Scripting no Processamento de Linguagem Natural

Trabalho Prático 1

João Carvalho - PG50496
Joaquim Roque - PG50502
Vasco Matos - PG50796

Módulos usados

PyTesseract

TextBlob

Pillow

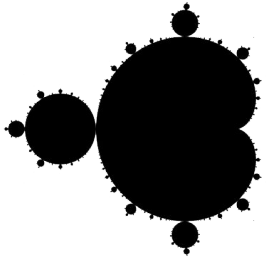
Argparse

PyTesseract

- É um *optical character recognition* (OCR), i.e. reconhece texto contido em imagens.
- *Binding* para a biblioteca da Google **Tesseract**
- Suporta múltiplos formatos de imagens (jpg, png, gif e outros)
- Produz informação relevante relativamente ao texto encontrado (i.e. *bounding boxes* dos caracteres, confiança)

TextBlob

- Biblioteca para processamento de dados sob a forma de texto.
- API orientada a tarefas comuns no Processamento de Linguagem Natural (i.e. tokenização, traduções, *sentiment analysis*, deteção de linguagem, entre outras)



TextBlob

Opções do Programa

Bounding Box Estimates
Retorna os caracteres conhecidos e as fronteiras do espaço onde está contido.

Verbose Data
Retorna os caracteres e palavras conhecidas, as fronteiras do espaço onde estão contidos, a confiança percentual daquele resultado, bem como outras informações.

Orientation
Retorna a informação sobre a orientação do texto, bem como a percentagem de confiança desta informação.

PDF
Cria um ficheiro pdf com todos os elementos presentes na imagem, convertendo o texto da imagem para texto real.

XML
Cria um ficheiro XML à semelhança do PDF.

HOCR
Cria um ficheiro hOCR com os mesmos elementos do ficheiro pdf.
O hOCR é um padrão aberto de representação de dados para texto formatado obtido a partir de reconhecimento ótico de caracteres.

Translate
Com a utilização do TextBlob o texto contido na imagem é traduzido para a linguagem pretendida.

Chapter 1

THE WORST BIRTHDAY

Not for the first time, an argument had broken out over breakfast at number four, Privet Drive. Mr. Vernon Dursley had been woken in the early hours of the morning by a loud, hooting noise from his nephew Harry's room.

"Third time this week!" he roared across the table. "If you can't control that owl, it'll have to go!"

Harry tried, yet again, to explain.

"She's *bored,*" he said. "She's used to flying around outside. If I could just let her out at night —"

"Do I look stupid?" snarled Uncle Vernon, a bit of fried egg dangling from his bushy mustache. "I know what'll happen if that owl's let out."

He exchanged dark looks with his wife, Petunia.

Harry tried to argue back but his words were drowned by a long, loud belch from the Dursleys' son, Dudley.

THE WORST BIRTHDAY

Not for the first time, an argument had broken out over breakfast at number four, Privet Drive. Mr. Vernon Dursley had been woken in the early hours of the morning by a loud, hooting noise from his nephew Harry's room.

"Third time this week!" he roared across the table. "If you can't control that owl, it'll have to go!"

Harry tried, yet again, to explain.

"She's bored," he said. "She's used to flying around outside. If I could just let her out at night —"

"Do I look stupid?" snarled Uncle Vernon, a bit of fried egg dangling from his bushy mustache. "I know what'll happen if that owl's let out."

He exchanged dark looks with his wife, Petunia.

Harry tried to argue back but his words were drowned by a long, loud belch from the Dursleys' son, Dudley.

Verbose data:

level	page_num	block_num	par_num	line_num	word_num	left	top	width	height	conf	text
1	1	0	0	0	0	0	1078	1800	-1		
2	1	1	0	0	0	337	62	421	550	-1	
3	1	1	1	0	0	337	62	421	550	-1	
4	1	1	1	1	0	337	62	421	550	-1	
5	1	1	1	1	1	337	62	421	550	95.000000	
2	1	2	0	0	0	243	691	594	35	-1	
3	1	2	1	0	0	243	691	594	35	-1	
4	1	2	1	1	0	243	691	594	35	-1	
5	1	2	1	1	1	243	691	107	34	96.225616	THE
5	1	2	1	1	2	367	691	185	35	96.479073	WORST
5	1	2	1	1	3	571	691	266	34	96.153099	BIRTHDAY
2	1	3	0	0	0	63	786	952	203	-1	

```
Bounding box estimates:
~ 337 1188 758 1738 0
T 243 1075 276 1109 0
H 279 1075 315 1108 0
E 318 1075 350 1108 0
W 367 1075 412 1108 0
O 414 1074 449 1108 0
R 452 1075 487 1108 0
S 489 1074 517 1108 0
T 519 1075 552 1109 0
B 571 1075 602 1108 0
I 606 1075 622 1108 0
R 625 1075 660 1108 0
T 660 1075 693 1109 0
H 695 1075 731 1108 0
D 735 1075 769 1108 0
A 770 1075 804 1108 0
Y 804 1075 837 1108 0
N 64 988 88 1013 0
o 77 987 96 1013 0
t 91 987 123 1011 0
f 136 988 150 1014 0
o 148 988 166 1005 0
r 168 988 183 1006 0
```

Translated text:
O pior aniversário

Não pela primeira vez, uma discussão havia surgido
Durante o café da manhã no número quatro, a Drive Drive. Senhor.
Vernon Dursley havia sido acordado nas primeiras horas de
a manhã por um barulho alto e girando de seu sobrinho
Quarto de Harry.

"Terceira vez nesta semana!" Ele rugiu do outro lado da mesa. "Se
Você não pode controlar essa coruja, terá que ir! "

Harry tentou, mais uma vez, explicar.

"Ela está entediada", disse ele. "Ela está acostumada a voar por aí
fora. Se eu pudesse deixá -la sair à noite - "

"Eu pareço estúpido?" tio vernon rosnado, um pouco de
Ovo frito pendurado em seu bigode espesso. "Eu sei
O que vai acontecer se essa coruja sair. "

Ele trocou olhares sombrios com sua esposa, Petúnia.

Harry tentou discutir de volta, mas suas palavras foram afogadas
Por um longo e alto Belch do filho dos Dursleys, Dudley.


```
Orientation and script detection:
Page number: 0
Orientation in degrees: 0
Rotate: 0
Orientation confidence: 27.89
Script: Latin
Script confidence: 23.06
```

Falhas PyTesseract

```
"WHEN I ney 5 wa TOLD | ME TO BE A BIG BOY.
WHEN I WAS 10 THEY TOLD ME I SHOULD BE MORE MATURE.
NOW THEY SAY IT'S TIME TO START ACTING LIKE AN ADULT.
AT THIS RATE, I'LL BE ELIGIBLE FOR SOCIAL SECURITY
BEFORE I GRADUATE FROM HIGH SCHOOL!"
```

Verbose Data:

```
94.466095       "WHEN
65.429123       I
36.132912       ney
96.594101       5
42.344177       wa
47.365131       TOLD
73.372459       |
95.584686       ME
```

Falhas PyTesseract

Falhas ao tentar reconhecer texto rodado