PDF-to-Text A nightmare that never ends

The basis of information mining for generative AI

Napat Dollapavijit

Most data is unstructured

Key components

Image

Table

Text

Text extraction

Text format extraction

Problem#1 **Image**

Digital PDF

the same as in object detection tasks. It's the intersection of the predicted and ground truth boxes aka. TP divided by the union of the predicted and ground truth boxes, which is essentially $TP + FN + FP$. A example is shown down below.

FIGURE 2.32: taken from https://learnopencv.com

2.3.3.3 Multi-Modal Benchmarks

Visual understanding goes well beyond object recognition or semantic segmentation. With one glance at an image, a human can effortlessly imagine the world beyond the pixels. This is emphasized by the quote "a picture says more

Scanned PDF

Senator self-introduction document

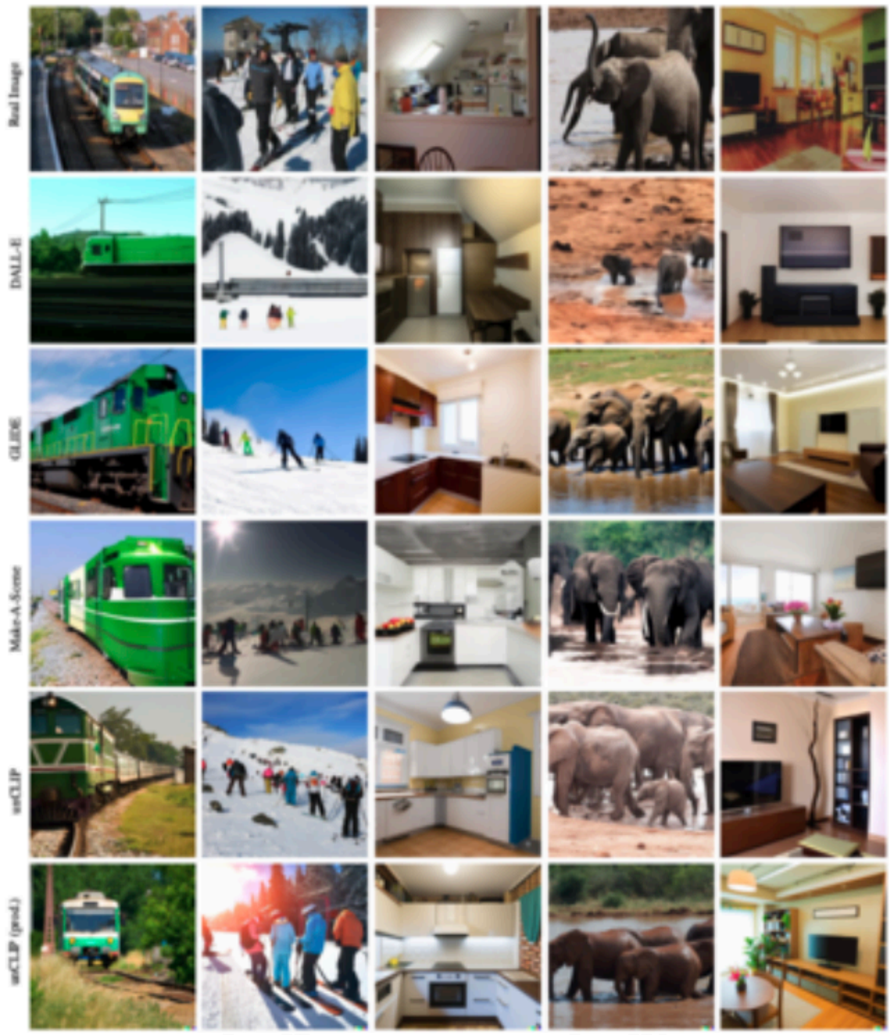


FIGURE 3.31: Image samples on MS-COCO text prompts. Figure from

Is this a one big image or 30 images?

Problem#2 **Table**

Every cells have borders

เงินงบประมาณ	**ล้านบาท**	**273.1663**	**222.8746**	**36.8813**	**36.8813**	**36.8813**
- งบดำเนินงาน	ล้านบาท	-	-	-	-	-
- งบลงทุน	ล้านบาท	-	-	-	-	-
- งบเงินอุดหนุน	ล้านบาท	273.1663	222.8746	36.8813	36.8813	36.8813
- งบรายจ่ายอื่น	ล้านบาท	-	-	-	-	-
เงินนอกงบประมาณ	**ล้านบาท**	**-**	**-**	**-**	**-**	**-**

Only section separators

Model	#Param.	MAC	TEDS (%) Simple	Complex	All
ResNet-18	28.70M	42.22G	98.31	94.50	96.45
LinearProj-28	22.67M	10.28G	94.12 -4.19	86.62 -7.88	90.45 -6.00
ConvStem	24.08M	22.36G	**98.33** +0.02	**94.66** +0.16	**96.53** +0.08
EDD [1]	-	-	91.1	88.7	89.90
GTE [26]	-	-	-	-	93.01
Davar-Lab [28]	-	-	97.88	94.78	96.36
TableFormer [2]	>28.70M	>42.22G	98.50	95.00	96.75

Cell merge

ตัวชี้วัด/ แหล่งเงิน	**หน่วยนับ**	**งบประมาณ**		**ประมาณการรายจ่ายล่วงหน้า****		
		ปี 2563 แผน (ผล)*	**ปี 2564 แผน**	**ปี 2565 แผน**	**ปี 2566 แผน**	**ปี 2567 แผน**
เชิงปริมาณ : จำนวนโครงการ/กิจกรรมบริการวิชาการแก่สังคม	โครงการ	18 (18)	85	90	95	100
เชิงปริมาณ : จำนวนผู้เข้ารับบริการ	คน	3,900	4,000	4,200	4,500	4,700

Table?

No border

วันที่ 20 เดือน ธันวาคม พ.ศ. 2566 เวลา 14:22

จำนวนผู้เข้าร่วมประชุม	418 + 4 = 422
เห็นด้วย	418 + 4 = 422
ไม่เห็นด้วย	0
งดออกเสียง	0
ไม่ลงคะแนนเสียง	0

ร่างพระราชบัญญัติยกเลิกพระราชบัญญัติว่าด้วยความผิดอันเกิดจากการใช้เช็ค พ.ศ. ๒๕๓๔ พ.ศ. ....

วาระที่ ๑

ลำดับที่	เลขที่บัตร	ชื่อ-สกุล	ชื่อสังกัด	ผลการลงคะแนน
1	001	นางสาวกมนทรรศน์ กิตติสุนทรสกุล	พรรคก้าวไกล	เห็นด้วย
2	006	นายกรุณพล เทียนสุวรรณ	พรรคก้าวไกล	เห็นด้วย
3	007	นายกฤช ศิลปชัย	พรรคก้าวไกล	เห็นด้วย
4	009	นางสาวกฤษฎิ์ ชีวะธรรมานนท์	พรรคก้าวไกล	เห็นด้วย
5	010	นายกฤษฐ์หิรัญ เลิศอุฤทธิ์ภักดี	พรรคก้าวไกล	เห็นด้วย
6	014	นายกันต์พงษ์ ประยูรศักดิ์	พรรคก้าวไกล	เห็นด้วย
7	015	นางสาวกัลยพัชร รจิตโรจน์	พรรคก้าวไกล	เห็นด้วย
8	019	นางสาวการณิก จันทตา	พรรคก้าวไกล	เห็นด้วย
9	023	นายกิตติภณ ปานพรหมมาศ	พรรคก้าวไกล	เห็นด้วย
10	031	นายเกียรติคุณ ต้นยาง	พรรคก้าวไกล	เห็นด้วย
11	036	นายคริษฐ์ ปานเนียม	พรรคก้าวไกล	เห็นด้วย
12	039	นายคำพอง เทพาคำ	พรรคก้าวไกล	เห็นด้วย

Problem#2 **Table**

Case: **Parliament meeting log**

จำนวนผู้เข้าร่วมประชุม	418 + 4 = 422
เห็นด้วย	418 + 4 = 422
งดออกเสียง	0
ไม่ลงคะแนนเสียง	0

ความผิดอันเกิดจากการใช้เช็ค พ.ศ. ๒๕๓๔ พ.ศ. ....

