

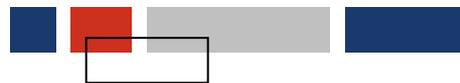


INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL

» Use of R to increase the Automatic Coding Rate in Census2021



21-05-2019



Rui Alves
Almiro Moreira



INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL



About me



- Sociology degree since 1993
- Working at Statistics Portugal (Data Collection and Management Department) since 2013
- Using with R since 2015
- Statistics Portugal Data Science Team since 2019



Overview



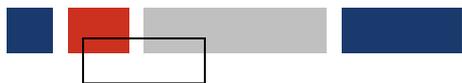
- Create and expand existing dictionaries
- Monitor performance and validate results
- Some results
- Next steps



Challenges



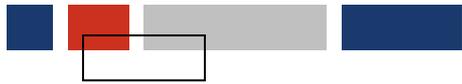
- **Economic Activity** and **Occupation** are two of the most commonly collected variables in household surveys
- Approximately 100.000 responses / per year
- Census 2011: about 50% (five million) manually coded: **250.000 working hours**
- Census 2021: **collect all data with CAWI**
- Text strings: very demanding computation



CREATE AND EXPAND EXISTING DICTIONARIES



Starting Point

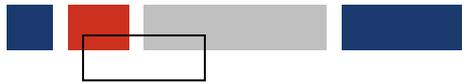


- Previous experience (Census 2011)
- Available data (Household Surveys 2011-...)
- An old idea (2012)
- Paradigm shift
- Coders should be part of the process
- Reproduce manual coding process



Household Surveys

2011-2019



- LFS - Labour Force Survey
- HFCS - Household Finance and Consumption Survey
- SILC - Income and Living Conditions Survey
- AES - Adult Education Survey
- TSR - Travel Survey of Residents
- ICT - Information and Communication Technologies



Household Surveys 2011-2019

```
glimpse(df_household_survs)
```

```
## Observations: 919,309
```

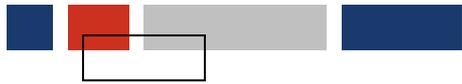
```
## Variables: 2
```

```
## $ cpp_dsg <chr> "pintor automoveis", "agente imobiliario", "operadora l...
```

```
## $ cpp_cod3 <chr> "713", "333", "522", "815", "812", "432", "751", "422",...
```



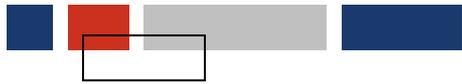
Household Surveys 2011-2019



```
# Household Survey global dataframe  
  
df_household_survs %>%  
  head() %>%  
  knitr::kable()
```

cpp_dsg	cpp_cod3
pintor automoveis	713
agente imobiliario	333
operadora loja faz tudo repoe material atende clientres	522
costureira textil	815
industrial fundicao operador maquinas fundicao	812
fiel armazem	432

Frequency rules



Eligible for dictionary test:

- A string with more than one code
 - $n \geq 10$ & $\geq 90\%$ code consistency
- A string with same code
 - $n \geq 5$ & $\geq 100\%$ code consistency



```
# my_id
df_household_survs$my_id <-
  str_c(df_household_survs$cpp_cod3,
        df_household_survs$cpp_dsg,
        sep = "_")

# count_my_id
count_my_id <- df_household_survs %>%
  count(my_id) %>%
  arrange(desc(n)) %>%
  rename(n_my_id = n)

# n_cpp_dsg
count_cpp_dsg <- df_household_survs %>%
  count(cpp_dsg) %>%
  arrange(desc(n)) %>%
  rename(n_cpp_dsg = n)

# joins n_cpp_dsg & n_my_id to df_household_survs
df_household_survs <- left_join(df_household_survs, count_cpp_dsg, by = "cpp_dsg") %>%
  left_join(count_my_id, by = "my_id")

# prob
df_household_survs <- df_household_survs %>%
  mutate(prob = n_my_id * 100 / n_cpp_dsg) %>%
  distinct(my_id, .keep_all = T) %>%
  arrange(cpp_dsg)
```

Frequency rules

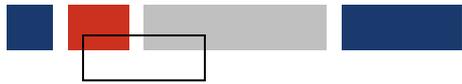
```
# create dic rule n>=10 & prob <=90%

rule_10_90 <-
  df_household_survs %>%
  distinct(my_id, .keep_all = T) %>%
  select(cpp_dsg, cpp_cod3, my_id:prob) %>%
  filter(n_cpp_dsg >= 10, prob >= 89.5) %>%
  arrange(desc(n_cpp_dsg  ))
```

```
bind_rows(head(rule_10_90, n = 5),
          tail(rule_10_90, n = 5)) %>%
  knitr::kable()
```

cpp_dsg	cpp_cod3	my_id	n_cpp_dsg	n_my_id	prob
empregada domestica	911	911_empregada domestica	16112	16100	99.92552
empregada limpeza	911	911_empregada limpeza	15229	15129	99.34336
pedreiro	711	711_pedreiro	11732	11656	99.35220
cozinheira	512	512_cozinheira	8084	8078	99.92578
enfermeira	222	222_enfermeira	6065	6012	99.12613
vendedora numa loja	522	522_vendedora numa loja	10	10	100.00000
vigilancia criancas	531	531_vigilancia criancas	10	10	100.00000
vigilante portaria	541	541_vigilante portaria	10	9	90.00000
vigilante praia	541	541_vigilante praia	10	10	100.00000
vigilante recepcionista	541	541_vigilante recepcionista	10	9	90.00000

Frequency rules



The same word can be written in a multitude of variations:

- spelling errors
- (mis)use of abbreviation
- typo's





String Distance

Levenshtein Distance

The **Optimal String Alignment distance** (method='osa') is like the Levenshtein distance but also allows **transposition of adjacent characters**.

Each substring may be edited only once.

