+

Data Scraping Demo

Scrape External Data

- Demo1: Scrap time series data
- Demo2: Scraping news data
- Demo3: Information extraction
- Demo4: Question answering

matplotlib

- Scrapy is a free and open source web crawling framework, written in Python. Originally designed for web scraping, extracting data from websites which is the general purpose of web crawler.
- Cross platform, Fast and Powerful
- Scrapy allows specifying which web page's elements would be scraped by using XPath.

Demo1: Scrap Time Series Data

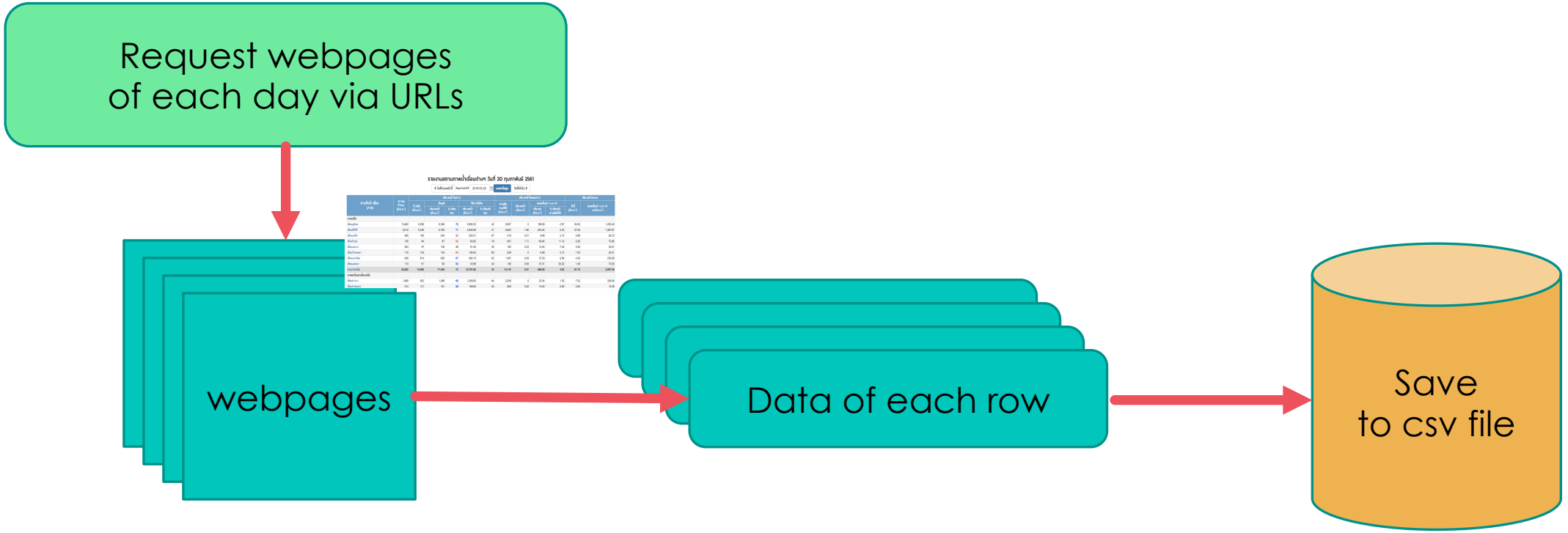
- From website http://www.thaiwater.net/DATA/REPORT/php/rid_bigcm.php reporting daily data of each dam in Thailand.

รายงานสถานภาพน้ำเขื่อนต่างๆ วันที่ 20 กุมภาพันธ์ 2561

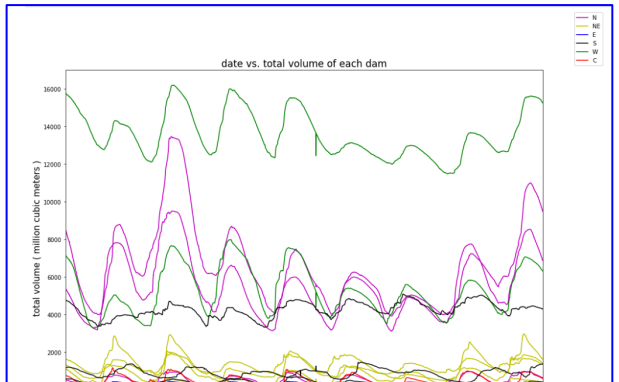
< วันที่ก่อนหน้านี้ | ข้อมูลล่าสุดวันที่ 2018-02-20 | แสดงข้อมูล | วันที่ถัดไป >