วาระที่ ๑

ลำดับที่	เลขที่บัตร	ชื่อ-สกุล	ชื่อสังกัด	ผลการลงคะแนน
1	001	นางสาวกมนทรรศน์ กิตติสุนทรสกุล	พรรคก้าวไกล	เห็นด้วย
2	006	นายกรุณพล เทียนสุวรรณ	พรรคก้าวไกล	เห็นด้วย
3	007	นายกฤช ศิลปชัย	พรรคก้าวไกล	เห็นด้วย
4	009	นางสาวกฤษฎิ์ ชีวะธรรมานนท์	พรรคก้าวไกล	เห็นด้วย
5	010	นายกฤษฐ์หิรัญ เลิศอุฤทธิ์ภักดี	พรรคก้าวไกล	เห็นด้วย
6	014	นายกันต์พงษ์ ประยูรศักดิ์	พรรคก้าวไกล	เห็นด้วย
7	015	นางสาวกัลยพัชร รจิตโรจน์	พรรคก้าวไกล	เห็นด้วย
8	019	นางสาวการณิก จันทดา	พรรคก้าวไกล	เห็นด้วย
9	023	นายกิตติภณ ปานพรหมมาศ	พรรคก้าวไกล	เห็นด้วย
10	031	นายเกียรติคุณ ต้นยาง	พรรคก้าวไกล	เห็นด้วย
11	036	นายคริษฐ์ ปานเนียม	พรรคก้าวไกล	เห็นด้วย

Problem#2 **Table**

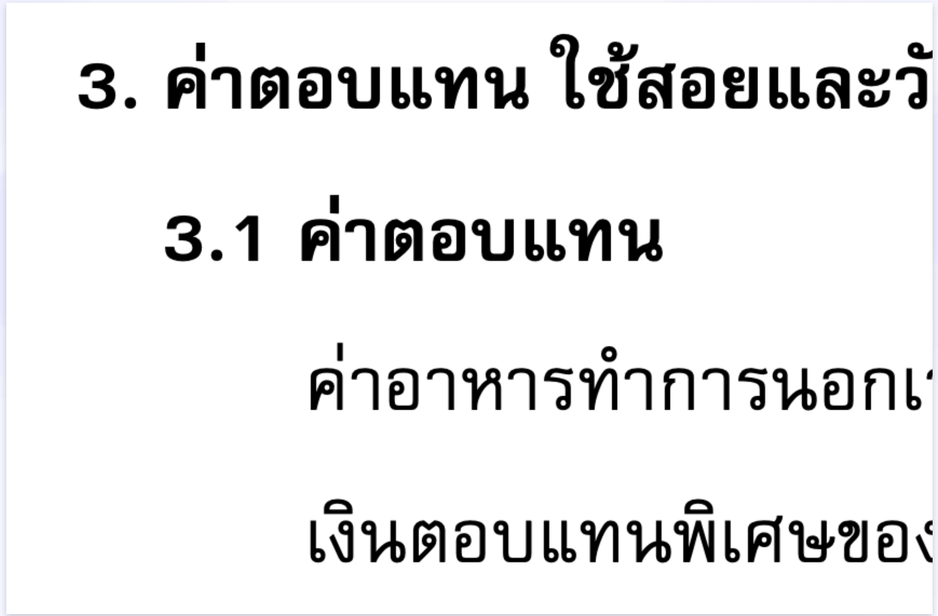
Case: **Bangkok budget**

สำนักปลัดกรุงเทพมหานคร

ด้าน/แผนงาน/งาน/โครงการ	เงินเดือนและค่าจ้างประจำ	ค่าจ้างชั่วคราว	ค่าตอบแทน ใช้สอย และวัสดุ	ค่าสาธารณูปโภค	ค่าครุภัณฑ์ ที่ดิน และสิ่งก่อสร้าง	เงินอุดหนุน	รายจ่ายอื่น	รวม
รายจ่ายประจำ								
ด้านการบริหารทั่วไป	**365,177,600**	**9,282,600**	**146,212,100**	**38,021,800**	**15,873,100**	**2,730,700**	**366,152,000**	**943,449,900**
แผนงานบริหารทั่วไป	**75,400,000**	**3,312,000**	**81,401,700**	**32,304,600**	**9,346,400**	**-**	**24,249,000**	**226,013,700**
งานบริหารทั่วไป	75,400,000	3,312,000	81,401,700	32,304,600	9,346,400	-	24,249,000	226,013,700
แผนงานบริหารงานปกครองและทะเบียน	**23,865,200**	**432,000**	**2,935,800**	**649,500**	**-**	**1,750,000**	**1,807,500**	**31,440,000**
งานปกครองและทะเบียน	23,865,200	432,000	2,935,800	649,500	-	1,750,000	1,807,500	31,440,000
แผนงานบริหารการคลัง	**25,325,700**	**288,000**	**2,471,600**	**163,500**	**206,800**	**-**	**486,800**	**28,942,400**
งานตรวจสอบภายใน	25,325,700	288,000	2,471,600	163,500	206,800	-	486,800	28,942,400
แผนงานบริหารงานบุคคล	**66,133,100**	**1,095,000**	**15,742,600**	**95,400**	**523,100**	**-**	**5,332,300**	**88,921,500**
งานการเจ้าหน้าที่	66,133,100	1,095,000	15,742,600	95,400	523,100	-	5,332,300	88,921,500
แผนงานส่งเสริมระบบบริหาร	**174,453,600**	**4,155,600**	**43,660,400**	**4,808,800**	**5,796,800**	**980,700**	**334,276,400**	**568,132,300**
งานพัฒนาบุคลากรและองค์การ	63,561,200	2,883,500	6,397,600	2,458,000	3,464,800	-	181,009,400	259,774,500
งานกฎหมายและคดี	24,361,300	288,000	7,817,800	84,000	831,500	-	10,200,000	43,582,600

