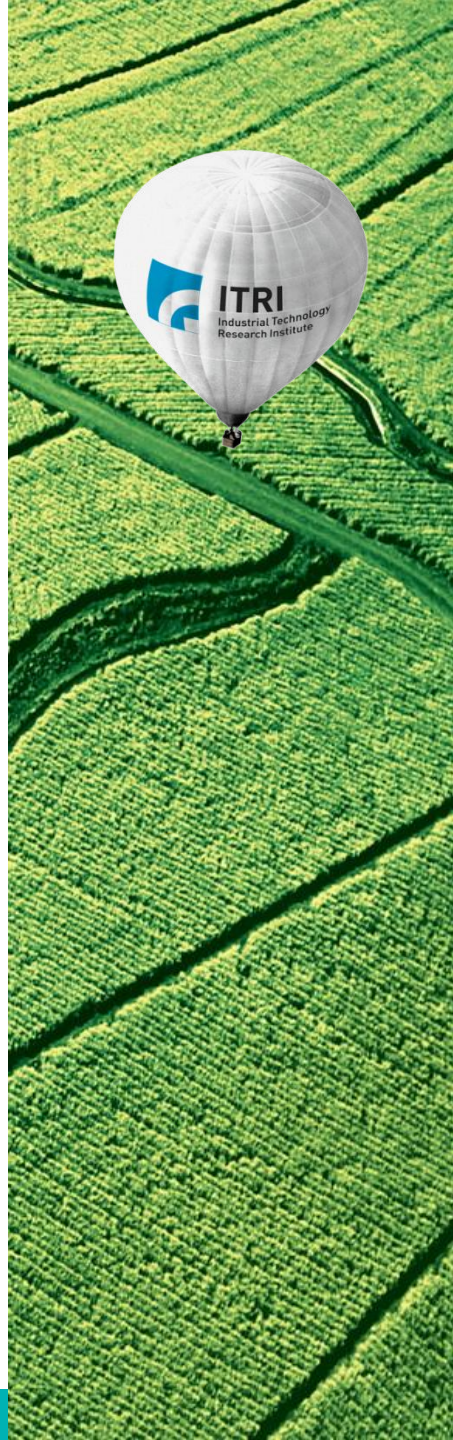
工業技術研究院
Industrial Technology
Research Institute

Construction of a Traditional Chinese Table Image and QA Dataset



Institute of Information and Communications W400 – Intern: SHU-CHI LIU

2024/12/20

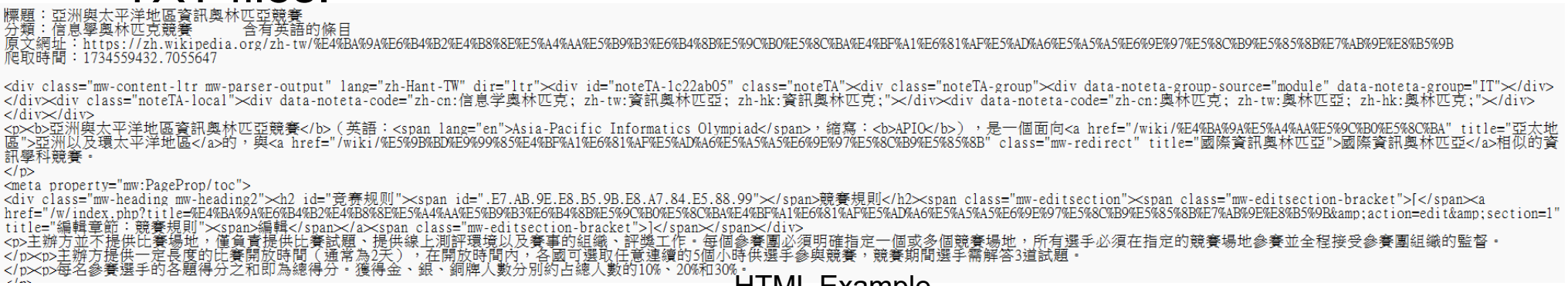
Project Steps

- 1.Web Crawling
  - Collect high-quality Traditional Chinese web pages containing tables.
- 2.Table Extraction
  - Extract tables and related information (titles, descriptions) from HTML pages
  - Save them as JSON and CSV files.
- 3.Data Cleaning and Special Processing
  - Handle special cases such as merged-cell tables
  - Remove low-quality tables.
- 4.Table Visualization
  - Render tables in a browser and convert them into image files.
- 5.Question-Answer Text Generation
  - Use an LLM to generate relevant questions and answers based on the tables.