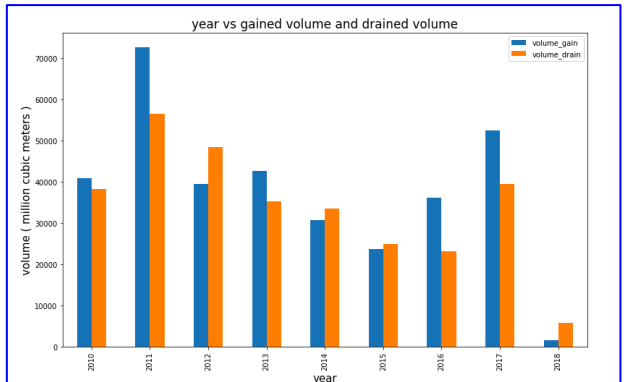
อ่างเก็บน้ำ เขื่อน (ภาค)	ความจุ ที่ รนก. (ล้าน ม.^{3})	ปริมาตรน้ำในอ่างฯ					ปริมาตรน้ำไหลลงอ่างฯ				ปริมาณน้ำระบาย	
		ปี 2560 (ล้าน ม.^{3})	ปัจจุบัน		ใช้การได้จริง		ค่าเฉลี่ย รวมทั้งปี (ล้าน ม.^{3})	ปริมาตรน้ำ (ล้าน ม.^{3})	สะสมตั้งแต่ 1 ม.ค. 61		วันนี้ (ล้าน ม.^{3})	สะสมตั้งแต่ 1 ม.ค. 61 ม.(ล้าน ม.^{3})
			ปริมาตรน้ำ (ล้าน ม.^{3})	% เทียบ รนก.	ปริมาตรน้ำ (ล้าน ม.^{3})	% เทียบกับ รนก.			ปริมาตร (ล้าน ม.^{3})	% เทียบกับ ค่าเฉลี่ยทั้งปี		
ภาคเหนือ												
เขื่อนภูมิพล	13,462	6,536	9,390	70	5,590.23	42	5,607	0	166.80	2.97	30.00	1,250.40
เขื่อนสิริกิติ์	9,510	6,239	6,793	71	3,942.66	41	5,683	1.66	240.24	4.23	27.08	1,367.97
เขื่อนแม่งัด	265	165	243	92	230.31	87	319	0.01	6.86	2.15	0.68	26.72
เขื่อนกิ่วลม	106	84	87	82	83.92	79	547	1.71	60.94	11.14	2.26	72.96
เขื่อนแม่กวง	263	97	106	40	91.84	35	190	0.23	14.24	7.49	0.36	35.91
เขื่อนกิ่วคอหมา	170	153	143	84	136.52	80	229	0	4.86	2.12	1.00	33.51
เขื่อนแควน้อย	939	614	625	67	582.15	62	1,387	0.93	37.25	2.69	4.32	235.98
เขื่อนแม่มอก	110	61	60	54	43.99	40	148	0.83	37.31	25.22	1.49	73.93
รวมภาคเหนือ	**24,825**	**13,950**	**17,446**	**70**	**10,701.62**	**43**	**14,110**	**5.37**	**568.50**	**4.03**	**67.19**	**3,097.38**
ภาคตะวันออกเฉียงเหนือ												
เขื่อนลำปาว	1,980	832	1,366	69	1,265.60	64	2,258	0	23.34	1.03	7.02	328.56
เขื่อนลำตะคอง	314	101	181	58	158.64	50	269	0.22	15.04	5.59	0.30	15.38