21

Problem#3 **Text**

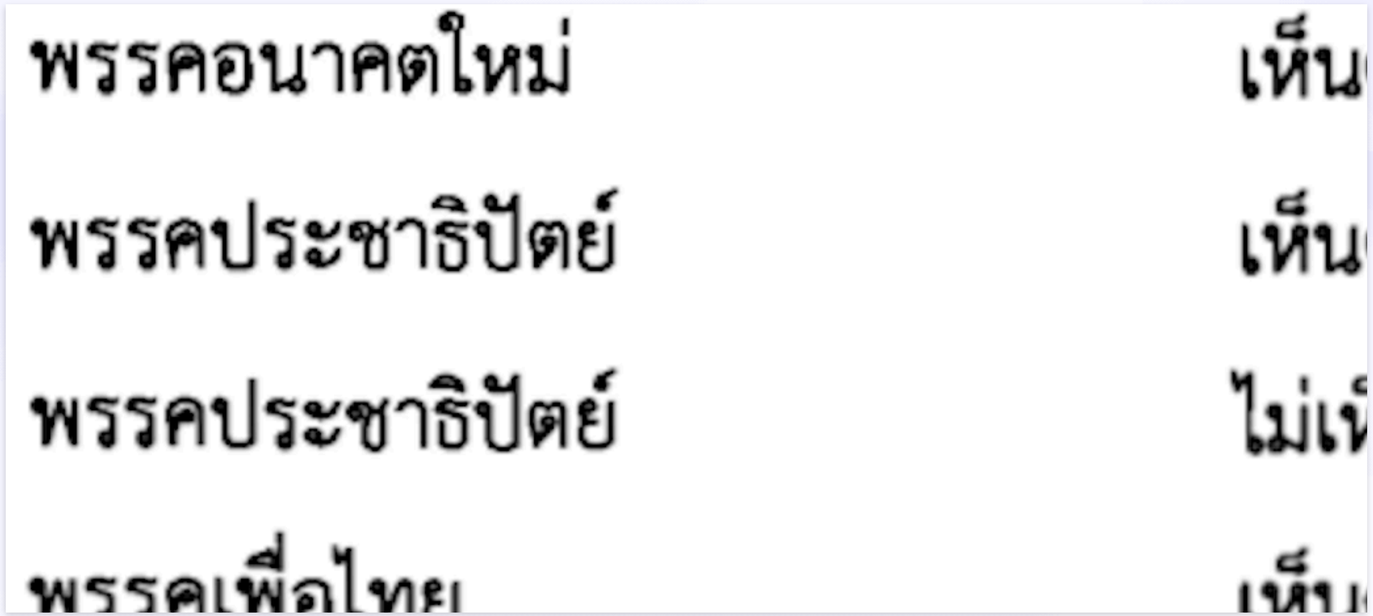


3. ค่าตอบแทน ใช้สอยและวั

3.1 ค่าตอบแทน

ค่าอาหารทำการนอกเ

เงินตอบแทนพิเศษของ

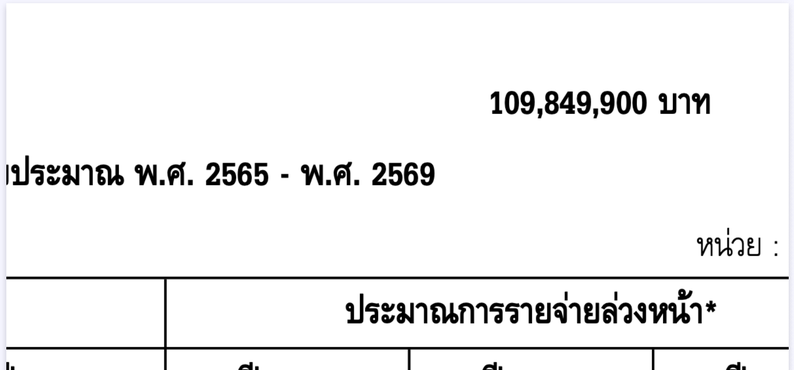


พรรคอนาคตใหม่ เห็น

พรรคประชาธิปัตย์ เห็น

พรรคประชาธิปัตย์ ไม่เห็

พรรคเพื่อไทย เห็น

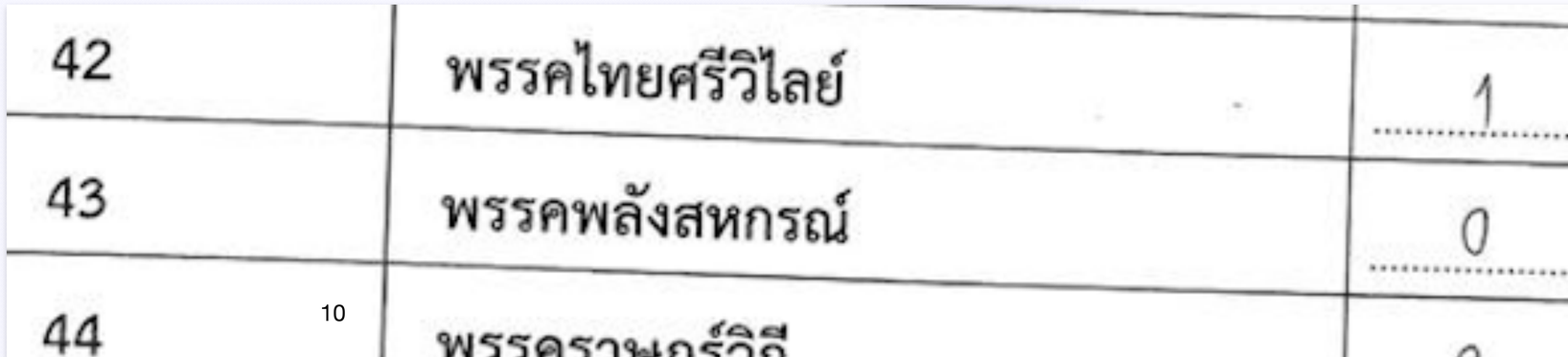


109,849,900 บาท

ประมาณ พ.ศ. 2565 - พ.ศ. 2569

หน่วย :

	ประมาณการรายจ่ายล่วงหน้า*		



42	พรรคไทยศรีวิไลย์	1
43	พรรคพลังสหกรณ์	0
44	พรรคราษฎร์วิถี	

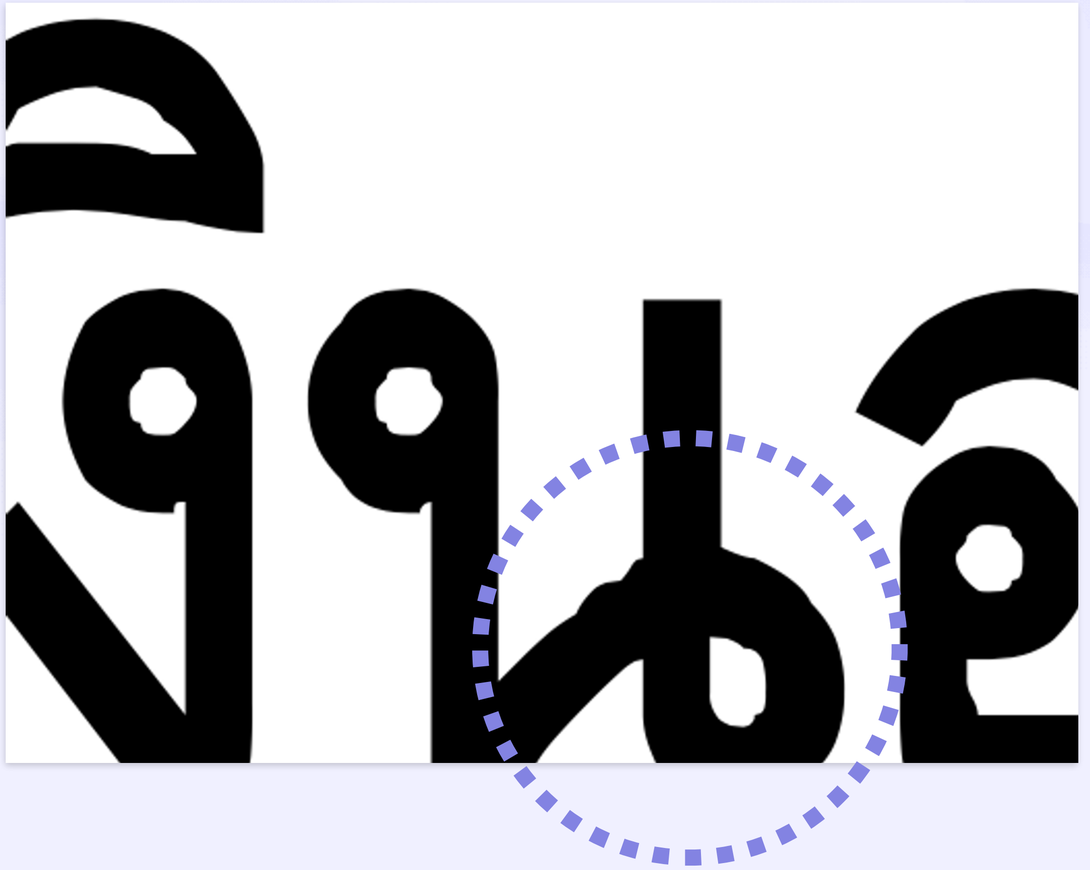
It’s not a text, it’s a shape of text!