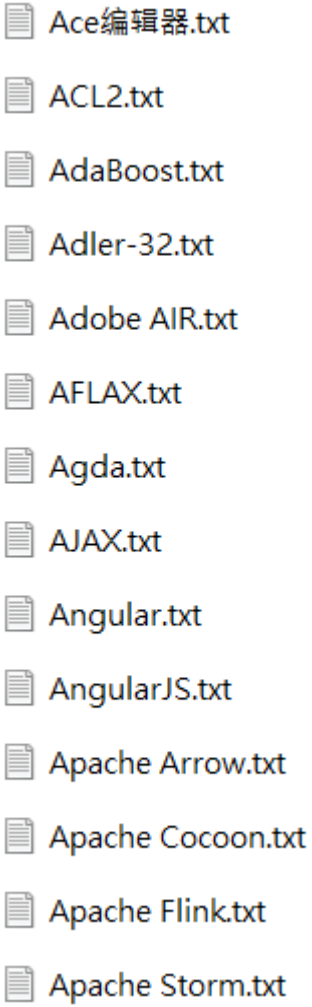
工業技術研究院
Industrial Technology Research Institute

Web Crawling

- High-quality Traditional Chinese table data: Traditional Chinese Wikipedia (https://zh.wikipedia.org/zh-tw/)
- Web scraping tool:
  - GitHub: Traditional Chinese Wikipedia Data Scraping and Preprocessing (https://github.com/wjn1996/scrapy_for_zh_wiki)
- Usage:
  - Provide a Wikipedia category page as the entry point.
  - Set filter keywords to limit scraping to specific topics.
- Outcome:Collected 10,000 Wikipedia pages' HTML, saved as TXT files.



HTML Example



List of scraped TXT files

工業技術研究院
Industrial Technology Research Institute

Table Extraction – Identifying Required Tables

- Traditional Chinese Wikipedia contains four types of tables. Ultimately, only WikiTables were selected for extraction because their structure is more standardized, they are easier to extract along with related descriptions, and they have higher relevance to the page's main topic.

Infobox vevent

誰是接班人 The Apprentice	
類型	實境電視影集
開創	馬克·伯內特
主持	唐納·川普
主演	唐納·川普 喬治·H·羅斯 卡羅琳·凱普徹 比爾·蘭西克 伊凡卡·川普 小唐納·川普 艾瑞克·川普 尚恩·亞茲貝克
國家／地區	美國
季數	15
集數	192
每集長度	60分鐘（第1–7季、第10季） 120分鐘（第8–9、11–15季）
配樂	肯尼斯·甘布爾 里昂·霍夫 安東尼·傑克森
片頭曲	歐傑斯合唱團《人為財死》

Sidebar

生物分類學

分類階元

域・界・門・綱
目・科・屬・種

三域系統

細胞生物（古菌域・細菌域・真核域）

其他話題

二名法・「非細胞生物」（病毒）・生物演化樹

閱・論・編

WikiTabe

舉例 [編輯]

常見的五種細胞生物（果蠅、人、豌豆、釀酒酵母和大腸桿菌）的名稱和分類如下：

中文	英文	拉丁文 (單數, 複數)	果蠅	人	豌豆	釀酒酵母	大腸桿菌
域；總界	domain; superkingdom	regio, regiones	真核域 Eukarya	真核域 Eukarya	真核域 Eukarya	真核域 Eukarya	細菌域 Bacteria
界	kingdom	regnum, regna	動物界 Animalia	動物界 Animalia	植物界 Plantae	真菌界 Fungi	細菌界 Bacteria
門	division（植物）; phylum（動物）	divisio, divisiones; phylum, phyla	節肢動物門 Arthropoda	脊索動物門 Chordata	被子植物門 Magnoliophyta	子囊菌門 Ascomycota	變形菌門 Proteobacteria
亞門	subdivision; subphylum	subdivisio, subdivisiones; subphylum, subphyla	六足亞門 Hexapoda	脊椎動物亞門 Vertebrata			
綱	class	classis,	昆蟲綱	哺乳綱	雙子葉植物綱	酵母綱	γ-變形菌綱