Demo1: Scrap Time Series Data (cont.)




```
date,region,name,capacity,volume_total,volume_usable,volume_gain,volume_drain
2010-01-03,N,เขื่อนภูมิพล,13462,8454,4654.00,0,18.00
2010-01-03,N,เขื่อนสิริกิติ์,9510,5361,2511.00,7.05,20.07
2010-01-03,N,เขื่อนแม่งัด,265,229,207.00,0.12,0.35
2010-01-03,N,เขื่อนกิ่วลม,112,86,82.00,0.44,0.12
2010-01-03,N,เขื่อนแม่กวง,263,80,66.00,0.12,0.03
2010-01-03,N,เขื่อนกิ่วคอหมา,170,123,116.00,0.04,0.25
2010-01-03,N,เขื่อนแควน้อย,769,534,498.00,0.27,6.05
2010-01-03,NE,เขื่อนลำปาว,1430,1131,1046.00,0.69,4.89
2010-01-03,NE,เขื่อนลำตะคอง,314,179,151.00,0.60,0.01
2010-01-03,NE,เขื่อนลำพระเพลิง,110,57,56.00,0,0.13
2010-01-03,NE,เขื่อนน้ำอูน,520,222,179.00,0,0.79
2010-01-03,NE,เขื่อนอุบลรัตน์,2264,1336,754.00,0,3.97
2010-01-03,NE,เขื่อนสิรินธร,1966,1628,797.00,0,0
```

