



Approximate Matching

Mark van der Loo and Edwin de Jonge

Statistics Netherlands Research & Development
@markvdloo @edwindjonge

useR!2019



Try the code together with your neighbour

```
02input/raw_to_input.R
```



String distance

Default (Optimal String Alignment distance)

Count number of character deletions, insertions, substitutions and transpositions (of adjacent characters)

```
library(stringdist)
stringdist("Ross Ihaka", "Robert Gentleman")

## [1] 12
```



Exact Matching with match

```
lookup <- c("Alice","Bob","Carol","Danny")
raw      <- c("Bob","Carl","Rob","bob","Dan","Alice")
i <- match(raw, lookup)
data.frame(raw=raw, matched=lookup[i])
```

```
##      raw matched
## 1   Bob      Bob
## 2  Carl    <NA>
## 3   Rob    <NA>
## 4   bob    <NA>
## 5   Dan    <NA>
## 6 Alice  Alice
```



Approximate Matching with `stringdist::amatch`

```
library(stringdist)
j <- amatch(raw, lookup, maxDist=2)
data.frame(raw=raw, matched=lookup[i], amatched=lookup[j])
```

##		raw	matched	amatched
##	1	Bob	Bob	Bob
##	2	Carl	<NA>	Carol
##	3	Rob	<NA>	Bob
##	4	bob	<NA>	Bob
##	5	Dan	<NA>	Danny
##	6	Alice	Alice	Alice

→ Match with closest match, and distance ≤ 2 .



Optimal string alignment?

```
stringdist("Robert Gentleman", "Gentleman, Robert")
```

```
## [1] 15
```

```
stringdist("Robert Gentleman", "Ross Ihaka")
```

```
## [1] 12
```

→ OSA will give a false match (if we allow maxDist of 12)



Alternative: cosine distance

```
stringdist("Robert Gentleman", "Gentleman, Robert",  
           , method="cosine", q=2)
```

```
## [1] 0.1608536
```

```
stringdist("Robert Gentleman", "Ross Ihaka",  
           , method="cosine", q=2)
```

```
## [1] 0.9139337
```

Notes

- Based on counting co-occurrence of character q -grams (here: pairs).
- Always between 0 and 1



More on amatch

```
amatch(x, table, method, maxDist,...)
```

x	character data to be matched
table	the lookup table with clean values
method	string distance type
maxDist	Maximum distance allowed (depends on "method"!)
...	Extra options depending on "method"

Example

```
amatch(raw, lookup, method="cosine", maxDist=0.5, q=3)
```



Assignment

Merge data from the `companies` dataset with data from `backbone.csv`.

- Using approximate matching on the "name" and "company" column.
- Think about and try different distance functions and `maxDist`
- Keep your best solution
- Remove rows that cannot be matched
- Write to `02input/myinput.csv`