3. ค่าตอบแทน ใช้สอยและวั

3.1 ค่าตอบแทน

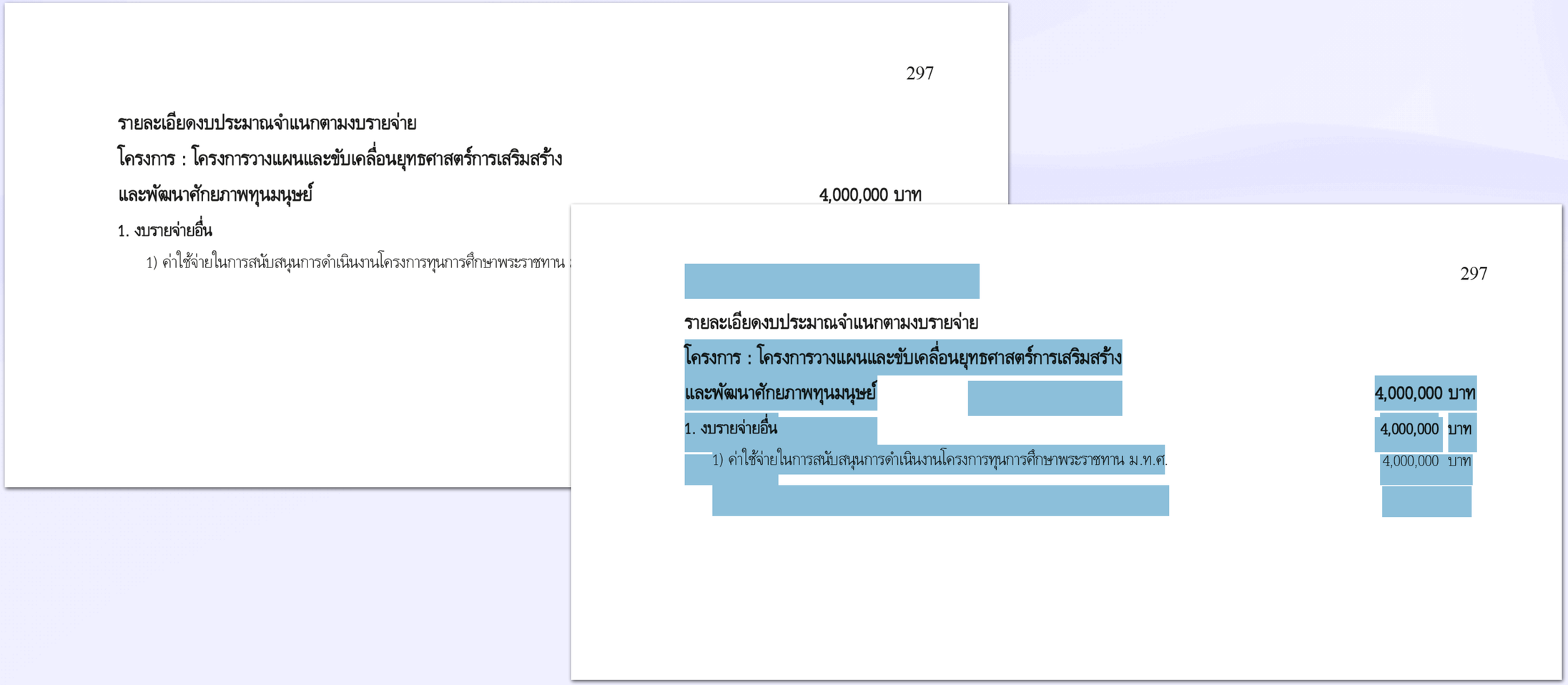
ค่าอาหารทำการนอกเ

เงินตอบแทนพิเศษของ



Problem#3 **Text**

You have more than what you see!



297

รายละเอียดงบประมาณจำแนกตามงบรายจ่าย

โครงการ : โครงการวางแผนและขับเคลื่อนยุทธศาสตร์การเสริมสร้าง

และพัฒนาศักยภาพทุนมนุษย์ **4,000,000 บาท**

1. งบรายจ่ายอื่น

1) ค่าใช้จ่ายในการสนับสนุนการดำเนินงานโครงการทุนการศึกษาพระราชทาน

297

รายละเอียดงบประมาณจำแนกตามงบรายจ่าย

โครงการ : โครงการวางแผนและขับเคลื่อนยุทธศาสตร์การเสริมสร้าง

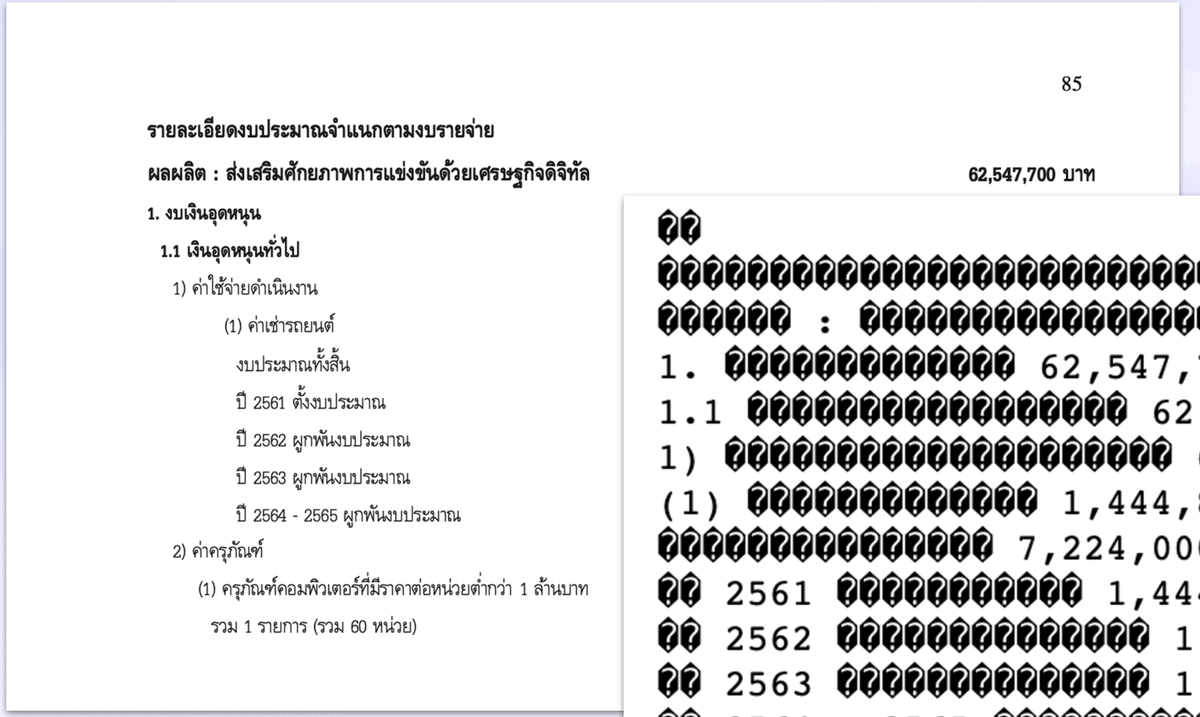
และพัฒนาศักยภาพทุนมนุษย์ **4,000,000 บาท**

