	2015			
X-Men 3	2006	X-Men 3	2006	Q221168

Let's practice!

INTERMEDIATE REGULAR EXPRESSIONS IN R

Congratulations

INTERMEDIATE REGULAR EXPRESSIONS IN R

Angelo Zehr
Data Journalist

A look back

1. Regular Expressions: Writing custom patterns

- `str_view()` , `str_match()` , `str_detect()` …

2. Creating strings with data

- `glue()` , `glue_collapse()` ,…

3. Extracting structured data from text

- `str_extract_all()` , `extract()` ,…

4. Similarities between strings

- `strindist()` , `amatch()` , `stringdist_join()`

Next courses

Thank you!

INTERMEDIATE REGULAR EXPRESSIONS IN R