ESTRAZIONE AUTOMATICA DI INFORMAZIONI DA SMS

Progetto IA

d'Argenio Mattia
Montefusco Alberto
Aquino Alessandro

CONTENTS

PROBLEMA

Sviluppo di un sistema di **Estrazione Automatica** di Informazioni da SMS in lingua inglese.

spaCy fornisce una serie di modelli pre-addestrati: nel seguente caso, il modello preso in esame è RoBERTa .

spaCy

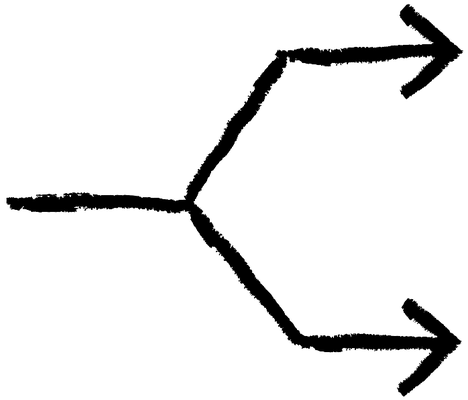
WORKFLOW

Individuazione Dataset

SMS-NER-Dataset-165-Annotations

Data cleaning

Training e testing

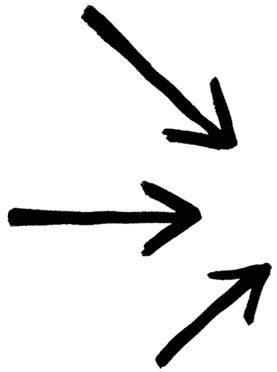


model-last: il modello addestrato nell'ultima iterazione

model-best: il modello che ha ottenuto il punteggio più alto sul dataset di test

Spacy-Config

Modello



1. en_core_web_sm
2. en_core_web_md
3. en_core_web_lg

robBERTa

IMPLEMENTAZIONE

SMS-NER-Dataset-165-Annotations.json

1. Classes
2. Annotations
3. Entities

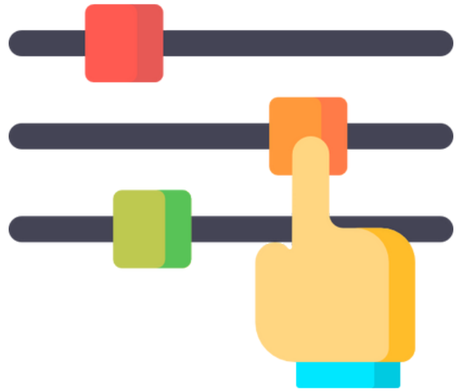
"MONEY", "TITLE", "OTP",
"TRANSAC", "TIME", "PURPOSE"

[19, 26, "TRANSAC"]

IMPLEMENTAZIONE

IPER-PARAMETRI: CONFIG.CFG

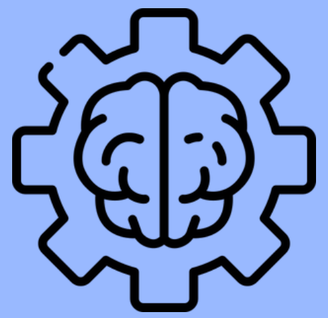
- **Modello Transformer:**
  - name = "roberta-base"
  - mixed_precision = false
- **Tokenizer:**
  - use_fast = true
- **Componente NER (Named Entity Recognition):**
  - batch_size = 128
  - dropout = 0.1
  - hidden_width = 64
  - maxout_pieces = 2
  - use_upper = false
- **Ottimizzatore:**
  - @optimizers = "Adam.v1"
- **Pianificazione del Tasso di Apprendimento:**
  - @schedules = "warmup_linear.v1"
  - warmup_steps = 250
  - total_steps = 20000



IMPLEMENTAZIONE

base_config.cfg

```
python -m spacy init fill-config
dataset/SMS-NER-Dataset-165-Annotations/
base_config.cfg config.cfg
```



Training e testing

```
python -m spacy train config.cfg –output
./output –paths.train train.spacy –paths.dev
test.spacy
```

Otteniamo le metriche

```
python -m spacy benchmark
accuracy model/large/model-best
model/large/test.spacy –output –code
–gold-preproc –gpu-id 0 –displacy-path
model/large
```

ANALISI DEI DATI

```
================================== Results ===

TOK      100.00
NER P    76.09
NER R    80.46
NER F    78.21
SPEED    496


============================== NER (per type)

                P        R        F
OTP         70.00    77.78    73.68
PURPOSE     62.50    65.22    63.83
TITLE       78.79    86.67    82.54
MONEY       81.82    81.82    81.82
TRANSAC    100.00    88.89    94.12
TIME        83.33   100.00    90.91
```

Medium

Small

```
================================== Results ====

TOK      100.00
NER P    72.92
NER R    80.46
NER F    76.50
SPEED    406


============================== NER (per type) :

                P        R        F
OTP         72.73    88.89    80.00
PURPOSE     62.50    65.22    63.83
TITLE       72.73    80.00    76.19
MONEY       69.23    81.82    75.00
TRANSAC    100.00   100.00   100.00
TIME        83.33   100.00    90.91
```

ANALISI DEI DATI

Trf

```
================================== Results ============

TOK      100.00
NER P    73.91
NER R    78.16
NER F    75.98
SPEED    483


============================== NER (per type) ========

                P        R        F
OTP         87.50    77.78    82.35
PURPOSE     60.00    65.22    62.50
TITLE       71.88    76.67    74.19
MONEY       75.00    81.82    78.26
TRANSAC    100.00   100.00   100.00
TIME        83.33   100.00    90.91
```

```
================================== Results ====:

TOK      100.00
NER P    75.82
NER R    79.31
NER F    77.53
SPEED    478


============================== NER (per type)

                P        R        F
OTP         87.50    77.78    82.35
PURPOSE     64.00    69.57    66.67
TITLE       71.88    76.67    74.19
MONEY       75.00    81.82    78.26
TRANSAC    100.00   100.00   100.00
TIME       100.00   100.00   100.00
```

Large

PROBLEMI RISCONTRATI

Metriche raggiunte

80% circa

Qualità dataset

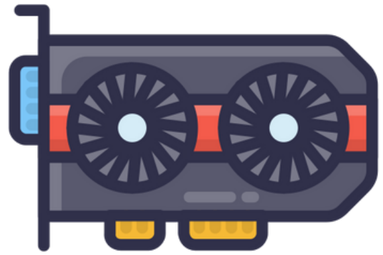
abbiamo allenato il modello su 165 sample dove 132 sono stati utilizzati per il training e 33

Dimensione dataset

100k sample di messaggi SM

Risorse hardware

Risorse limitate
3/4 ore per completare
Crush del sistema



DISPONIBILITÀ DEI DATI

Tutto il codice è disponibile sulla nostra repo di GitHub

Thank You

d’Argenio Mattia
Montefusco Alberto
Aquino Alessandro