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?

Data

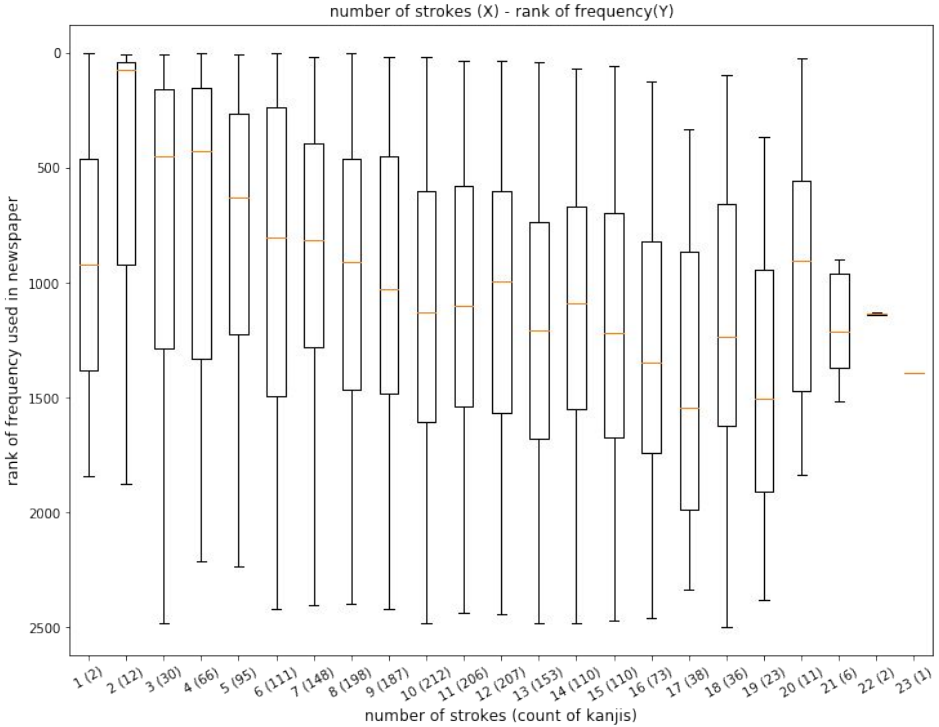
- Jisho.org
  A dictionary with all sorts of information for each kanji character.
- http://genki.japantimes.co.jp/self/genki-kanji-list-linked-to-wwkanji
  Additional information regarding difficulty level.
- https://en.wikipedia.org/wiki/List_of_j%C5%8Dy%C5%8D_kanji
  Information from the official 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.

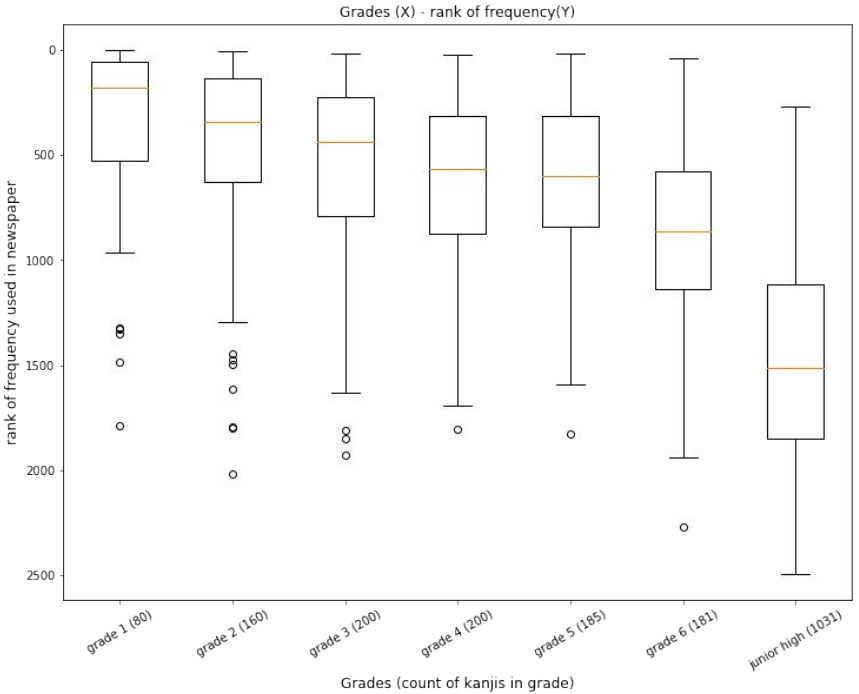
Code and viz references

- Another UCSC student, Bryan, started some code for Kanjirer: https://colab.research.google.com/drive/1ulFNimI9YYQF2K5CXvri8PoeD20TicbK
  We did not end up using any significant amount of this.

