Scraping Japanese Kanji for an Educational Game

Dustin Seltz • Mengdi Wei • Jimmy Wang • Xudong Guo

Questions:

- What are the most common characters, and information about them such as: frequency of use, meanings, readings, number of strokes, radicals, and difficulty level.
- A single kanji character could have various readings when formed into different words. Given a kanji, what words can it form, and what are the readings of this kanji in each of those words?
- Kanjis have various information: number of strokes, difficulty levels, and frequency of use in news, and WaniKani levels (assigned by WaniKani, a Japanese learning website). Pairing them with each other, do they show some interesting relationships?
- Taking into account the student's progress and goals, what is the best set of kanji / vocab to teach to them next?

Using data from:

- Jisho.org
  A dictionary with all sorts of information for each kanji character.
- http://genki.japantimes.co.jp/self/genki-kanji-list-linked-to-wwkanji
  Additional information regarding difficulty level.
- https://en.wikipedia.org/wiki/List_of_j%C5%8Dy%C5%8D_kanji
  Information from the official Jōyō table.
- https://scriptin.github.io/kanji-frequency/
  A comparison of Kanji frequency from various sources.
- https://www.wanikani.com/
  Additional difficulty information.
- https://en.wiktionary.org/wiki/Wiktionary:Frequency_lists/Japanese
  https://en.wiktionary.org/wiki/Wiktionary:Frequency_lists/Japanese10001-20000
  Frequency information for entire words.
- The app
  The user selects their goal to determine the order of kanji to learn.

Code references:

- Another UCSC student, Bryan, started some code for Kanjirer: https://colab.research.google.com/drive/1ulFNimI9YYQF2K5CXvri8PoeD20TicbK

Viz references:

- https://scriptin.github.io/kanji-frequency/
  A visual comparison of Kanji frequency from various sources.

Answers provided or final results obtained:

- Question 1

  We obtained a large csv file, containing 2136 frequently used kanjis with its meanings, readings, number of strokes, radicals, wanikani levels, compounds, rank of frequency in news, etc.

- Question 2

  We obtained two csv files: one includes the most frequently used 20K words, the other obtains the 20K words' meanings and readings. Also finished a query function.

- Question 3

  From Jisho visualization:

  - Kanjis with less number of strokes seems to appear more frequent than those of more strokes.
  - As the students go into higher grades, they study kanjis that are much less frequently used.
  - We could hardly find some relationships from frequency of these kanji with their radicals.
  - Japanese junior high students are learning more kanji than elementary students per year
  - An N2 learner might have the same level as Japan's elementary school graduate, while N1 learner would have the same level as a junior high graduate.

  From WaniKani visualization:

  - Wanikani distributed kanjis in a quite uniform way. Each level contains approximately 45 kanjis.

- Wanikani tends to introduce the most frequently used kanjis first. As the kanji's level goes higher, their frequency of use is becoming less.
- After level 10, the number of strokes may not be a categorizing standard.
- Lower WaniKani level kanji are taught in grade 1 or 2, while the higher level ones mostly taught in junior high. They have somewhat a correlation.

- Question 4

  Did a comparison of frequency and learning sequences from various sources to attempt to determine the optimal learning sequence to follow to learn how to read a given source. Found WaniKani to be the best choice for all sources.

  Has easy to use queries for the level information. For example, given that a user has completed "N5" what are the next 6 kanji that the user should learn?