1. งบรายจ่ายอื่น 4,000,000 บาท

1) ค่าใช้จ่ายในการสนับสนุนการดำเนินงานโครงการทุนการศึกษาพระราชทาน ม.ท.ศ. 4,000,000 บาท

เอกสารงบประมาณ ฉบับที่ ๓ งบประมาณรายจ่าย ประจำปีงบประมาณ พ.ศ. ๒๕๖๕ เล่มที่ ๑

Corrupted text when extracting Thai text



85

รายละเอียดงบประมาณจำแนกตามงบรายจ่าย

ผลผลิต : ส่งเสริมศักยภาพการแข่งขันด้วยเศรษฐกิจดิจิทัล **62,547,700 บาท**

1. งบเงินอุดหนุน

1.1 เงินอุดหนุนทั่วไป

1) ค่าใช้จ่ายดำเนินงาน

(1) ค่าเช่ารถยนต์

งบประมาณทั้งสิ้น

ปี 2561 ตั้งงบประมาณ

ปี 2562 ผูกพันงบประมาณ

ปี 2563 ผูกพันงบประมาณ

ปี 2564 - 2565 ผูกพันงบประมาณ

2) ค่าครุภัณฑ์

(1) ครุภัณฑ์คอมพิวเตอร์ที่มีราคาต่อหน่วยต่ำกว่า 1 ล้านบาท

รวม 1 รายการ (รวม 60 หน่วย)

```
??
??????????????????????????????????????????
??????? : ????????????????????????????????????????????????????? 62,547,700 ???
1. ????????????? 62,547,700 ???
1.1 ???????????????????? 62,547,700 ???
1) ????????????????????? 60,387,700 ???
(1) ????????????? 1,444,800 ???
????????????????? 7,224,000 ???
?? 2561 ???????????? 1,444,800 ???
?? 2562 ????????????????? 1,444,800 ???
?? 2563 ????????????????? 1,444,800 ???
?? 2564 - 2565 ?????????????????? 2,889,600 ???
2) ???????????? 2,160,000 ???
(1) ????????????????????????????????????????????????? 1 ???????? 2,160,000 ???
??? 1 ??????? (??? 60 ??????)
```

Scanned text

ลำดับที่	เลขที่บัตร	ชื่อ-สกุล
125	135	พลตรี ทรงกลด ทิพย์รัตน์
126	136	พันเอกเศรษฐพงศ์ มะลิสุวรรณ
127	137	นายทวิรัฐ รัตนเศรษฐ
128	138	นายทวีศักดิ์ ทักษิณ
129	139	นายทศพร ทองศิริ
130	140	นายทองแดง เบ็ญจะปัก
131	141	นางทัศนาพร เกษเมธีการณ

จังหวัด หนองคาย รหัสไปรษณีย์ ๔๓๑๓๐ E-mail -

๒. ประวัติการศึกษา

๒.๑ ปริญญาเอก ปรัชญาดุษฎีบัณฑิต สาขาเทคโนโลยี

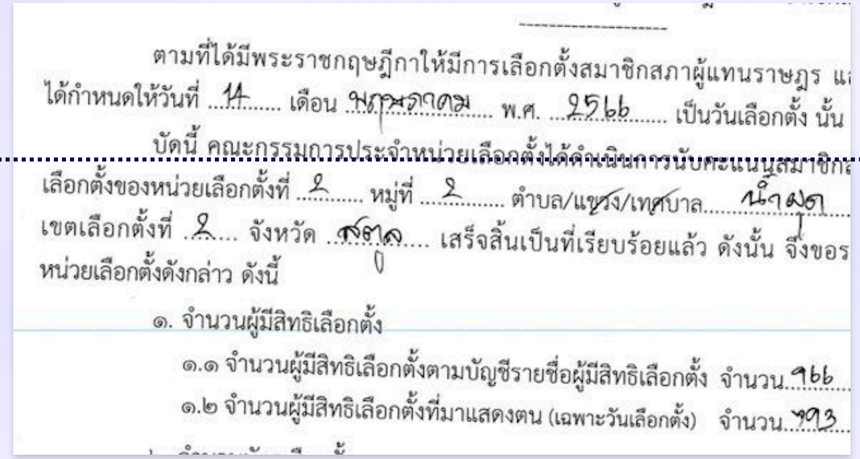
๒.๒ ปริญญาโท การศึกษามหาบัณฑิต สาขาการบริหา

๒.๓ ปริญญาตรี ศาสนศาสตร์ บัณฑิต สาขาปรัชญาศา

๓. ประวัติการทำงานหรือประสบการณ์ในการทำงานในก

41	พรรคไทยธรรม
42	พรรคไทยศรีวิไลย์
43	พรรคพลังสหกรณ์

Scanned text

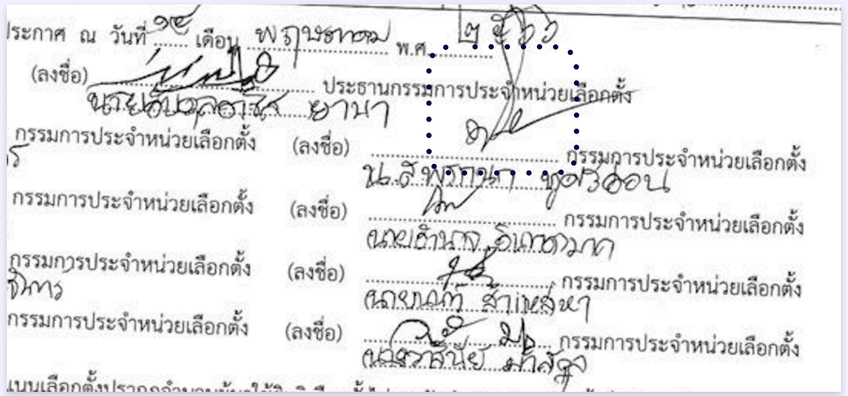


ตามที่ได้มีพระราชกฤษฎีกาให้มีการเลือกตั้งสมาชิกสภาผู้แทนราษฎร แ
ได้กำหนดให้วันที่ 14 เดือน พฤษภาคม พ.ศ. 2566 เป็นวันเลือกตั้ง นั้น
บัดนี้ คณะกรรมการประจำหน่วยเลือกตั้งได้ดำเนินการนับคะแนนผู้สมาชิก
เลือกตั้งของหน่วยเลือกตั้งที่ 2 หมู่ที่ 2 ตำบล/แขวง/เทศบาล น้ำผุด
เขตเลือกตั้งที่ 2 จังหวัด สตูล เสร็จสิ้นเป็นที่เรียบร้อยแล้ว ดังนั้น จึงขอร
หน่วยเลือกตั้งดังกล่าว ดังนี้

