

Data mining, praktični zadatak, 2021.

Zadatak je zasnovan na *ICDAR 2019 Robust Reading Challenge on Scanned Receipts OCR and Information Extraction*.

Reference,

<https://rrc.cvc.uab.es/?ch=13&com=tasks>

<https://medium.com/analytics-vidhya/extracting-structured-data-from-invoice-96cf5e548e40>

<https://www.taylorfrancis.com/chapters/edit/10.1201/b19007-13/text-classification-using-python-david-colton>

Zadatak se sastoji u ekstrakciji ključnih polja sa skeniranih računa, kao što je prikazano na slici. Ovdje su ključna polja *company*, *date*, *address* i *total*.

tan woon yann
company: **BOOK TA .K (TAMAN DAYA) SDN BHD**
789-117-W
address: **NO.53 55,57 & 59, JALAN SAGU 18,
TAMAN DAYA,
81100 JOHOR BAHRU,
JOHOR.**

Document No : TD01167104
Date : **25/12/2018** 6:13:39 PM
Cashier : MANIS
Member :

CASH BILL

CODE/DESC	PRICE	Disc	AMOUNT
QTY	RM		RM
9556939040116 KF MODELLING CLAY KIDDY FISH			
1 PC *	9.000	0.00	9.00
Total :			9.00
Rounding Adjustment :			0.00
Round d Total (RM):			total 9.00
Cash			10.00
CHANGE			1.00

GOODS SOLD ARE NOT RETURNABLE OR EXCHANGEABLE
凡售出之貨物恕不退換或更換
如有不便, 敬請原諒, 謝謝!

THANK YOU
PLEASE COME AGAIN!

```
{  
  "company": "BOOK TA .K (TAMAN DAYA) SDN BHD",  
  "date": "25/12/2018",  
  "address": "NO.53 55,57 & 59, JALAN SAGU 18,  
            TAMAN DAYA, 81100 JOHOR BAHRU, JOHOR.",  
  "total": "9.00"  
}
```

Dataset za rad sastoji se od 625 slika. Adresa je <https://github.com/zzzDavid/ICDAR-2019-SROIE/tree/master/data>.

Struktura direktorijma poslije pružimanja je prikazana na sljedećoj slici.

```
data/  
  img/  
    000.jpg  
    001.jpg  
  box/  
    000.csv  
    001.csv  
  key/  
    000.json  
    001.json
```

Za svaku sliku iz *img* direktorijma, u *box* direktorijumu postoji odgovarajuća *csv* datoteka koja sadrži niz *bounding box*-ova sa pripadajućim tekstom. Format je prikazan na slici.

```
x1_1, y1_1, x2_1, y2_1, x3_1, y3_1, x4_1, y4_1, transcript_1
```

```
x1_2, y1_2, x2_2, y2_2, x3_2, y3_2, x4_2, y4_2, transcript_2
```

```
x1_3, y1_3, x2_3, y2_3, x3_3, y3_3, x4_3, y4_3, transcript_3
```

...

Bounding box je pravougaonik koji je određen sa četiri tačke koje su navedene počevši od gornjeg lijevog tjemena. Transkript je tekst koji je sadržan u pravougaoniku.

U direktorijumu *key* je za svaku sliku u posebnoj *json* datoteci navedeno koje informacije treba da budu pronađene. Primjer je dat na slici.

```
{"company": "STARBUCKS STORE #10208",  
  "date": "14/03/2015",  
  "address": "11302 EUCLID AVENUE, CLEVELAND, OH (216) 229-0749",  
  "total": "4.95",  
}
```

Dataset je potrebno podijeliti na *train* i *test* tako da prvih 500 slika čine skup za treniranje. Za svaku sliku iz *test* skupa provjerava se da li se pronađena informacija poklapa sa sadržajem odgovarajuće *json* datoteke. Performanse se iskazuju sa Precision (računato prema ukupnom broju polja za ekstrakciju), Recall i F1.

Jezik za implementaciju je Python 3.

Zadatak se radi samostalno. Rok za predaju je 31. maj do 12 sati. Rješenje se šalje na mejl analitickaobradapodataka@gmail.com.