Very demanding in terms of computation

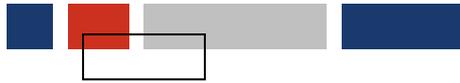
$$d_W(s, t) = \begin{cases} 0 & \text{if } s = t = \varepsilon \\ \min\{ & \\ & d_W(s, t_{1:|t|-1}) + w_1, \\ & d_W(s_{1:|s|-1}, t) + w_2, \\ & d_W(s_{1:|s|-1}, t_{1:|t|-1}) + [1 - \delta(s_{|s|}, t_{|t|})]w_3 \\ & \} & \text{otherwise.} \end{cases}$$

- Substitution: 'foo' → 'boo'
- Deletion: 'foo' → 'oo'
- Insertion: 'foo' → 'floo'
- **Transposition**: 'foo' → 'ofu'

stringdist R package by M.P.J. van der Loo (2014)



Levenshtein Distance



$$d_V(s, t) = \begin{cases} 0 & \text{if } s = t = \epsilon \\ \min\{ & \\ & d_V(s, t_{1:|t|-1}) + w_1, \\ & d_V(s_{1:|s|-1}, t) + w_2, \\ & d_V(s_{1:|s|-1}, t_{1:|t|-1}) + [1 - \delta(s_{|s|}, t_{|t|})]w_3 \\ & \} & \text{otherwise.} \end{cases}$$

- Stringdistance:
 - Soldado (*soldier*)
 - Soldador (*welder*)
- $d_V = 1!$



Vladimir Levenshtein



```
bind_rows(head(rule_10_90, n = 5),
          tail(rule_10_90, n = 5)) %>%
  knitr::kable()
```

cpp_dsg	cpp_cod3	my_id	n_cpp_dsg	n_my_id	prob
empregada domestica	911	911_empregada domestica	16112	16100	99.92552
empregada limpeza	911	911_empregada limpeza	15229	15129	99.34336
pedreiro	711	711_pedreiro	11732	11656	99.35220
cozinheira	512	512_cozinheira	8084	8078	99.92578
enfermeira	222	222_enfermeira	6065	6012	99.12613
vendedora numa loja	522	522_vendedora numa loja	10	10	100.00000
vigilancia criancas	531	531_vigilancia criancas	10	10	100.00000
vigilante portaria	541	541_vigilante portaria	10	9	90.00000
vigilante praia	541	541_vigilante praia	10	10	100.00000
vigilante recepcionista	541	541_vigilante recepcionista	10	9	90.00000

```

library(tidystringdist)

tidy_comb("enfermeira",
          df_household_survs$cpp_dsg[df_household_survs$cpp_cod3 == "222"]) %>%
  tidy_stringdist(method = "osa") %>%
  filter(osa <= 3) %>%
  arrange(osa) %>%
  distinct(V1) %>%
  knitr::kable()

```

enfermeira	enferemira	enfermeira	efermeiro	enfremeira
3nfermeira	enferimeira	3nfermeira	ehfermeiro	engermeira
efermeira	enfermaeira	efermeira	emfermeiro	infermeira
emfermeira	enfermaira	emfermeira	enefermeiro	nfermeira
emnfermeira	enfermeia	emnfermeira	enfemeiro	unfermeira
enefermeira	enfermeiar	enefermeira	enfemerira	e nfermeiro
enfeermeira	enfermeiira	enfeermeira	enfereima	efermeiro
enfeirmeira	enfermeir	enfeirmeira	enfereiro	ehfermeiro
enfemeira	enfermeiraq	enfemeira	enferemeiro	emfermeiro
enfemreira	enfermeiras	enfemreira	enferemiro	enefermeiro



Let's see a wordcloud

```
# create a df: only cpp code "222"
```

```
df_household_survs_cpp222 <- filter(df_household_survs, cpp_cod3 == "222")
```

```
# results in a wordcloud
```

```
library(tidystringdist)
```

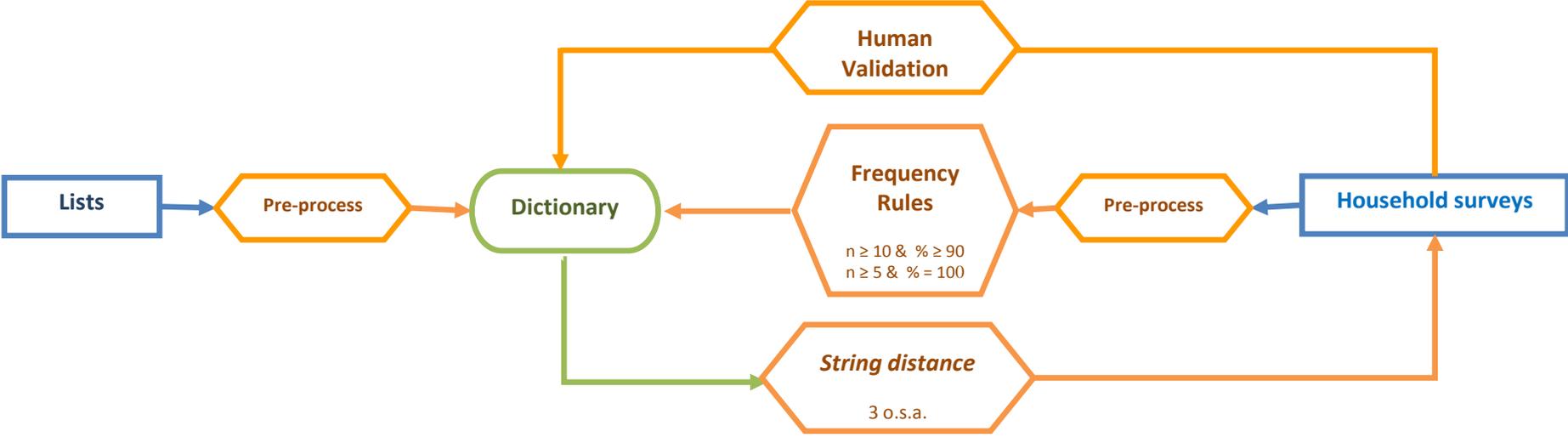
```
wc <- tidy_comb("enfermeira", df_household_survs_cpp222$cpp_dsg) %>%  
  tidy_stringdist(method = "osa") %>%  
  filter(osa <= 3)
```

```
wespal <- wesanderson::wes_palette(name = "Zissou1", 5, type = "continuous")  
wordcloud::wordcloud(wc$V1, wc$osa, scale = c(2, .1), min.freq = 1, colors = wespal)
```