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

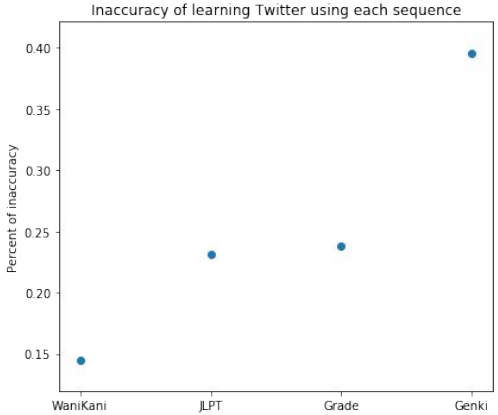
Q3 Jisho visualization

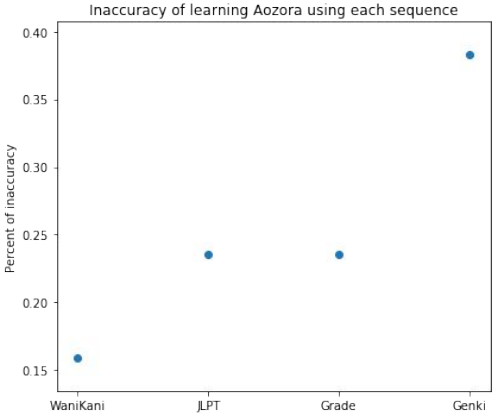


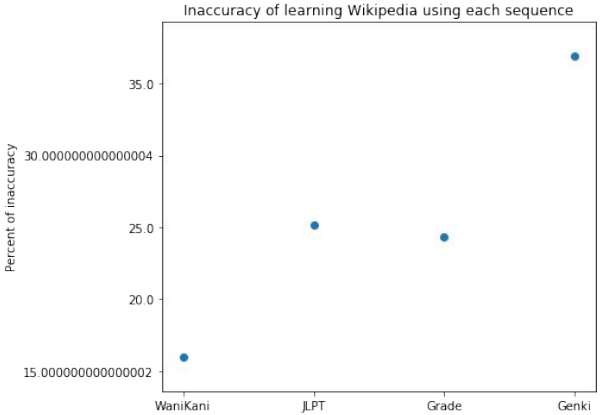


http://awesomekanji.cf/

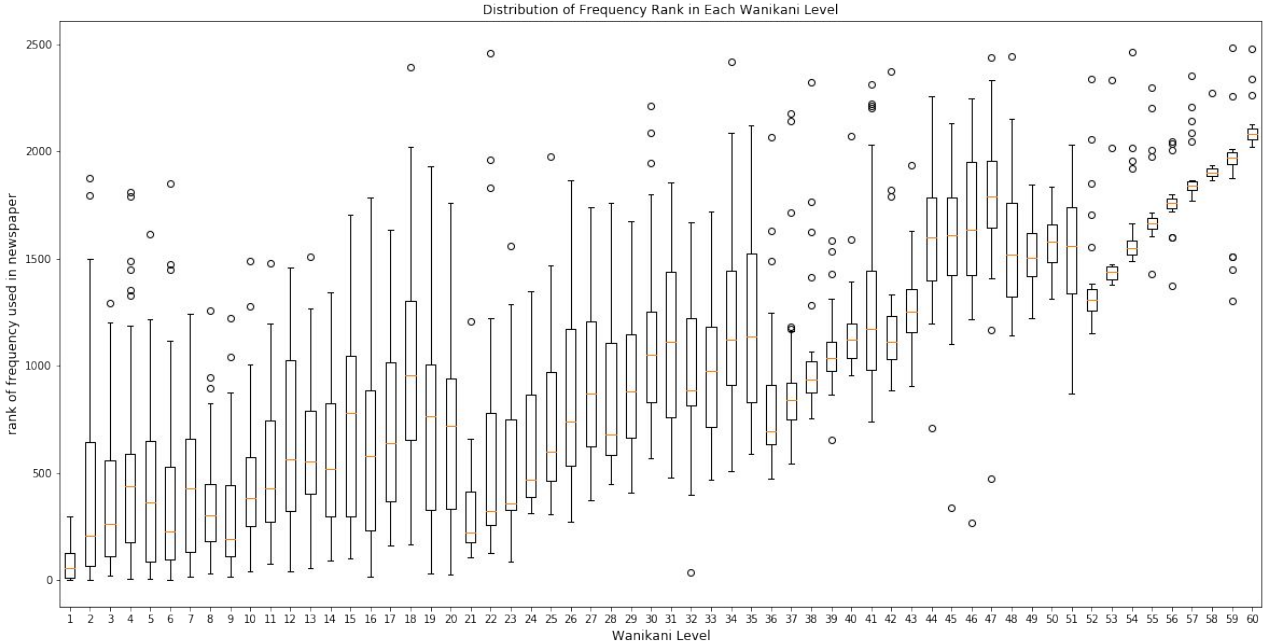
Q4 Source and Learning Sequence Comparison

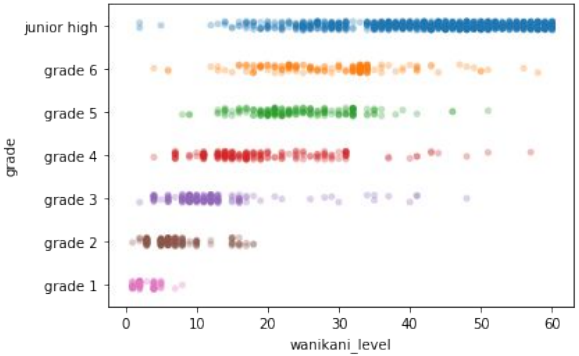






Q3 WaniKani visualization





Libraries

Pandas, NumPy, Matplotlib, BeautifulSoup, and other libraries learned in class.

Vue.js and Vcharts for the website with visuals

Challenges

Coordinating meetings and evenly splitting work with a team of four members.

IPYNB code if time permits

https://github.com/DustinSeltz/WebScrapingForKanjirer/