In [1]:
```
# Name of creator
CREATOR_NAME = "Jingheng Wang"
```

This notebook is the very first file of the project. It is the first step towards answering Question 1: "What are the most frequently used Kanjis?"

On 2010, the Agency of Cultural Affairs, Government of Japan published a list of "Joyo Kanjis" (frequently used kanjis) as a pdf file. Wikipedia contributors transfered the pdf file into a chart, and put on a Wikipedia page (https://en.wikipedia.org/wiki/List_of_j%C5%8Dy%C5%8D_kanji (https://en.wikipedia.org/wiki/List_of_j%C5%8Dy%C5%8D_kanji)).

The task of this file is to scrape all of the Kanjis on that page, in total 2136 of them, and save in a csv file for further scraping.

To use:

```
Run the whole notebook. A file "kanji_list.csv" will be created in the same directory
with this notebook.
```

In [1]:
```
# Initialization
# HTML GET
from requests import get

# Web Scraping
from bs4 import BeautifulSoup

# Data Wrangling
import numpy as np
import pandas as pd
```

First, request the content of the target page.

In [2]:
```
# Target URL
url = 'https://en.wikipedia.org/wiki/List_of_j%C5%8Dy%C5%8D_kanji'

# HTML get, store the webpage to response
response = get(url)

# Test if scraped down the website
print(response.text[:200])
```

```
<!DOCTYPE html>
<html class="client-nojs" lang="en" dir="ltr">
<head>
<meta charset="UTF-8"/>
<title>List of jōyō kanji - Wikipedia</title>
<script>document.documentElement.className="client-js";RLCON
```

Use BeautifulSoup to analyze

In [3]:
```
# Use BeautifulSoup to analyze the code
soup = BeautifulSoup(response.text, 'html.parser')

# Locate the big table on that website
kanji_table = soup.find('table', class_='sortable wikitable')

print(kanji_table)
```

```
<table class="sortable wikitable" style="font-family:'ヒラギノ角ゴ ProN W3','ヒラギ
ノ角ゴ ProN','Hiragino Kaku Gothic ProN','メイリオ',Meiryo,'新ゴ Pr6N R','A-OTF 新ゴ
Pr6N R','小塚ゴシック Pr6N M','IPAexゴシック','Takaoゴシック','XANO明朝U32','XANO明朝
','和田研中丸ゴシック2004絵文字','和田研中丸ゴシック2004ARIB','和田研中丸ゴシック2004P4','
和田研細丸ゴシック2004絵文字','和田研細丸ゴシック2004ARIB','和田研細丸ゴシック2004P4','和田
研細丸ゴシックProN','IPA Pゴシック','MS Pゴシック';">
<tbody><tr style="line-height:1.4em">
<th>#
</th>
<th><a href="/wiki/Shinjitai" title="Shinjitai">New</a>
</th>
<th><a href="/wiki/Ky%C5%ABjitai" title="Kyūjitai">Old</a>
</th>
<th>Radical
</th>
<th>Strokes
</th>
<th>Grade
</th>
<th>Year added
</th>
<th>English meaning
</th>
<th>Readings
</th></tr>
<tr>
<td>1</td>
<td style="font-size:2em"><a class="extiw" href="https://en.wiktionary.org/wiki
/%E4%BA%9C" title="wikt:亜">亜</a></td>
<td style="font-size:2em"><a class="extiw" href="https://en.wiktionary.org/wiki
/%E4%BA%9E" title="wikt:亞">亞</a></td>
<td style="font-size:2em"><a href="/wiki/Radical_7" title="Radical 7">二</a></td>
<td>7</td>
<td>S</td>
<td></td>
<td>sub-</td>
<td>ア<br/>a
</td></tr>
<tr>
<td>2</td>
<td style="font-size:2em"><a class="extiw" href="https://en.wiktionary.org/wiki
/%E5%93%80" title="wikt:哀">哀</a></td>
<td></td>
<td style="font-size:2em"><a href="/wiki/Radical_30" title="Radical 30">口</a></t
d>
<td>9</td>
<td>S</td>
<td></td>
<td>pathetic</td>
<td>アイ、あわ-れ、あわ-れむ<br/>ai, awa-re, awa-remu
</td></tr>
<tr>
<td>3</td>
<td style="font-size:2em"><a class="extiw" href="https://en.wiktionary.org/wiki
/%E6%8C%A8" title="wikt:挨">挨</a></td>
<td></td>
<td style="font-size:2em"><a href="/wiki/Radical_64" title="Radical 64">手</a></t
d>
<td>10</td>
<td>S</td>
<td>2010</td>
<td>push open</td>
<td>アイ<br/>ai
</td></tr>
```

In [4]:
```
# tr_list is an iterator of all <tr>s
tr_list = kanji_table.tbody.children


# container of data
container = pd.Series()

# for each single tr
for single_tr in tr_list:

#    print(single_tr)
    try:
        # scrape the sequence number
        seq_num = single_tr.td.get_text()

        # scapre the content (the kanji itself)
        contents = single_tr.td.next_sibling.next_sibling.a.get_text()

        # store it into our series
        container[seq_num] = contents

    except:
        continue


container
```

Out[4]:
```
1       亜
2       哀
3       挨
4       愛
5       曖
       ..
2132    脇
2133    惑
2134    枠
2135    湾
2136    腕
Length: 2136, dtype: object
```

In [5]:
```
# refurbrish the series to dataframe with index and contents
df = pd.DataFrame(container, index = container.index, columns = ['kanji'])
df.index.name = 'index'

# write to csv file
df.to_csv('kanji_list.csv')
```

In [6]:
```
# FOR DISPLAY ONLY: show the kanjis written dataframe
df.dropna()
df
```

Out[6]:

index	kanji
1	亜
2	哀
3	挨
4	愛
5	曖
...	...
2132	脇
2133	惑
2134	枠
2135	湾
2136	腕

2136 rows × 1 columns

In [7]:
```
# FOR DISPLAY ONLY: read the dataframe just written
df2 = pd.read_csv("kanji_list.csv")
df2.head()
```

Out[7]:

	index	kanji
0	1	亜
1	2	哀
2	3	挨
3	4	愛
4	5	曖

In [ ]:
```

```