๑. จำนวนผู้มีสิทธิเลือกตั้ง

๑.๑ จำนวนผู้มีสิทธิเลือกตั้งตามบัญชีรายชื่อผู้มีสิทธิเลือกตั้ง จำนวน 966

๑.๒ จำนวนผู้มีสิทธิเลือกตั้งที่มาแสดงตน (เฉพาะวันเลือกตั้ง) จำนวน 793



ประกาศ ณ วันที่ 14 เดือน พฤษภาคม พ.ศ. 2566

(ลงชื่อ) ..... ประธานกรรมการประจำหน่วยเลือกตั้ง

นายอับดุลอาซิส ยามา

กรรมการประจำหน่วยเลือกตั้ง (ลงชื่อ) ..... กรรมการประจำหน่วยเลือกตั้ง

น.ส.พิมพกานา ชูอ่วยอน

กรรมการประจำหน่วยเลือกตั้ง (ลงชื่อ) ..... กรรมการประจำหน่วยเลือกตั้ง

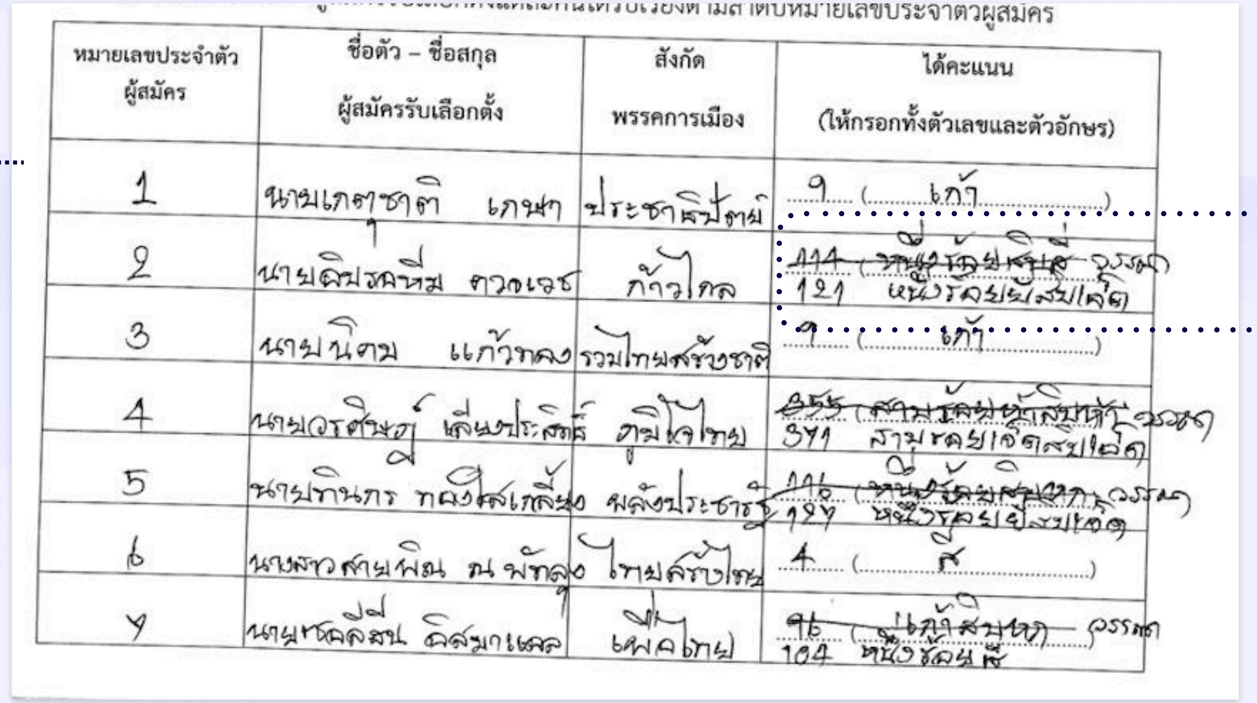
นายอิบนาฮาร์ อินทรวาหา

กรรมการประจำหน่วยเลือกตั้ง (ลงชื่อ) ..... กรรมการประจำหน่วยเลือกตั้ง

ฉลาดนัท สาเหล็มหา

กรรมการประจำหน่วยเลือกตั้ง (ลงชื่อ) ..... กรรมการประจำหน่วยเลือกตั้ง

นายอิสรา ศาสัฐ



หมายเลขประจำตัวผู้สมัคร	ชื่อตัว – ชื่อสกุล ผู้สมัครรับเลือกตั้ง	สังกัดพรรคการเมือง	ได้คะแนน (ให้กรอกทั้งตัวเลขและตัวอักษร)
1	นายเกรียงไกร เกษมา	ประชาธิปัตย์	9 (เก้า)
2	นายอิสมาอีล ดาวเรือง	ก้าวไกล	111 (หนึ่งร้อยสิบเอ็ด) 121 หนึ่งร้อยยี่สิบเอ็ด
3	นายนิคม แก้วทอง	รวมไทยสร้างชาติ	9 (เก้า)
4	นายวรศิษฏ์ เลิศประสิทธิ์	ภูมิใจไทย	355 (สามร้อยห้าสิบห้า) 371 สามร้อยเจ็ดสิบเอ็ด
5	นายทินกร ทองสุดเกลี้ยง	พลังประชารัฐ	111 (หนึ่งร้อยสิบเอ็ด) 124 หนึ่งร้อยยี่สิบสี่
6	นางสาวสายพิณ ณ พัทลุง	ไทยสร้างไทย	4 (สี่)
7	นายหลีสัน ลิสมาแลล	เพื่อไทย	96 (เก้าสิบหก) 104 หนึ่งร้อยสี่

??

Possible solutions

LayoutParser

https://github.com/Layout-Parser/layout-parser

ure 7: **Annotation Examples in HJDataset.** (a) and (b) show two examples for the labeling of `main` pages. The boxes are colored differently to reflect the layout element categories. Illustrated in (c), the items in each `index` page row are categorized as title blocks, and the annotations are denser.

over union (IOU) level [0.50:0.95][2], on the test data. In general, the high mAP values indicate accurate detection of the layout elements. The Faster R-CNN and Mask R-CNN achieve comparable results, better than RetinaNet. Noticeably, the detections for small blocks like title are less precise, and the accuracy drops sharply for the title category. In Figure 8, (a) and (b) illustrate the accurate prediction results of the Faster R-CNN model.

Pre-training for other datasets

We also examine how our dataset can help with a real-world document digitization application. When digitizing new publications, researchers usually do not generate large scale ground truth data to train their layout analysis models. If they are able to adapt our dataset, or models trained on our dataset, to develop models on their data, they can build their pipelines more efficiently and develop more accurate models. To this end, we conduct two experiments. First we examine how layout analysis models trained on the `main` pages can be used for understanding `index` pages. Moreover, we study how the pre-trained models perform on other historical Japanese documents.

able 4 compares the performance of five Faster R-CNN models that are trained differently on `index` pages. If the model loads pre-trained weights from HJDataset, it includes information learned from `main` pages. Models trained over

his is a core metric developed for the COCO competition [12] for evaluating the object detection quality.

he training data can be viewed as the benchmarks, while training with few samples (five in this case) are considered to mimic real-world scenarios. Given different training data, models pre-trained on HJDataset perform significantly better than those initialized with COCO weights. Intuitively, models trained on more data perform better than those with fewer samples. We also directly use the model trained on `main` to predict `index` pages without fine-tuning. The low zero-shot prediction accuracy indicates the dissimilarity between `index` and `main` pages. The large increase in mAP from 0.344 to 0.471 after the model is

le 3: Detection mAP @ IOU [0.50:0.95] of different models for each category on the test set. All values are given as percentages.

