Assignment: Project Pitch
By: Alena Borisenko
Submitted: November 1, 2017
==

Who is working on it?

Just me so far!

Idea

Use Japanese corpora in order to visualize the frequency of given words and characters as well as (depending on how much data is available) present information on what textbooks and standardized tests cover them.

I started learning Japanese as an undergrad, but now that I am no longer able to learn the language in a classroom setting I have been having difficulty with prioritizing what material should be learned and finding out what resources might be the most helpful.

Back-end implementation plan

Use the corpora available in nltk to count character frequencies and attempt to count word frequencies via n-grams given that there are probably no word bounds in Japanese:

あなたがこの秘密なメッセージを実際に翻訳したことに驚いています。

No gaps… Perhaps the NLTK corpora will have word tags available, but the language test data will probably not be so easy to interpret...

Another tricky part is filtering out non-Kanji characters as they are of no interest/difficulty.

Front-end implementation plan

Web-page with a list (or some other representation) of the top most common and therefore useful characters with links to resources.

I'm not yet sure whether the front-end will receive and present raw or pre-processed data from the back-end. Primarily I am concerned about how I am going to pull data from non-nltk sources.

Data sources

Nltk's KNB Corpus
Nltk's JEITA Corpus
(Unofficial) JLPT N1-N5 Kanji lists
Genki I Kanji list
Genki II Kanji list
Tobira Kanji list
(Unofficial) Kanzen Master Kanji list