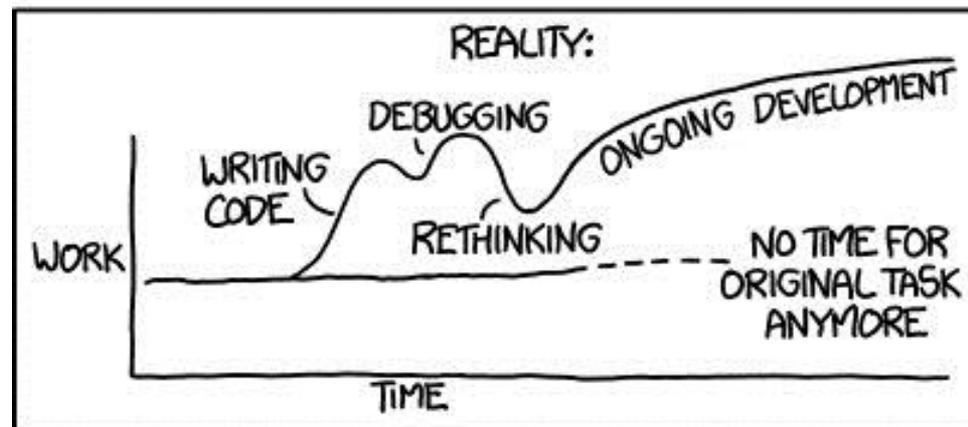
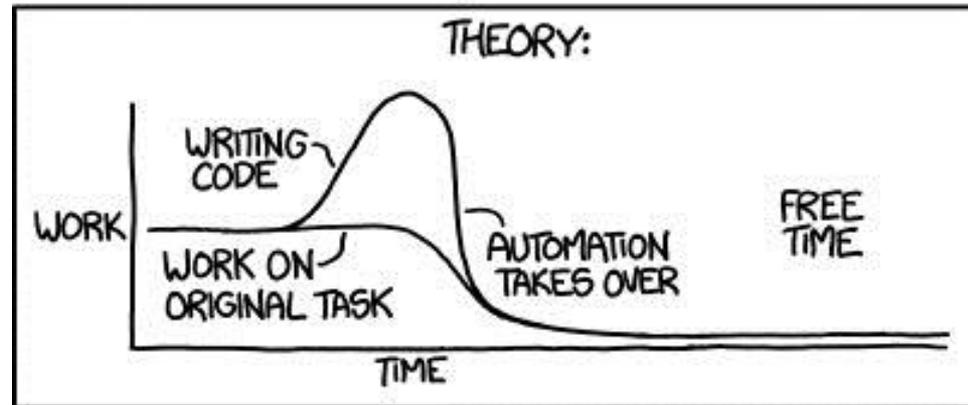


Create and expand dictionaries



Motivation

"I SPEND A LOT OF TIME ON THIS TASK.
I SHOULD WRITE A PROGRAM AUTOMATING IT!"



XKCD

Ongoing development...

```
library(purrr)
```

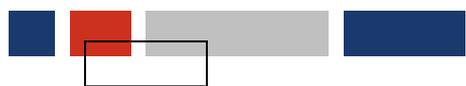
```
# dataframe to add and / or to check
```

```
ok_to_add_check <- data.frame(  
  cpp_dsg = c("1 escrituraria", "1 escritorario"),  
  correct_cod3 = c("431", "431"))
```

```
# search dic for strings within 3osa , compare cod
```

```
map2_df(list(ok_to_add_check$cpp_dsg), dic_cpp$cpp_dsg, tidy_comb) %>%  
tidy_stringdist(method = "osa") %>%  
filter(osa <= 3) %>%  
arrange(osa) %>%  
rename(cpp_dsg = V1) %>%  
left_join(dic_cpp) %>%  
left_join(ok_to_add_check, by = c("V2" = "cpp_dsg")) %>%  
filter(cpp_cod3 != correct_cod3) %>%  
head(n = 15) %>%  
knitr::kable(.)
```



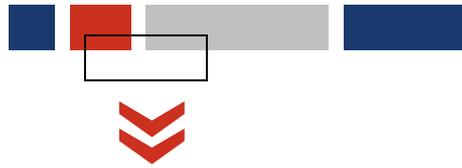


More data to analyze...

cpp_dsg	V2	osa	cpp_cod3	cpp_fonte3	correct_cod3
1 escrituraria	1 escrituraria	0	411	coders	431
1 escriturario	1 escriturario	0	411	coders	431
1 escrituraria	1 escriturario	1	411	coders	431
1 escriturario	1 escrituraria	1	411	coders	431
1o escriturario	1 escriturario	1	411	coders	431
2 escriturario	1 escriturario	1	411	coders	431
3 escrituraria	1 escrituraria	1	411	coders	431
1 escrituraria	1 escrituraria	1	411	coders	431
2 escrituraria	1 escrituraria	1	411	coders	431
3 escriturario	1 escriturario	1	411	coders	431



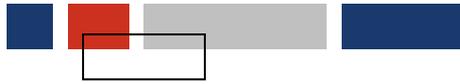
Algorithm



- Classify
 - Occupation and Economic Activity
 - Exact matching
 - 2 and 3 digits level
- Compare with manual coding (preliminary stage)
- Communicate results



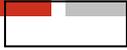
Automation and Integration



- R Package: classify, plot and summary report
- RMarkdown: classify, plot extended report, output files
- Integrated in our management system: full automation



Packages



- tidyverse
- janitor
- stringi
- tidystringdist
- knitr
- qdap
- readxl
- rio
- wesanderson
- wordcloud





MONITOR PERFORMANCE AND VALIDATE RESULTS

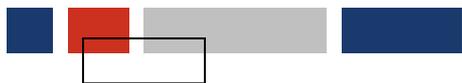


Monitor performance and validate results



- Plot and report
- Rmarkdown report
- Validation:
 - AES automatic classification
 - Dictionaries





Report & Plot

Report

=====

Summary Automatic Classification of Economic Activity

=====

Total cases: 25607

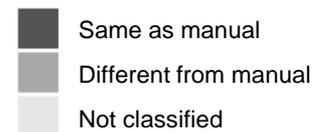
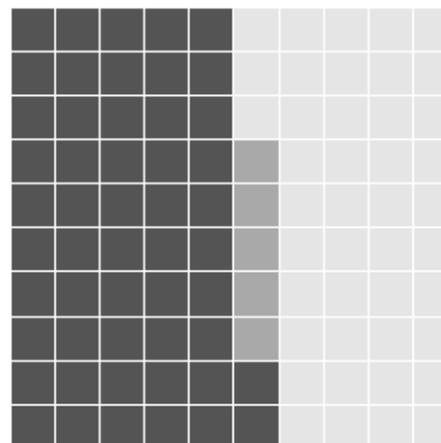
Classified cases: 14637 (57.16 %)

* 13333 (52.07 %) equal to manual classification

* 1304 (5.09 %) different from manual classification

Not classified: 10970 (42.84 %)

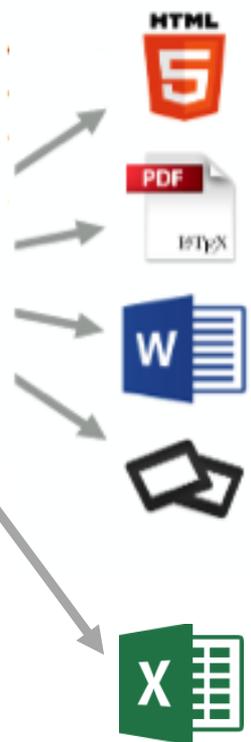
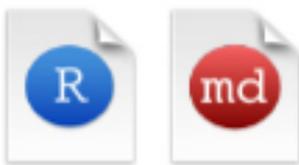
Plot



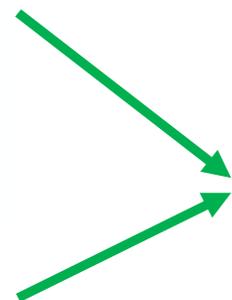
1 quadrado = 1% (aprox.)