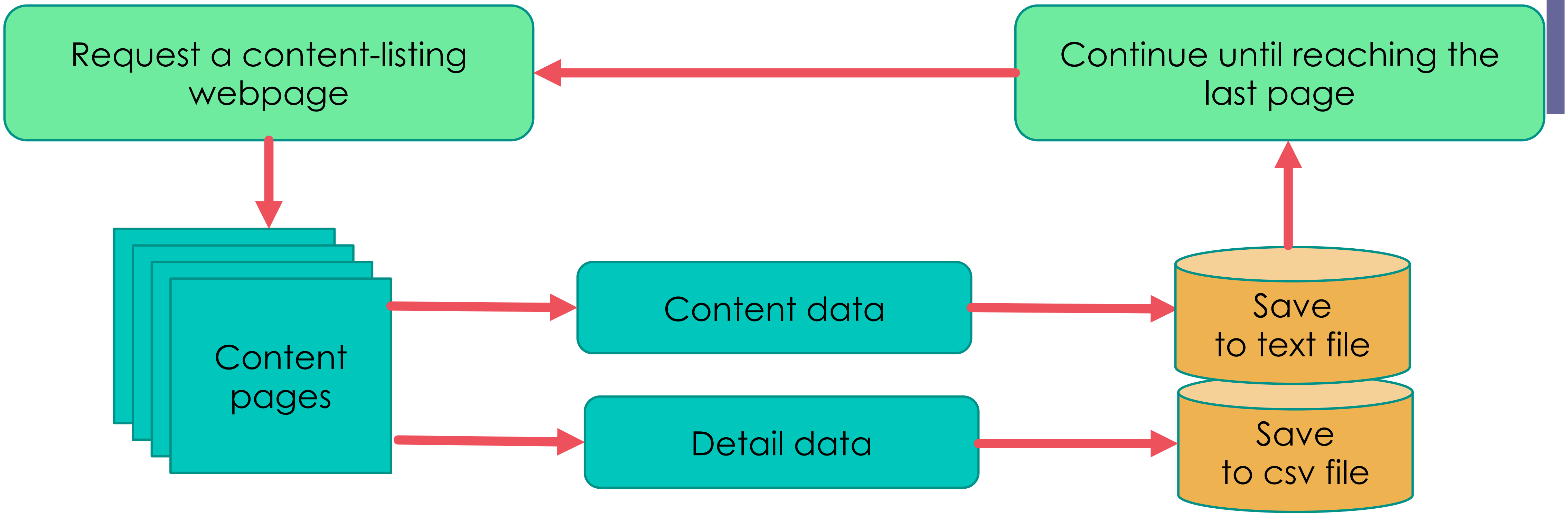




Demo2: Scraping news data

- Scraping the news contents from https://news.mthai.com/economy-news

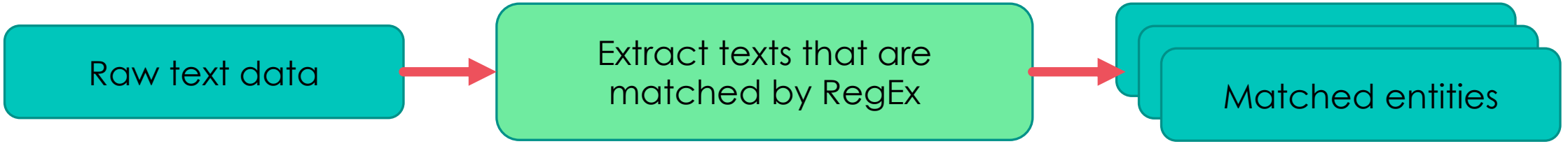
Demo2: Scraping news data (cont.)




```
path,id,author,title,date
output\619203.txt,619203,ร รุ้ง,คลังหนุนเซเว่นฯ เป็นตัวแทนแบงก์ เปิดบริการรับฝาก-ถอนเงิน,2018-02-18
output\619322.txt,619322,จรัญญา,คืบหน้าโปรเจคยักษ์ 6 หมื่นล้าน รถไฟฟ้าสายสีน้ำตาล แคราย-ลำสาลี,2018-02-19
output\619435.txt,619435,Lheam-Thong,MONO29  จัดหนัก คืนชีพป้ายโฆษณายักษ์ สู่อีสานและภาคตะวันออก,2018-02-19
output\619264.txt,619264,จรัญญา,ทอง เปิดตลาดวันนี้ ปรับขึ้น 50 บาท,2018-02-19
output\619287.txt,619287,จรัญญา,นักวิเคราะห์มอง หุ้นไทยวันนี้ขึ้นต่อ แรงขายจากต่างชาติเริ่มชะลอ,2018-02-19
output\619575.txt,619575,จรัญญา,รถจักรยานยนต์ ราคาจ่อขยับ สรรพสามิตเล็งเก็บภาษีเพิ่ม,2018-02-20
output\619549.txt,619549,จรัญญา,"ทอง ปรับลง 50 บาท รูปพรรณขาย 20,500 บาท",2018-02-20
output\619709.txt,619709,จรัญญา,"ครม. ไฟเขียวงบ 5,547 ล้าน ดัน 19 โปรเจคหนุนเศรษฐกิจในประเทศ",2018-02-20
output\619837.txt,619837,จรัญญา,ทอง เปิดตลาดวันนี้ ร่วง 100 บาท,2018-02-21
```

Demo3: Information extraction (IE) Extraction company entities

- Using Regular Expression patterns to extract entities from the news data



โดยบริษัทที่ทำรายได้สูงสุดอันดับที่ 1 อยู่ที่จังหวัดสงขลา ได้แก่ ตระกูล “สินเจริญกุล” ผู้ถือหุ้นใหญ่บริษัท ศรีตรังแอโกรอินดัสทรี จำกัด (มหาชน) หรือ STA ผู้ประกอบการยางธรรมชาติแบบครบวงจรเบอร์ 1 ของโลก มีรายได้รวมในปี 2559 ถึง 38,950.26 ล้านบาท

ขณะที่อันดับ 2 เป็นของตระกูล เกิดวงศ์บัณฑิต จังหวัดภูเก็ต เจ้าของ บจ.วงศ์บัณฑิต มีรายได้ 25,297.64 ล้านบาท

อันดับ 3 ตระกูลธีรศานต์วงศ์ จังวหัดสงขลา เจ้าของ บจ.เซาท์แลนด์รีสอร์ต มีรายได้ 24,480.80 ล้านบาท

อันดับ 4 ตระกูลสุริยวนากุล จ.ร้อยเอ็ด เจ้าของ บมจ.สยามโกลบอลเฮ้าส์ มีรายได้ 19,474 ล้านบาท