外部連結 [編輯]

- （英文）PhyloCode
- （英文）NCBI分類（頁面存檔備份，存於網際網路檔案館）
- （英文）國際動物命名法規
- （英文）國際植物命名法規，2000年聖路易斯（頁面存檔備份，存於網際網路檔案館）
- 《細菌名稱》

Navbox

開・論・編	分類階元								[合併]
病毒域（病毒專用）			高綱		高目	派（動物專用）			
病毒亞域（病毒專用）	總門／超門	總綱／超綱		總目／超目	總科／超科	總族／超族			總種／超種
域／總界／超界	門	綱	部	目	科	族	屬		種
界	亞門	亞綱	群	亞目	亞科	亞族	亞屬		亞種
亞界	下門	下綱		下目	下科	下族	節／組（植物專用）	變種（植物專用）	
下界／支	小門	小綱		小目			系（植物專用）	變型（植物專用）	
開・論・編	演化生物學								[展開]
開・論・編	生物學								[展開]

工業技術研究院
Industrial Technology Research Institute

Table Extraction – Extract and Save as JSON and CSV

- Extraction: Use XPath to locate and extract WikiTables within the HTML.
- JSON Format
  - Entity: Wikipedia page title
  - Category: Wikipedia page category
  - Url: Wikipedia page URL
  - Table_id: Unique identifier
  - Table_size: rows*columns
  - Description: Section title where the table is located
  - Data: Table content
- CSV: Store only the table content in tabular CSV format


```
"entity": "亞洲與太平洋地區資訊奧林匹亞競賽",
"category": "信息學奧林匹克競賽, 含有英語的條目",
"url": "https://zh.wikipedia.org/zh-tw/%E4%
 %E5%85%8B%E7%AB%9E%E8%B5%9B",
"table_id": 6,
"table_size": "17*3",
"is_complex_table": false,
"description": "中國區歷屆比賽概況",
"data": [
    {
        "年份": "2007",
        "承辦單位": "中國人民大學",
        "比賽日期": "5月12日"
    },
    {
        "年份": "2008",
        "承辦單位": "東北大學",
        "比賽日期": "5月10日"
    },
    {
        "年份": "2009",
        "承辦單位": "天津大學",
        "比賽日期": "5月9日"
    },
```


JSON Data Example

	A	B	C	D	E	F	G	H	I	J	K	L
1		搭檔	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012
2	女子單打	—	64強	—	64強	—	64強	第一輪(6	—	—	—	小組第二名
3	女子雙打	Sara Jonsd	—	—	64強	—	—	—	—	—	—	—

CSV Data Example

工業技術研究院
Industrial Technology Research Institute

Special Data Processing

- Special Data Processing – Merged-Cell Tables:
  - Merged-cell tables refer to structures where a single cell spans across multiple columns or rows.
  - In HTML, this is represented using the colspan or rowspan attributes.
  - In JSON and CSV, each row corresponds to one record with a fixed number of columns, which cannot directly represent merged cells.
  - Processing Method: Replicate the spanned value across all corresponding cells until the structure is normalized.

年份	賽事	公開賽級別	項目	搭檔	成績
2009年	賽普勒斯羽球國際賽	國際系列賽	混合雙打	Helgi Johannesson	準決賽
	冰島羽球國際賽	國際系列賽	女子單打	–	冠軍
			女子雙打	Snjólaug Jóhannsdóttir	冠軍
2010年	賽普勒斯羽球國際賽	國際系列賽	女子單打	–	準決賽
	冰島羽球國際賽	未來系列賽	女子單打	–	冠軍
			女子雙打	Katrín Atladóttir	冠軍

	A	B	C	D	E	F
1	年份	賽事	公開賽級別	項目	搭檔	成績
2	2009年	賽普勒斯羽球國際賽	國際系列賽	混合雙打	Helgi Johannesson	準決賽
3	2009年	冰島羽球國際賽	國際系列賽	女子單打	–	冠軍
4	2009年	冰島羽球國際賽	國際系列賽	女子雙打	Snjólaug Jóhannsdóttir	冠軍
5	2010年	賽普勒斯羽球國際賽	國際系列賽	女子單打	–	準決賽
6	2010年	冰島羽球國際賽	未來系列賽	女子單打	–	冠軍

工業技術研究院
Industrial Technology Research Institute

Data Cleaning

- Unqualified tables can be removed based on the following criteria:
  - More than half of the table's cells are empty values.
  - The table contains only one row or one column.
- Example: Since images cannot be stored, they are treated as empty values. Therefore, the following table is discarded because more than half of its cells are empty.