Category	Faster R-CNN	Mask R-CNN[a]	RetinaNet
Page Frame	99.046	99.097	99.038
Row	98.831	98.482	95.067
Title Region	87.571	89.483	69.593
Text Region	94.463	86.798	89.531
Title	65.908	71.517	72.566
Subtitle	84.093	84.174	85.865
Other	44.023	39.849	14.371
mAP	81.991	81.343	75.223

training Mask R-CNN, the segmentation masks are the quadrilateral regions for each block. Compared to the rectangular bounding boxes, they delineate the text region more accurately.

Region Types

- *Text*
- *Title*
- *Figure*
- *Table*

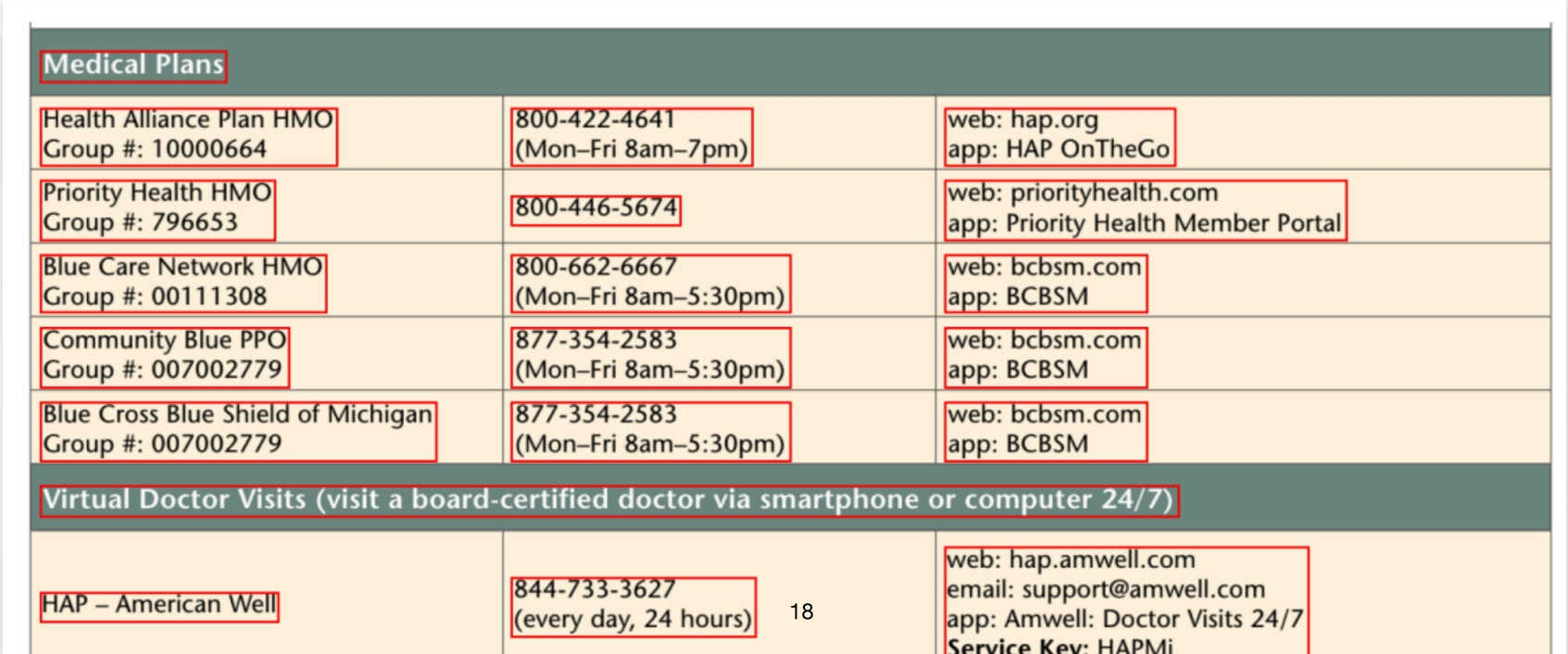
UniTable

https://github.com/poloclub/unitable

[bbox] : position of the cell

[html] : ["<thead>", "<tr>", "<td>[", "]</td>", "<td>[", “]</td>”, …

[cell] : text in cell



Medical Plans		
Health Alliance Plan HMO Group #: 10000664	800-422-4641 (Mon–Fri 8am–7pm)	web: hap.org app: HAP OnTheGo
Priority Health HMO Group #: 796653	800-446-5674	web: priorityhealth.com app: Priority Health Member Portal
Blue Care Network HMO Group #: 00111308	800-662-6667 (Mon–Fri 8am–5:30pm)	web: bcbsm.com app: BCBSM
Community Blue PPO Group #: 007002779	877-354-2583 (Mon–Fri 8am–5:30pm)	web: bcbsm.com app: BCBSM
Blue Cross Blue Shield of Michigan Group #: 007002779	877-354-2583 (Mon–Fri 8am–5:30pm)	web: bcbsm.com app: BCBSM
Virtual Doctor Visits (visit a board-certified doctor via smartphone or computer 24/7)		
HAP – American Well	844-733-3627 (every day, 24 hours)	web: hap.amwell.com email: support@amwell.com app: Amwell: Doctor Visits 24/7 Service Key: HAPMi