RMarkdown Report

Labour Force Survey



Internal
Wikipage
Dashboard



Coders
(validation)

RMarkdown Report

Relatório de Classificação CAE e CPP IECAT

DRI | SPR
2017-07-17

Introdução

O presente relatório ¹ tem como objectivo apresentar um resumo das diferenças entre a classificação manual e a automática da CAE e da CPP no âmbito do IECAT (semanas 2016.SMN.01, 2016.SMN.02, 2016.SMN.03, 2016.SMN.04). Apenas são consideradas as situações em que se verificaram diferenças entre a classificação manual e classificação automática e em que CAE e CPP foram ambos codificados manualmente. Pretende ainda facilitar a interpretação dos ficheiros de resultados em formato XLSX (IECAT_CAE_compara_2017_07_17_16_57_35.xlsx e IECAT_CPP_compara_2017_07_17_16_57_35.xlsx).

A classificação automática é realizada com recurso a um algoritmo desenvolvido na linguagem R. Este algoritmo classifica descritivos CAE ou CPP a 2 ou 3 dígitos de acordo com dicionários de classificação (mais documentação sobre os dicionários pode ser consultada no final deste documento). O resultado da classificação é devolvido num formato tabular onde constam os dados originais acrescidos de novas variáveis:

- **cae_dsg** ou **cpp_dsg**: cópia normalizada do descritivo a classificar
- **cae_cod** ou **cpp_cod**: código a 2 ou 3 dígitos atribuído automaticamente
- **cae_fonte** ou **cpp_fonte**: origem da entrada do dicionário
- **caem3** ou **cppm3**: classificação manual simplificada a 3 dígitos
- **cae_igual** ou **cpp_igual**: classificação manual e automática são iguais? (TRUE/FALSE)

O algoritmo normaliza a variável a classificar (descritivo CAE ou CPP) através dos seguintes passos:

- Remove acentuação e aplica o **encoding** UTF-8
- Coloca o texto em letra minúscula
- Remove **pontuação**
- Remove **stopwords**
- Remove espaços em branco (duplicados e do início/fim do campo)

¹ Este relatório em formato DOC, os ficheiros de resultados em XLSX e todo o código subjacente foram produzidos em RMarkdown.

RMarkdown Report

Relatório de Classificação CAE

DRI | SPR
2017-07-17

Introdução

O presente relatório ¹ tem como objetivo apresentar um a classificação manual e a automática da CAE e da CPP no 2016.SMN.01, 2016.SMN.02, 2016.SMN.03, 2016.SMN.04 situações em que se verificaram diferenças entre a classificação automática e em que CAE e CPP foram ambos codificados ainda facilitar a interpretação dos ficheiros de resultados (IECAT_CAE_compara_2017_07_17_16_57_35.xlsx e IECAT_CPP_compara_2017_07_17_16_57_35.xlsx).

A classificação automática é realizada com recurso a um linguagem R. Este algoritmo classifica descritivos CAE ou acordo com dicionários de classificação (mais documentado pode ser consultada no final deste documento). O resultado devolvido num formato tabular onde constam os dados e variáveis:

- **cae_dsg** ou **cpp_dsg**: cópia normalizada do descritivo
- **cae_cod** ou **cpp_cod**: código a 2 ou 3 dígitos atribuído
- **cae_fonte** ou **cpp_fonte**: origem da entrada do dicionário
- **caem3** ou **cppm3**: classificação manual simplificada
- **cae_igual** ou **cpp_igual**: classificação manual e automática (TRUE/FALSE)

O algoritmo normaliza a variável a classificar (descritivo) seguintes passos:

- Remove acentuação e aplica o **encoding** UTF-8
- Coloca o texto em letra minúscula
- Remove **pontuação**
- Remove stopwords
- Remove espaços em branco (duplicados e do início)

¹ Este relatório em formato DOC, os ficheiros de resultado subjacente foram produzidos em RMarkdown.

Análise dos resultados do classificador automático

Classificação CAE

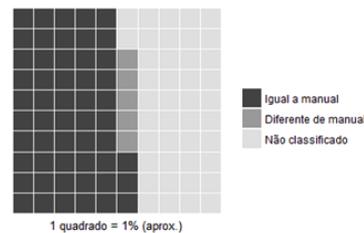
Os resultados detalhados da classificação automática podem ser consultados no ficheiro "IECAT_CAE_compara_2017_07_17_16_57_35.xlsx". De seguida apresenta-se um resumo.

Total de observações: 3751
Casos classificados: 2166 (57.74%):

- 1979 (52.76%) igual a classificação manual
- 187 (4.99%) diferente da classificação manual

Sem classificação: 1585 (42.26%)

Gráfico "waffle" da classificação CAE



Nas semanas de referência (2016.SMN.01, 2016.SMN.02, 2016.SMN.03, 2016.SMN.04) foram identificados 5 descritivos com as seguintes características: com classificação manual diferente da automática e com uma frequência igual ou superior a 4.

Na tabela em baixo podemos observar um resumo destes resultados.

|

RMarkdown Report

Relatório de Classificação CAE

DRI | SPR
2017-07-17

Introdução

O presente relatório ¹ tem como objetivo apresentar um a classificação manual e a automática da CAE e da CPP no 2016.SMN.01, 2016.SMN.02, 2016.SMN.03, 2016.SMN.04 situações em que se verificaram diferenças entre a classificação automática e em que CAE e CPP foram ambos codificados, ainda facilitar a interpretação dos ficheiros de resultado (IECAT_CAE_compara_2017_07_17_16_57_35.xlsx e IECAT_CPP_compara_2017_07_17_16_57_35.xlsx).

A classificação automática é realizada com recurso a um linguagem R. Este algoritmo classifica descritivos CAE ou acordo com dicionários de classificação (mais document pode ser consultada no final deste documento). O resultado devolvido num formato tabular onde constam os dados e variáveis:

- **cae_dsg** ou **cpp_dsg**: copia normalizada do descritivo
- **cae_cod** ou **cpp_cod**: código a 2 ou 3 dígitos atribuído
- **cae_fonte** ou **cpp_fonte**: origem da entrada do dicionário
- **caem3** ou **cppm3**: classificação manual simplificada
- **cae_igual** ou **cpp_igual**: classificação manual e automática (TRUE/FALSE)

O algoritmo normaliza a variável a classificar (descritivo) seguintes passos:

- Remove acentuação e aplica o **encoding** UTF-8
- Coloca o texto em letra minúscula
- Remove **pontuação**
- Remove stopwords
- Remove espaços em branco (duplicados e do início,

¹ Este relatório em formato DOC, os ficheiros de resultado subjacente foram produzidos em RMarkdown.