節目總結 [編輯]

圖例:

- 主持人（Host）
- 主場董事長（Board member）
- 客隊董事長（Guest board member）
- 選手（Contestant）

Cast	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
唐納·川普															
阿諾·史瓦辛格															
喬治·H·羅斯															
凱洛琳·凱普謝															
比爾·蘭西克															
伊凡卡·川普															
小唐納·川普															
艾瑞克·川普															
瓊·瑞佛斯															
泰拉·班克斯															
派屈克·M·克納普·史瓦辛格															

互通樞紐及服務設施列表 [編輯]

地區	里程	類型	名稱	連接	備註
赤峰市 克什克騰旗			經棚	S105 105省道	
			西拉木倫		
			樺木溝		
			樺木溝		
			烏蘭布統	S304 304省道	
			五彩山		
			蔡木山		
蒙冀界			蒙冀界		
承德市 圍場滿族蒙古族自治縣			塞罕壩西	X508 508縣道	
			御道口草原		
			御道口	G510 510國道	
			木蘭林海		
			半截塔	G111 111國道	
			道壩子		
			龍頭山	G111 111國道	

工業技術研究院
Industrial Technology Research Institute

Table Visualization

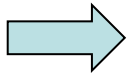
- Ordinary tables can be stored in DataFrame format, but merged-cell tables (cross-column/row) require separate handling.
- For ordinary tables:
  - Read the CSV file into DataFrame format.
  - Use the Playwright library to render images. By simulating the Chromium browser to process HTML and CSS, screenshots of the web content are captured.
  - This method allows customization of backgrounds, borders, fonts, etc., enabling the generation of diverse table images.
- For merged-cell tables:
  - Read the HTML from the webpage TXT file.
  - Use Selenium WebDriver to render images.
  - The resulting merged-cell table images are almost identical to those displayed on Wikipedia pages.

工業技術研究院
Industrial Technology Research Institute

Table Visualization

Merged-cell tables:

年份	賽事	公開賽級別	項目	搭檔	成績
2009年	賽普勒斯羽球國際賽	國際系列賽	混合雙打	Helgi Johannesson	準決賽
	冰島羽球國際賽	國際系列賽	女子單打	–	冠軍
			女子雙打	Snjólaug Jóhannsdóttir	冠軍
2010年	賽普勒斯羽球國際賽	國際系列賽	女子單打	–	準決賽
	冰島羽球國際賽	未來系列賽	女子單打	–	冠軍
			女子雙打	Katrín Atladóttir	冠軍
2011年	立陶宛羽球公開賽	國際系列賽	女子單打	–	亞軍
	冰島羽球國際賽	國際系列賽	女子單打	–	冠軍
	威爾斯羽球國際賽	國際系列賽	女子單打	–	亞軍



年份	賽事	公開賽級別	項目	搭檔	成績
2009年	賽普勒斯羽球國際賽	國際系列賽	混合雙打	Helgi Johannesson	準決賽
	冰島羽球國際賽	國際系列賽	女子單打	–	冠軍
			女子雙打	Snjólaug Jóhannsdóttir	冠軍
2010年	賽普勒斯羽球國際賽	國際系列賽	女子單打	–	準決賽
	冰島羽球國際賽	未來系列賽	女子單打	–	冠軍
			女子雙打	Katrín Atladóttir	冠軍
2011年	立陶宛羽球公開賽	國際系列賽	女子單打	–	亞軍
	冰島羽球國際賽	國際系列賽	女子單打	–	冠軍
	威爾斯羽球國際賽	國際系列賽	女子單打	–	亞軍

Ordinary tables:

年份	主辦國	比賽日期	網站連結
2007	澳大利亞	5月12日	
2008	泰國	5月10日	網站
2009	印度	5月9日	網站（頁面存檔備份，存於網際網路檔案館）
2010	中國	5月8日	網站（頁面存檔備份，存於網際網路檔案館）
2011	伊朗	5月7日	網站（頁面存檔備份，存於網際網路檔案館）
2012	日本	5月12日	網站（頁面存檔備份，存於網際網路檔案館）
2013	新加坡	5月11日	網站
2014	哈薩克斯坦	5月3日-4日	網站
2015	印度尼西亞	5月9日-10日	網站
2016	韓國	5月7日-8日	網站
2017	澳大利亞	5月13日-14日	網站
2018	俄羅斯	5月12日-13日	網站（頁面存檔備份，存於網際網路檔案館）
2019	俄羅斯	5月18日-19日	網站（頁面存檔備份，存於網際網路檔案館）
2020	印度尼西亞	8月14日-21日	網站（頁面存檔備份，存於網際網路檔案館）
2021	印度尼西亞	5月19日-23日	網站（頁面存檔備份，存於網際網路檔案館）
2022	埃及	5月26日-30日	網站（頁面存檔備份，存於網際網路檔案館）

年份	主辦國	比賽日期	網站連結
2007	澳大利亞	5月12日	
2008	泰國	5月10日	網站
2009	印度	5月9日	網站 （ 頁面存檔備份 ， 存於 網際網路檔案館 ）
2010	中國	5月8日	網站 （ 頁面存檔備份 ， 存於 網際網路檔案館 ）
2011	伊朗	5月7日	網站 （ 頁面存檔備份 ， 存於 網際網路檔案館 ）
2012	日本	5月12日	網站 （ 頁面存檔備份 ， 存於 網際網路檔案館 ）
2013	新加坡	5月11日	網站
2014	哈薩克斯坦	5月3日-4日	網站
2015	印度尼西亞	5月9日-10日	網站
2016	韓國	5月7日-8日	網站
2017	澳大利亞	5月13日-14日	網站
2018	俄羅斯	5月12日-13日	網站 （ 頁面存檔備份 ， 存於 網際網路檔案館 ）
2019	俄羅斯	5月18日-19日	網站 （ 頁面存檔備份 ， 存於 網際網路檔案館 ）
2020	印度尼西亞	8月14日-21日	網站 （ 頁面存檔備份 ， 存於 網際網路檔案館 ）
2021	印度尼西亞	5月19日-23日	網站 （ 頁面存檔備份 ， 存於 網際網路檔案館 ）
2022	埃及	5月26日-30日	網站 （ 頁面存檔備份 ， 存於 網際網路檔案館 ）

究院

Question–Answer Generation

- Referring to the paper ChartLlama: A Multimodal LLM for Chart Understanding and Generation, we adopt a large language model (LLM) to generate text-based questions and answers related to the table images.
- Model Used: gpt-4o-mini
- Input to the model: Three attributes from the JSON file — Entity, Description, and Data. The model is prompted to generate 3 sets of questions and answers based on this input.
- The generated questions and answers are then saved as a JSON file.

處理第 1 個數據項：元件物件模型

生成的問題和答案：

Q1: 什麼是「Legacy STA (ThreadingModel=Single或空)」中的COM物件的特點？

A1: 「Legacy STA」中的COM物件屬於處理程序的第一個STA執行緒，通常是UI介面的執行緒，在過去單核CPU時代沒有遺留下來。

Q2: 在「多執行緒套間 (MTA)，(ThreadingModel=Free)」中，為什麼COM物件需要自己實現同步控制？

A2: 在「多執行緒套間 (MTA)」中，COM執行時不提供同步，多個MTA執行緒可以同時呼叫同一個MTA的COM物件，因此COM物件需要自己實現同步控制以避免競態條件或死結。

Q3: 「自動選擇套間 (Both)」中的COM物件的套間類別與建立它的執行緒的套間類別有什麼特點？

A3: 在「自動選擇套間 (Both)」中，COM物件的套間類別與建立它的執行緒的套間類別一致，避免了許多marshalling開銷，例如一個MTA伺服器被一個STA執行緒呼叫。

工業技術研究院
Industrial Technology Research Institute

Question–Answer Generation Process

- 1.Input the pre-processed JSON dataset into a large language model (via ChatGPT API) with a designed prompt and few-shot examples. The model generates questions for each table (referencing TableBench categories such as Comparison, Arithmetic Calculation, Ranking, etc.), and the question type is determined by whether the table contains time- or number-related terms.
- 2. Feed the generated questions back into the model for evaluation (criteria: clarity, answerability, independence, non-leading, etc.). Low-scoring questions are automatically refined into better versions.
- 3. Use the refined questions to generate answers with the model.
- 4. Evaluate the correctness of the answers with the model, refine them if needed, and extract the final outputs.