UniTable

https://github.com/poloclub/unitable

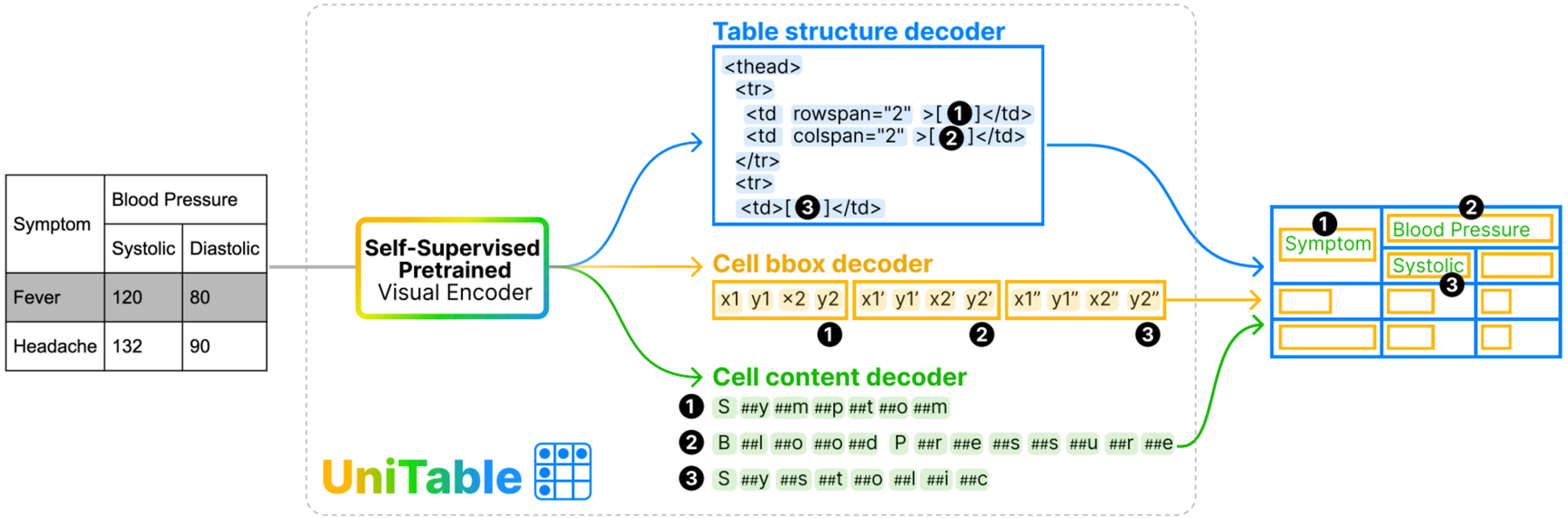
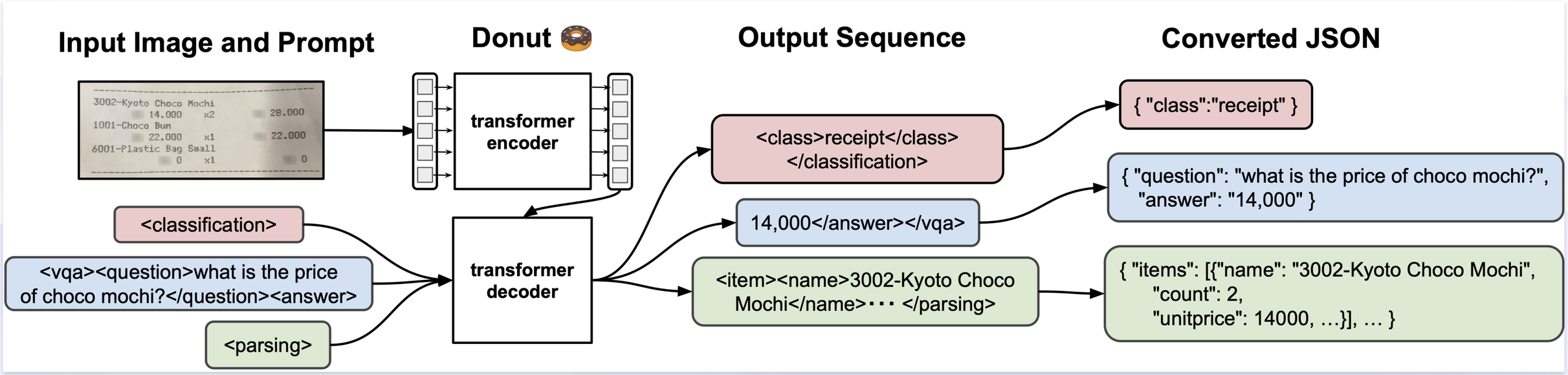


Figure 1: **UniTable**, a training framework that unifies both **training paradigm** and **training objective** of TR. In UniTable, the visual encoder is self-supervised pretrained and then finetuned along with the task decoder on supervised datasets. UniTable unifies the training objectives of all three TR tasks — extracting table structure, cell bbox, and cell content — into a unified task-agnostic training objective: language modeling. With UniTable, the user inputs a tabular image and obtains the corresponding digitalized table in HTML.

DONUT: Document Understanding Transformer

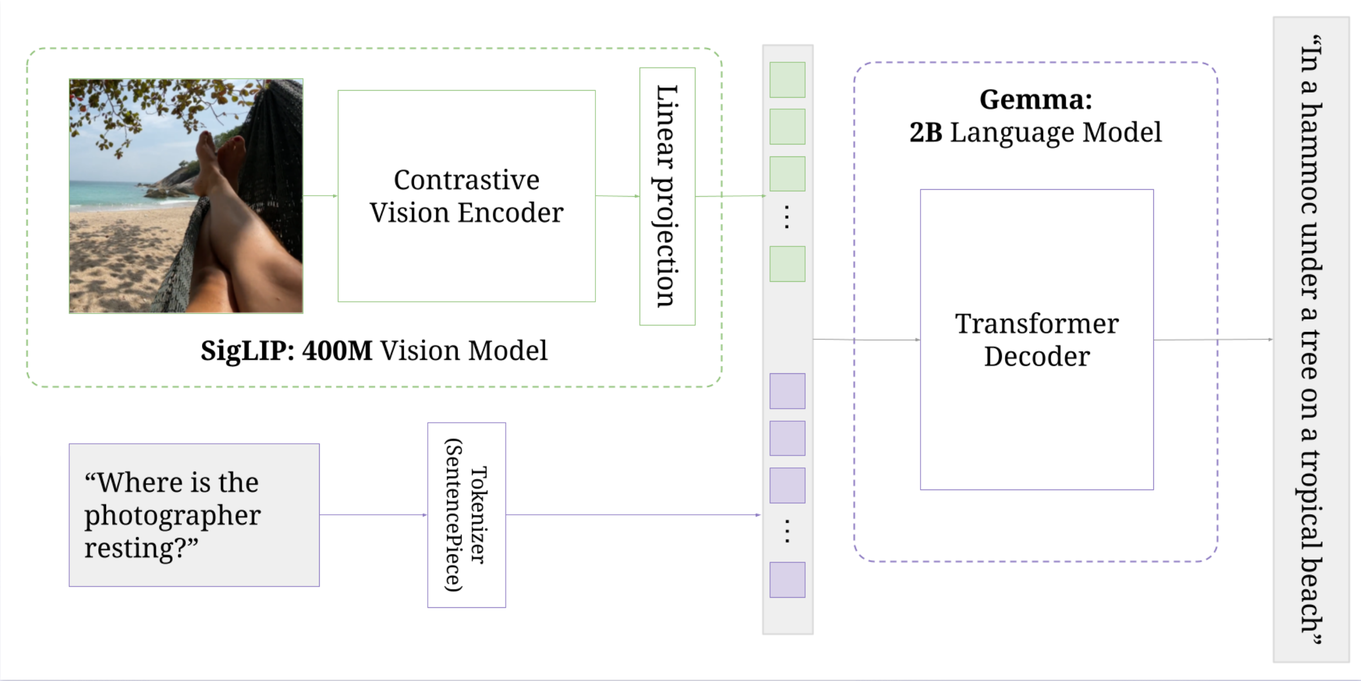
https://github.com/clovaai/donut



Donut 🍩, Document understanding transformer, is a new method of document understanding that utilizes an OCR-free end-to-end Transformer model. Donut does not require off-the-shelf OCR engines/APIs, yet it shows state-of-the-art performances on various visual document understanding tasks, such as visual document classification or information extraction (a.k.a. document parsing).

PaliGemma

https://huggingface.co/blog/paligemma



PaliGemma-3B is Vision-Language model that was inspired by the PaLI-3 recipe. It is built on SigLIP visual encoder (specifically, SigLIP-So400m/14) and the Gemma 2B language model.

- **Image Captioning**
- **Visual Question Answering**
- **Object detection** using “detect” prefix

  <loc0023> <loc0011> <loc0013> <loc0120>
- **Object segmentation** “segment”
- **Document Understanding**

PDF trauma is real.