Análise dos resultados do classificador automático

Classificação CAE

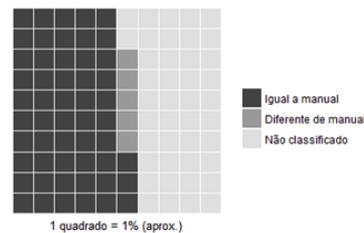
Os resultados detalhados da classificação automática podem ser consultados ficheiro "IECAT_CAE_compara_2017_07_17_16_57_35.xlsx". De seguida apresentamos um resumo.

Total de observações: 3751
Casos classificados: 2166 (57.74%):

- 1979 (52.76%) igual a classificação manual
- 187 (4.99%) diferente da classificação manual

Sem classificação: 1585 (42.26%)

Gráfico "waffle" da classificação CAE



Nas semanas de referência (2016.SMN.01, 2016.SMN.02, 2016.SMN.03, 2016.SMN.04) foram identificados 5 descritivos com as seguintes características: com classificação manual diferente da automática e com uma frequência igual ou superior a 4.

Na tabela em baixo podemos observar um resumo destes resultados.

Descritivos CAE com classificações manuais e automáticas distintas

cae_dsg	total	n_cods	n_reg
agropecuaria	14	1	1
agricultura subsistencia	5	1	2
clinica fisioterapia sem internamento	5	2	3
escola basica integrada	4	1	1
snack bar	4	1	2

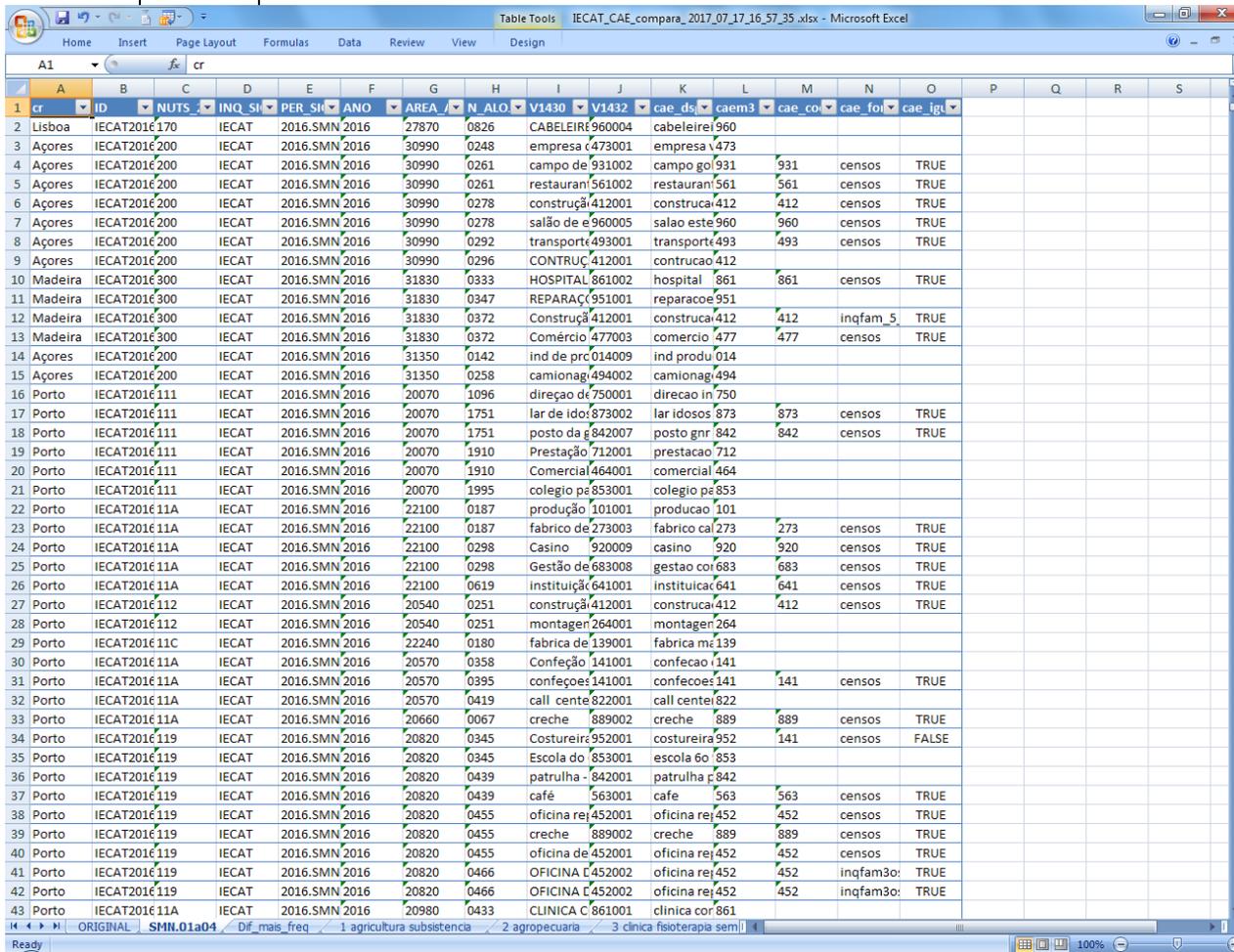
Legenda

- **cae_dsg**: descritivo normalizado
- **total**: número de vezes que o descritivo surge nas semanas em análise
- **n_cods**: número de diferentes códigos atribuídos manualmente ao mesmo descritivo
- **n_reg**: número de regiões que utilizaram código(s) diferente(s) do classificador automático

Esta tabela indica-nos, por exemplo, que o descritivo "agropecuaria" é utilizado 14 vezes nas semanas de referência, com 1 código(s) diferente(s) da classificação automática, sendo que esta situação ocorreu em 1 região(ões).

Outro exemplo, o descritivo "snack bar" é utilizado 4 vezes nas semanas de referência, com 1 código(s) diferente(s) da classificação automática, sendo que esta situação ocorreu em 2 região(ões).