工業技術研究院
Industrial Technology Research Institute

Traditional Chinese Table and Q&A Dataset

- Sample Data

question	answer
拉娜·比約·因戈爾夫斯多蒂爾在2009年冰島羽球國際賽中的成績如何？(How did Ragna Björg Ingólfsdóttir perform at the 2009 Iceland International badminton tournament?)	拉娜·比約·因戈爾夫斯多蒂爾在2009年冰島羽球國際賽中，女子單打項目中獲得了冠軍，女子雙打項目中與搭檔Snjólaug Jóhannsdóttir一起獲得了冠軍。(At the 2009 Iceland International, Ragna Björg Ingólfsdóttir won the women's singles title and, together with her partner Snjólaug Jóhannsdóttir, also won the women's doubles title.)
拉娜·比約·因戈爾夫斯多蒂爾在2010年的賽普勒斯羽球國際賽中取得了怎樣的成績？(How did Ragna Björg Ingólfsdóttir perform at the 2010 Cyprus International badminton tournament?)	在2010年的賽普勒斯羽球國際賽中，拉娜·比約·因戈爾夫斯多蒂爾在女子單打項目中取得了準決賽的成績。(At the 2010 Cyprus International, Ragna Björg Ingólfsdóttir reached the semifinals in the women's singles event.)
拉娜·比約·因戈爾夫斯多蒂爾在2011年的比賽中獲得了哪些獎項？(What titles did Ragna Björg Ingólfsdóttir win in tournaments in 2011?)	拉娜·比約·因戈爾夫斯多蒂爾在2011年的比賽中，分別在冰島羽球國際賽和威爾斯羽球國際賽的女子單打項目中分別獲得了冠軍和亞軍的成績。(In 2011, Ragna Björg Ingólfsdóttir won the women's singles title at the Iceland International and finished as the runner-up in the women's singles event at the Welsh International.)

年份	賽事	公開賽級別	項目	搭檔	成績
2009年	賽普勒斯羽球國際賽	國際系列賽	混合雙打	Helgi Johannesson	準決賽
	冰島羽球國際賽	國際系列賽	女子單打	–	冠軍
			女子雙打	Snjólaug Jóhannsdóttir	冠軍
2010年	賽普勒斯羽球國際賽	國際系列賽	女子單打	–	準決賽
	冰島羽球國際賽	未來系列賽	女子單打	–	冠軍
			女子雙打	Katrín Atladóttir	冠軍
2011年	立陶宛羽球公開賽	國際系列賽	女子單打	–	亞軍
	冰島羽球國際賽	國際系列賽	女子單打	–	冠軍
	威爾斯羽球國際賽	國際系列賽	女子單打	–	亞軍

工業技術研究院
Industrial Technology Research Institute


thank you
danke
謝謝
ngiyabonga
teşekkür ederim
gracias
спасибо
dank je
tapadh leat
bedankt
dziękuję
obrigado
merci
go raibh maith agat
mochchakkeram
sukriya
kop khun krap
terima kasih
grazie
arigatō
dakujem
takk
hvala
mauruuru
sagolun
dhanyavadagalu
мерси
xiexie
감사합니다
ευχαριστώ
рахмат
Баярлалаа
faafetai lava
kiitos
dankie
dhanyavad
enkosi
mesi
dekuji
sobodi
didi madloba
najis tuke
rahmat
kam sah hamnida
তোমাকে ধন্যবাদ
chnorakaloutioun
gratias ago
gracies
suipay
ありがとう
tanemirt
rahmet
diolch
shukriya
merce
djiere dieuf
tau
дякую
mamnun
trugarez
asante
manana
obrigada
murakoze
ebokwano
tenki
хвала
paldies
grazzi
misaotra
matondo
mahalo
wetalin
tack
barka
mersi
blagodaram
спасибі
vinaka
akun dankon
aciū
koszönöm
bayarlalaa
gracie
nanmi
nandri
dhanyavad