In [1]:
```
NAME = "Dustin Seltz"
```

Purpose:

Scrape frequency data of different sources from https://scriptin.github.io/kanji-frequency/ (https://scriptin.github.io/kanji-frequency/)

Input:

None.

Output:

KanjiFrequencyOnWikipedia.csv

KanjiFrequencyOnNews.csv

KanjiFrequencyOnTwitter.csv

KanjiFrequencyOnAozora.csv

In [2]:
```
from bs4 import BeautifulSoup
from urllib.request import urlopen

import re

import numpy as np
import pandas as pd
```

In [3]:
```
#There's probably an easier way to load data from JSON,
#    but I could use more practice with web scraping anyway.
#(I scraped this data before we covered anything about JSON)
sourceNames = ["Aozora", "News", "Twitter", "Wikipedia"]
urlsToScrape = []
for name in sourceNames:
    name = name.lower()
    urlsToScrape.append("https://raw.githubusercontent.com/scriptin/kanji-frequency
/master/data/"+name+".json")
```

In [ ]:
```
htmls = [urlopen(urlToScrape) for urlToScrape in urlsToScrape]
soups = [BeautifulSoup(html, 'lxml') for html in htmls]
```

In [ ]:
```
wikiSoup = soups[3]
wikiSoup
```

In [ ]:
```
#Testing regex on this stuff
str1 = """["年",21066593,0.02685131991368414],"""
expr = """\["(.)",(.*),(.*)\]"""
match = re.match(expr, str1)
if(match):
    print(match.group())
    print(match.group(1))
    print(match.group(2))
    print(match.group(3))
```

In [ ]:
```
#Scrape from the soup
expr = """\["(.)",(.*),(.*)\]"""
paragraph = wikiSoup.find_all("p")[0].text
paragraphLines = paragraph.splitlines()
```

In [ ]:
```
matches = [re.findall(expr, line) for line in paragraphLines]
matches
```

In [ ]:
```
#Turn the result into something we can easily make into a dataframe
matchesList = []
for entry in matches:
    #Each entry is a list of length 1.
    for tup in entry:
        if(len(tup) == 3):
            character = tup[0]
            numberOfAppearances = tup[1]
            percentage = tup[2]
            matchesList.append([character, numberOfAppearances, percentage])
matchesList
```

In [ ]:
```
#Store the data in a dataframe
colNames = ["Character", "Number of Appearances", "%"]
df = pd.DataFrame(matchesList, columns=colNames)
df
```

In [ ]:
```
#Store the dataframe in a file
file_name = "KanjiFrequencyOnWikipedia"
df.to_csv(file_name, index=False)
```

In [ ]:
```
#Test that it worked
df = pd.read_csv(file_name)
df
```

In [ ]:
```
#Looks good, but I want to make all four output files at once. Lets do all that stu
ff but in a loop.
for (soup,name) in zip(soups, sourceNames):
    #Scrape from the soup
    expr = """\["(.)",(.*),(.*)\]"""
    paragraph = soup.find_all("p")[0].text
    paragraphLines = paragraph.splitlines()
    matches = [re.findall(expr, line) for line in paragraphLines]
    #Turn the result into something we can easily make into a dataframe
    matchesList = []
    for entry in matches:
        #Each entry is a list of length 1.
        for tup in entry:
            if(len(tup) == 3):
                character = tup[0]
                numberOfAppearances = tup[1]
                percentage = tup[2]
                matchesList.append([character, numberOfAppearances, percentage])
    #Store the data in a dataframe
    colNames = ["Character", "Number of Appearances", "%"]
    df = pd.DataFrame(matchesList, columns=colNames)
    #Store the dataframe in a file
    file_name = "KanjiFrequencyOn"+name
    df.to_csv(file_name, index=False)
```