RMarkdown Report



1	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	cr	ID	NUTS	INQ_S	PER_S	ANO	AREA	N_ALO	V1430	V1432	cae_ds	caem3	cae_co	cae_fo	cae_ig				
2	Lisboa	IECAT2016170	IECAT	2016.SMN	2016	27870	0826	CABELEIRE	960004	cabeleire	960								
3	Açores	IECAT2016200	IECAT	2016.SMN	2016	30990	0248	empresa	473001	empresa	473								
4	Açores	IECAT2016200	IECAT	2016.SMN	2016	30990	0261	campo de	931002	campo go	931	931	censo	TRUE					
5	Açores	IECAT2016200	IECAT	2016.SMN	2016	30990	0261	restauran	561002	restauran	561	561	censo	TRUE					
6	Açores	IECAT2016200	IECAT	2016.SMN	2016	30990	0278	construçã	412001	construca	412	412	censo	TRUE					
7	Açores	IECAT2016200	IECAT	2016.SMN	2016	30990	0278	salão de e	960005	salao este	960	960	censo	TRUE					
8	Açores	IECAT2016200	IECAT	2016.SMN	2016	30990	0292	transporte	493001	transporte	493	493	censo	TRUE					
9	Açores	IECAT2016200	IECAT	2016.SMN	2016	30990	0296	CONTRUC	412001	contrucao	412								
10	Madeira	IECAT2016300	IECAT	2016.SMN	2016	31830	0333	HOSPITAL	861002	hospital	861	861	censo	TRUE					
11	Madeira	IECAT2016300	IECAT	2016.SMN	2016	31830	0347	REPARAÇ	951001	reparacao	951								
12	Madeira	IECAT2016300	IECAT	2016.SMN	2016	31830	0372	Construçã	412001	construca	412	412	inqfam_5	TRUE					
13	Madeira	IECAT2016300	IECAT	2016.SMN	2016	31830	0372	Comércio	477003	comercio	477	477	censo	TRUE					
14	Açores	IECAT2016200	IECAT	2016.SMN	2016	31350	0142	ind de prc	014009	ind produ	014								
15	Açores	IECAT2016200	IECAT	2016.SMN	2016	31350	0258	camionag	494002	camionag	494								
16	Porto	IECAT2016111	IECAT	2016.SMN	2016	20070	1096	direção d	750001	direcao in	750								
17	Porto	IECAT2016111	IECAT	2016.SMN	2016	20070	1751	lar de idos	873002	lar idosos	873	873	censo	TRUE					
18	Porto	IECAT2016111	IECAT	2016.SMN	2016	20070	1751	posto da g	842007	posto gnr	842	842	censo	TRUE					
19	Porto	IECAT2016111	IECAT	2016.SMN	2016	20070	1910	Prestação	712001	prestacao	712								
20	Porto	IECAT2016111	IECAT	2016.SMN	2016	20070	1910	Comercial	464001	comercial	464								
21	Porto	IECAT2016111	IECAT	2016.SMN	2016	20070	1995	colegio pe	853001	colegio pe	853								
22	Porto	IECAT201611A	IECAT	2016.SMN	2016	22100	0187	produção	101001	producao	101								
23	Porto	IECAT201611A	IECAT	2016.SMN	2016	22100	0187	fabrico de	273003	fabrico ca	273	273	censo	TRUE					
24	Porto	IECAT201611A	IECAT	2016.SMN	2016	22100	0298	Casino	920009	casino	920	920	censo	TRUE					
25	Porto	IECAT201611A	IECAT	2016.SMN	2016	22100	0298	Gestão de	683008	gestao cor	683	683	censo	TRUE					
26	Porto	IECAT201611A	IECAT	2016.SMN	2016	22100	0619	instituiçã	641001	instituicao	641	641	censo	TRUE					
27	Porto	IECAT2016112	IECAT	2016.SMN	2016	20540	0251	construçã	412001	construca	412	412	censo	TRUE					
28	Porto	IECAT2016112	IECAT	2016.SMN	2016	20540	0251	montagen	264001	montagen	264								
29	Porto	IECAT201611C	IECAT	2016.SMN	2016	22240	0180	fabrica de	139001	fabrica m	139								
30	Porto	IECAT201611A	IECAT	2016.SMN	2016	20570	0358	Confecção	141001	confecao	141								
31	Porto	IECAT201611A	IECAT	2016.SMN	2016	20570	0395	confeçoes	141001	confecoes	141	141	censo	TRUE					
32	Porto	IECAT201611A	IECAT	2016.SMN	2016	20570	0419	call cente	822001	call cente	822								
33	Porto	IECAT201611A	IECAT	2016.SMN	2016	20660	0067	creche	889002	creche	889	889	censo	TRUE					
34	Porto	IECAT2016119	IECAT	2016.SMN	2016	20820	0345	Costureire	952001	costureira	952	141	censo	FALSE					
35	Porto	IECAT2016119	IECAT	2016.SMN	2016	20820	0345	Escola do	853001	escola do	853								
36	Porto	IECAT2016119	IECAT	2016.SMN	2016	20820	0439	patrulha	842001	patrulha	842								
37	Porto	IECAT2016119	IECAT	2016.SMN	2016	20820	0439	café	563001	cafe	563	563	censo	TRUE					
38	Porto	IECAT2016119	IECAT	2016.SMN	2016	20820	0455	oficina re	452001	oficina re	452	452	censo	TRUE					
39	Porto	IECAT2016119	IECAT	2016.SMN	2016	20820	0455	creche	889002	creche	889	889	censo	TRUE					
40	Porto	IECAT2016119	IECAT	2016.SMN	2016	20820	0455	oficina de	452001	oficina re	452	452	censo	TRUE					
41	Porto	IECAT2016119	IECAT	2016.SMN	2016	20820	0466	OFICINA	452002	oficina re	452	452	inqfam3o	TRUE					
42	Porto	IECAT2016119	IECAT	2016.SMN	2016	20820	0466	OFICINA	452002	oficina re	452	452	inqfam3o	TRUE					
43	Porto	IECAT201611A	IECAT	2016.SMN	2016	20980	0433	CLINICA	861001	clinica cor	861								



RMarkdown Report

The screenshot displays two overlapping Excel windows. The top window shows a pivot table with the following data:

cae_dsg	total	n_cods	n_reg
agropecuaria	14	1	1
agricultura subsistencia	5	1	2
clinica fisioterapia sem internamento	5	2	3
escola basica integrada	4	1	1
snack bar	4	1	2

The bottom window shows a list of data rows with columns: cr, ID, NUTS, INQ_S, and PER_S. The rows are numbered 1 through 43, with the first few rows being: 1 cr, 2 Lisboa, 3 Açores, 4 Açores, 5 Açores, 6 Açores, 7 Açores, 8 Açores, 9 Açores, 10 Madeira, 11 Madeira, 12 Madeira, 13 Açores, 14 Açores, 15 Açores, 16 Porto, 17 Porto, 18 Porto, 19 Porto, 20 Porto, 21 Porto, 22 Porto, 23 Porto, 24 Porto, 25 Porto, 26 Porto, 27 Porto, 28 Porto, 29 Porto, 30 Porto, 31 Porto, 32 Porto, 33 Porto, 34 Porto, 35 Porto, 36 Porto, 37 Porto, 38 Porto, 39 Porto, 40 Porto, 41 Porto, 42 Porto, 43 Porto.



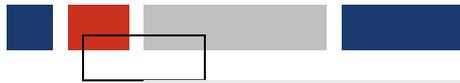
RMarkdown Report

The image displays a Microsoft Excel spreadsheet with a data table. The table has the following columns: ID, NUTS, INQ, SI, PER, SM, ANO, AREA, N_ALO, V1430, V1432, cr, caem3, cae_dsg, cae_co, cae_foi. The data rows contain various identifiers and codes, such as 'IECAT2016170', 'IECAT2011119', and 'IECAT201111E'. The spreadsheet is titled 'IECAT_CAE_compara_2017_07_17_16_57_35.xlsx'.