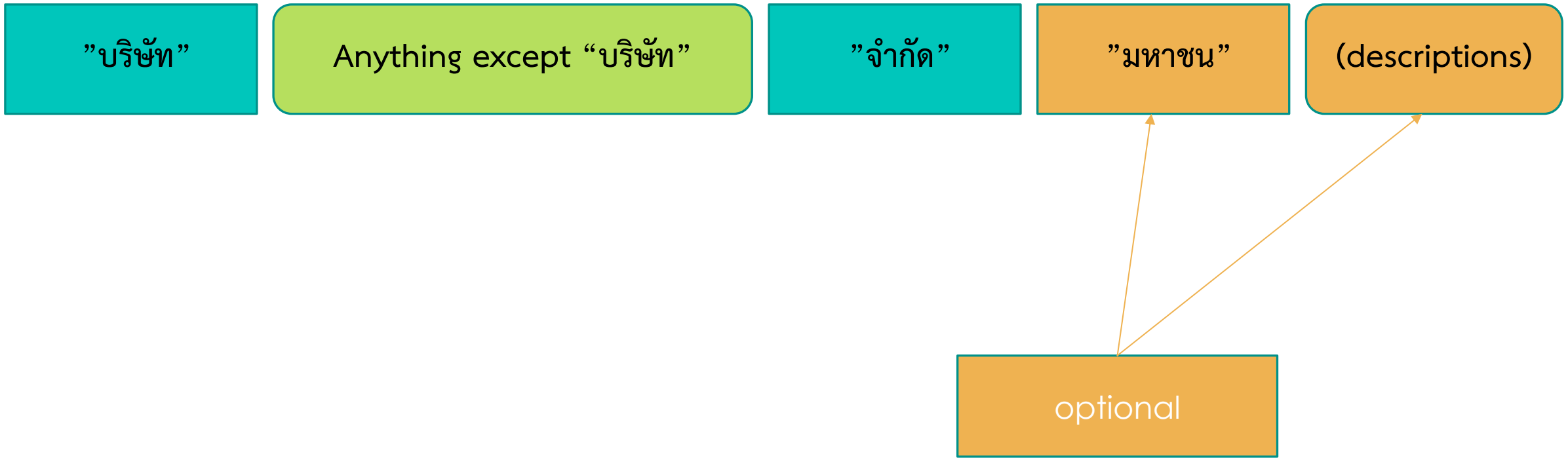
อันดับ 5 ตระกูล “เฉลิมวุฒินันท์” เจ้าของบริษัท เอเชีย โกลเด้น ไรซ์ จำกัด ซึ่งถือเป็นผู้ส่งออกข้าวรายใหญ่ที่สุดของประเทศ มีรายได้รวม 19,275.96 ล้านบาท

```
ปตท.: บริษัท ปตท.จำกัด (มหาชน), บริษัท ปตท. จำกัด(มหาชน), บริษัท ปตท. จำกัด (มหาชน), บริษัท ปตท.
(มหาชน), บริษัท ปตท.จำกัด (มหาชน), บริษัท ปตท.จำกัด (มหาชน), บริษัท ปตท. จำกัด (มหาชน) ...

บางจาก คอร์ปอเรชั่น: บริษัท บางจาก คอร์ปอเรชั่น จำกัด (มหาชน), บริษัทบางจาก คอร์ปอเรชั่น จำกัด(มหาชน), บริษัทบางจาก คอ

สยามพาร์คซิตี้: บริษัทสยามพาร์คซิตี้ จำกัด

เอกชัยการแพทย์: บริษัท เอกชัยการแพทย์ จำกัด (มหาชน)

เอกชัย อินเตอร์เนชั่นแนล: บริษัท เอกชัย อินเตอร์เนชั่นแนล จำกัด

หลักทรัพย์กสิกรไทย: บริษัทหลักทรัพย์กสิกรไทย จำกัด, บริษัทหลักทรัพย์กสิกรไทย จำกัด (มหาชน), บริษัทหลักทรัพย์กสิกรไทย จำกัด
```

```
ปตท.: [617895, 618598, 616864, 615448, 611637, 609300, 607635, 605718, 60
592125, 590461, 590838, 589888, 589472, 587693, 587129, 584451, 585345,
1 ...

บางจาก คอร์ปอเรชั่น: [617895, 618598, 616864, 607184, 600547, 597933, 595474, 59
574017, 572511, 569757, 568888, 567550, 565657, 563694, 562053, 562467,
```

Demo3: Information extraction (cont.) Extraction company entities

```
r'บริษัท(?P<key_0>(?:.(?!บริษัท|ที่|หรือ|ใน))+?)จำกัด(?:\s*\(\s*มหาชน\s*\))?(?:\s*\(\s*.{2,20}\s*\))?'
```



Demo4: Facebook Graph API

https://developers.facebook.com/docs/graph-api/