Deploying a text classification service

Instituto Nacional de Estadísticas - Chile

November 2021

Contents

2.
Motivation

Duplicated efforts

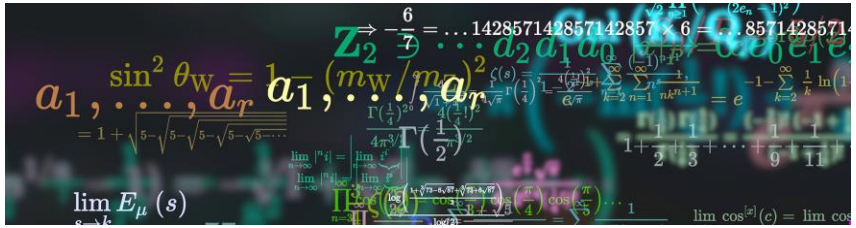
Many teams are carring out similar classification tasks

- Occupation
- Economic activity
- Production
- Consumption

The challenge

- **To create one single platform for text classification**
  - datasets
  - methodology (algorithm, quality measures, etc.)
  - programming language

```
import random

def game(n):
    '''An addition game that prompts user to answer simple addition problems
    using random numbers from 0-9 and prints whether they are correct or not.'''
    wrk = True
    cnt = 0
    for i in range(n):
        val1 = random.randrange(1,10)
        val2 = random.randrange(1,10)
        wrk = True
        while wrk == True:
            try:
                ans = int(input(str(val1)+ " + "+ str(val2)+ ": "))
                if(ans == val1+val2):
                    print("Correct!!!")
                    wrk = False
                    cnt += 1
                else:
                    print("Incorrect...")
                    wrk = False
            except:
                print("That's not a valid value :( Try again!!!")
    print("You got "+str(cnt)+" out of "+str(n)+" correct!!!!!")
```



Building of the datasets

- We created 2 train datasets
  - Economic activity
  - Occupation

- Trained analysts labelled records

- ~ 50.000 records for economic activity
- ~ 30.000 records for occupation

- 90% of records was labelled twice

2.
The model

The model

- 4 strategies were tested and the selected one was:
  - Neural net with GRU architecture
  - Spanish word embeddings for text representation

Economic activity (1 digit classification)

modelo	acc	macro	micro	weighted
seq_1d	0.9384	0.8757	0.9384	0.9386
tfidf_1d	0.9327	0.8658	0.9327	0.9330
emb_simple_1d	0.9274	0.8641	0.9274	0.9280
emb_gru_1d	0.9327	0.8694	0.9327	0.9328

https://github.com/dccuchile/spanish-word-embeddings

The model

Occupation (1 digit classification)

modelo	acc	macro	micro	weighted
seq_1d	0.8858	0.8599	0.8858	0.8855
tfidf_1d	0.8684	0.8362	0.8684	0.8686
emb_simple_1d	0.8793	0.8519	0.8793	0.8807
emb_gru_1d	0.8989	0.8796	0.8989	0.8990

- 4 models were created
  - Occupation 1 digit
  - Occupation 2 digits
  - Economic activity 1 digit
  - Economic activity 2 digits

https://github.com/inesscc/ineclassifiers

Sharing the models

```
keras::save_model_hdf5(model, "awesome_model")
```

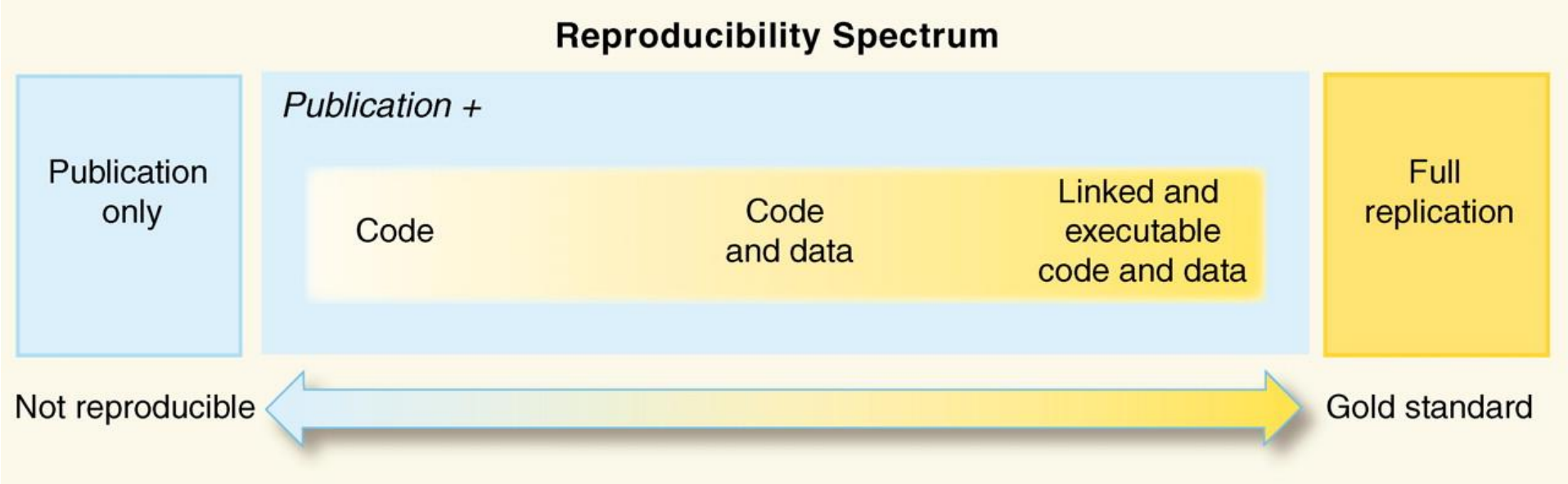
Dependencies

- Also, sometimes it is neccesary to create a **virtualenv** in a specific directory

Dependencies

- Even in the case the user is able to install the dependencies, the output may not be the same
  - R version
  - Packages version
  - OS

Dependencies and reproducibility



2.
Deployment

First idea: R package

R package (library) containing the model

Better than sharing the raw files

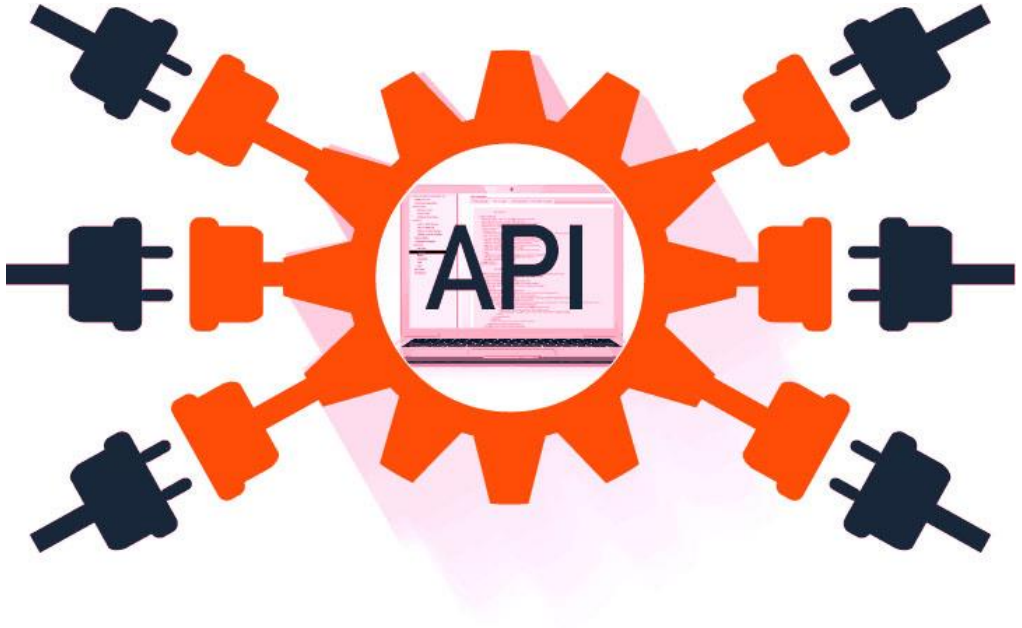
This strategy still relies on the user setup

Final idea: API

Upload the model to a server with all the dependencies

We can provide a service through an API

The user needs only a tool to make a requests



Final idea: API

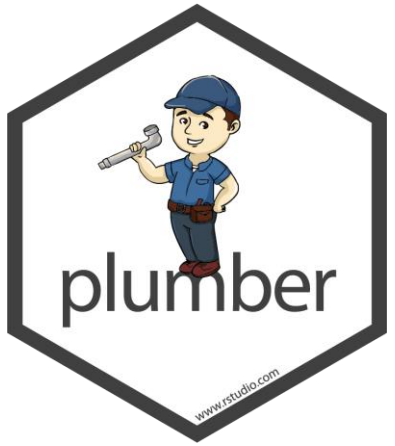

```
["aseador en oficinas","médico clínica Las Condes"]
```

```r
library(httr)
library(feather)
caenes <- read_feather("src/data/split_train_test/test.feather")
request <-  httr::POST("http://143.198.79.143:8080/predict",
                encode = "json",
                body =  list(text = caenes$glosa_caenes[1:10],
                             classification = "caenes",
                             digits = 1)
)

httr::status_code(request)
response <- httr::content(request)
```

```
[[1]]
[[1]]$codigo_int
[1] 13

[[1]]$cod_final
[1] "N"
```

Adventages of this solution

- The user does not requiere any dependence
- The output will be always the same (reproducibility)
- Different programming languages

RStudio Server +  + 

Final thoughts

Efficiency considerations:

- The goverment has a duty to reduce the public expenditure
- We have to avoid duplicated efforts and expand the automation
- We are exploring the posibility to open this API to provide a service to the general public (statistical offices, universities, NGOs, private sector).

Technical considerations:

- We have to keep in mind the harmonization across the official statistics
- It is very important monitoring changes in the fieldwork methodologies (questionnaires, data collection conditions, etc).
- Our predictions can be affected by those changes.

Contact

✉ kilehmannm@ine.gob.cl: Klaus Lehmann - Senior Analyst
✉ ifaglonij@ine.gob.cl: Ignacio Agloni - Project Manager

https://github.com/inesscc

Instituto Nacional de Estadísticas · Chile

THANKS
GRACIAS

ine.gob.cl