ID	NUTS	INQ	SI	PER	SM	ANO	AREA	N_ALO	V1430	V1432	cr	caem3	cae_dsg	cae_co	cae_foi
1	IECAT2016170	IECAT	2016.S				01000	0154	agricultura subsistencia	011001	Porto	011	agricultur,015		censos
2	IECAT2011119	IECAT	2011.SMN	2011			03970	0053	agricultura de subsistencia	011001	Porto	011	agricultur,015		censos
3	IECAT201111E	IECAT	2011.SMN	2011			03970	0053	agricultura de subsistencia	011001	Porto	011	agricultur,015		censos
4	IECAT201116G	IECAT	2011.SMN	2011			05560	0452	agricultura de subsistencia	011001	Coimbra	011	agricultur,015		censos
5	IECAT201116D	IECAT	2011.SMN	2011			04310	0163	AGRICULTURA DE SUBSISTENCIA	011001	Coimbra	011	agricultur,015		censos
6	IECAT201116F	IECAT	2011.SMN	2011			05430	0375	agricultura de subsistencia	011001	Coimbra	011	agricultur,015		censos
7	IECAT201116F	IECAT	2011.SMN	2011			05430	0375	agricultura de subsistencia	011001	Coimbra	011	agricultur,015		censos
8	IECAT2011185	IECAT	2011.SMN	2011			10790	0373	agricultura de subsistencia	011025	Évora	011	agricultur,015		censos
9	IECAT2011185	IECAT	2011.SMN	2011			10620	0237	agricultura de subsistencia	011025	Évora	011	agricultur,015		censos
10	IECAT2011185	IECAT	2011.SMN	2011			10840	0185	agricultura de subsistencia	011025	Évora	011	agricultur,015		censos
11	IECAT2011118	IECAT	2011.SMN	2011			04100	0210	agricultura de subsistencia	011001	Porto	011	agricultur,015		censos
12	IECAT2011111	IECAT	2011.SMN	2011			00050	0377	agricultura de subsistencia	011001	Porto	011	agricultur,015		censos
13	IECAT2011119	IECAT	2011.SMN	2011			01010	0302	AGRICULTURA DE SUBSISTENCIA	011001	Porto	011	agricultur,015		censos
14	IECAT2011300	IECAT	2011.SMN	2011			13651	5340	AGRICULTURA DE SUBSISTENCIA	011001	Madeira	011	agricultur,015		censos
15	IECAT2011119	IECAT	2011.SMN	2011			00800	0425	agricultura de subsistencia	011001	Porto	011	agricultur,015		censos
16	IECAT201116J	IECAT	2011.SMN	2011			05980	0189	agricultura de subsistencia	011001	Coimbra	011	agricultur,015		censos
17	IECAT201116J	IECAT	2011.SMN	2011			05980	0189	agricultura de subsistencia	011001	Coimbra	011	agricultur,015		censos
18	IECAT201116J	IECAT	2011.SMN	2011			05980	0532	Agricultura de Subsistencia	011001	Coimbra	011	agricultur,015		censos
19	IECAT201116E	IECAT	2011.SMN	2011			04591	5103	agricultura de subsistencia	011001	Coimbra	011	agricultur,015		censos
20	IECAT201116E	IECAT	2011.SMN	2011			04591	5103	agricultura de subsistencia	011001	Coimbra	011	agricultur,015		censos
21	IECAT201116E	IECAT	2011.SMN	2011			04591	5103	agricultura de subsistencia	011001	Coimbra	011	agricultur,015		censos
22	IECAT201116I	IECAT	2011.SMN	2011			05840	0898	agricultura subsistencia	011001	Coimbra	011	agricultur,015		censos
23	IECAT201116I	IECAT	2011.SMN	2011			05840	0911	AGRICULTURA DE SUBSISTENCIA	011001	Coimbra	011	agricultur,015		censos
24	IECAT201116E	IECAT	2011.SMN	2011			04591	5253	agricultura de subsistencia	011001	Coimbra	011	agricultur,015		censos
25	IECAT201116E	IECAT	2011.SMN	2011			04591	5253	agricultura de subsistencia	011001	Coimbra	011	agricultur,015		censos
26	IECAT2011300	IECAT	2011.SMN	2011			13661	5156	AGRICULTURA SUBSISTENCIA	011001	Madeira	011	agricultur,015		censos
27	IECAT2011300	IECAT	2011.SMN	2011			13661	5158	AGRICULTURA SUBSISTENCIA	011001	Madeira	011	agricultur,015		censos
28	IECAT2011200	IECAT	2011.SMN	2011			12820	0355	agricultura de subsistencia	011001	Açores	011	agricultur,015		censos
29	IECAT2011111	IECAT	2011.SMN	2011			00060	0847	agricultura subsistencia	011001	Porto	011	agricultur,015		censos
30	IECAT2011300	IECAT	2011.SMN	2011			13661	5463	agricultura subsistencia	011001	Madeira	011	agricultur,015		censos
31	IECAT2011300	IECAT	2011.SMN	2011			13661	5463	agricultura subsistencia	011001	Madeira	011	agricultur,015		censos
32	IECAT2011300	IECAT	2011.SMN	2011			13601	5140	agricultura de subsistencia	011001	Madeira	011	agricultur,015		censos
33	IECAT2011300	IECAT	2011.SMN	2011			13660	0352	agricultura subsistencia	011001	Madeira	011	agricultur,015		censos
34	IECAT2011300	IECAT	2011.SMN	2011			13790	0119	AGRICULTURA DE SUBSISTENCIA	011001	Madeira	011	agricultur,015		censos
35	IECAT201116F	IECAT	2011.SMN	2011			04990	0307	agricultura subsistencia	011001	Coimbra	011	agricultur,015		censos
36	IECAT201111A	IECAT	2011.SMN	2011			03270	0140	AGRICULTURA SUBSISTENCIA	011001	Porto	011	agricultur,015		censos
37	IECAT201116E	IECAT	2011.SMN	2011			04910	0339	agricultura de subsistencia	011001	Coimbra	011	agricultur,015		censos
38	IECAT201116E	IECAT	2011.SMN	2011			04910	0346	AGRICULTURA SUBSISTENCIA	011001	Coimbra	011	agricultur,015		censos
39	IECAT201116G	IECAT	2011.SMN	2011			05470	0675	AGRICULTURA SUBSISTENCIA	015001	Coimbra	015	agricultur,015		censos
40	IECAT201116F	IECAT	2011.SMN	2011			04990	0069	agricultura subsistencia	011001	Coimbra	011	agricultur,015		censos
41	IECAT201116F	IECAT	2011.SMN	2011			04990	0069	agricultura subsistencia	011001	Coimbra	011	agricultur,015		censos
42	IECAT201116G	IECAT	2011.SMN	2011			05480	0532	agricultura de subsistencia	011001	Coimbra	011	agricultur,015		censos
43	IECAT201116G	IECAT	2011.SMN	2011			05480	0532	agricultura de subsistencia	011001	Coimbra	011	agricultur,015		censos



Wikipage dashboard



You are here: [rss](#) > [drgd](#) > [codificação](#) > [dashboard_ie](#)
Trace: - [dashboard_ie](#)

DGRD

[Quem somos?](#)
[Plano de Atividades](#)
[Reuniões internas](#)

Serviços:

[SDE - Dados Empresariais](#)
[SIE - Inquéritos por Entrevista](#)
[SIP - Integração de Processos](#)

Projetos:

[IE - Inquérito ao Emprego](#)
[Outros Inquéritos às Famílias](#)
[IPC - Índice de Preços no Consumidor](#)
[Outros Inquéritos](#)
[CATI - Recolha telefónica](#)
[CAWI - Recolha eletrónica](#)
[Data Science](#)
[Codificação Automática](#)
[Modos Mistos](#)
[Comunicação com as Empresas](#)
[Comunicação com as Famílias](#)
[Imob](#)
[Scanner Data](#)
[Web Scraping](#)

Divulgação e investigação

[Artigos e apresentações externas](#)
[UNECE Data Collection Workshop](#)
[2018 BDCM Workshop](#)
[Jornadas de Produção Estatística - Gestão da Recolha e Análise de Dados](#)
[Big Data](#)
[WS Produção Estatística e as Empresas - 2013](#)
[WS Custos de Contexto - 2015](#)

 [drgd](#)
 [cati](#)

drgd:codificacao:dashboard_ie

DRGD | SIP

2019-05-17 08:00:04

Classificação automática de apoio ao Inquérito ao Emprego

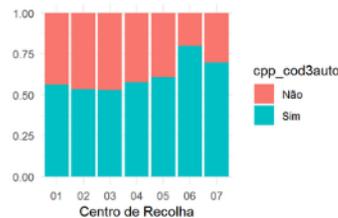
Para visualizar esta página utilize o Firefox ou o Chrome

[Classificação das Atividades Económicas](#)

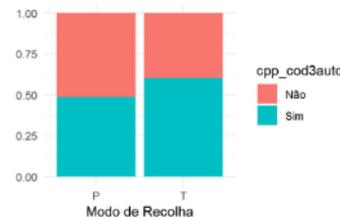
[Classificação das Profissões](#)

[Legenda de variáveis](#)

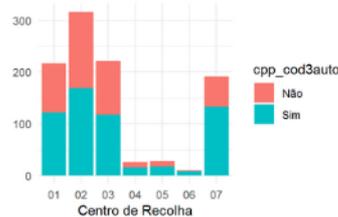
Classificação automática (%)
Por Centro de Recolha



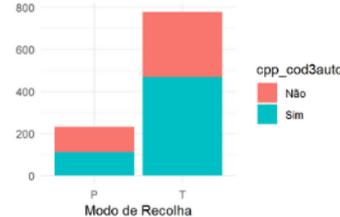
Classificação automática (%)
Por Modo de Recolha



Classificação automática (n)
Por Centro de Recolha



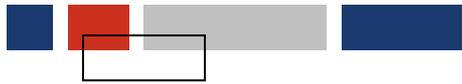
Classificação automática (n)
Por Modo de Recolha



Monitor performance and validate results

- Adult Education Survey (2016):
 - Automatic classification thoroughly reviewed by the best team of coders
- Dictionaries
 - Occupation and Economic Activity Dictionaries reviewed by all the 7 regional teams of coders
- Data collection
 - Feedback from / to coders



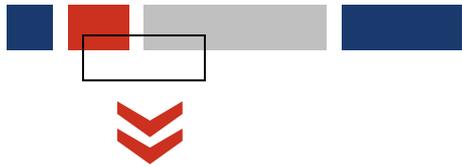


Occupation (3 digits)

```
dic_cpp %>%  
  tabyl(cpp_fonte3) %>%  
  knitr::kable()
```

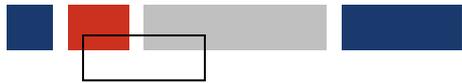
cpp_fonte3	n	percent
3osa	16810	0.4048359
coders	1336	0.0321749
list	20213	0.4867905
rule	3164	0.0761987

Packages



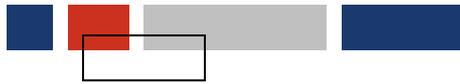
- tidyverse
- formattable
- xda
- ROracle
- Openxlsx
- knitr
- waffle
- lubridate
- janitor
- png
- grid
- gridExtra
- RColorBrewer





SOME RESULTS

Information and Communication Technologies 2018 (CAWI)



```
left_join(sictu_cawi, dic_cpp, "cpp_dsg") %>%  
  mutate(coda = !is.na(cpp_fonte3)) %>%  
  tabyl(coda) %>% knitr::kable()
```

coda	n	percent
FALSE	402	0.3089931
TRUE	899	0.6910069

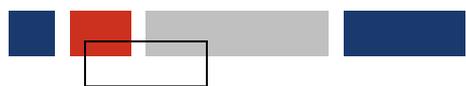


Information and Communication Technologies 2018 (CAWI)



```
left_join(sictu_cawi, dic_cpp, "cpp_dsg") %>%  
  tabyl/cpp_fonte3) %>%  
  knitr::kable()
```

cpp_fonte3	n	percent	valid_percent
3osa	166	0.1275942	0.1846496
coders	110	0.0845503	0.1223582
list	166	0.1275942	0.1846496
rule	457	0.3512683	0.5083426
NA	402	0.3089931	NA



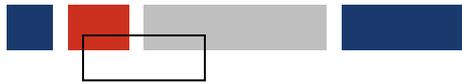
Limitations and shortcomings



- Ineffective with new strings
- Dictionary expansion depends on human validation
- Changes may occur with new interviewers and new coders
- Changes may occur when coders change their mind



Advantages and benefits



- Consistent
- Error management
- Speed
- Effective





NEXT STEPS



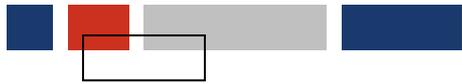
Next steps



- Every new [string + code] is added to dictionary
- spelling + stemming
- Different distance metrics for short and long strings
- n-grams



Next steps



- Special attention on specific cases
- Use auxiliary variables
- Provide an extensive and properly validated dataset for machine learning
- uRos2020 8th international conference





Use of R in Official Statistics 2019

7th international conference

ruivalves@ine.pt

THANK YOU FOR YOUR ATTENTION!



INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL

