Japanese NLTK Support

Flat Iron Capstone Experience

Christopher Hyek

9/25/2019

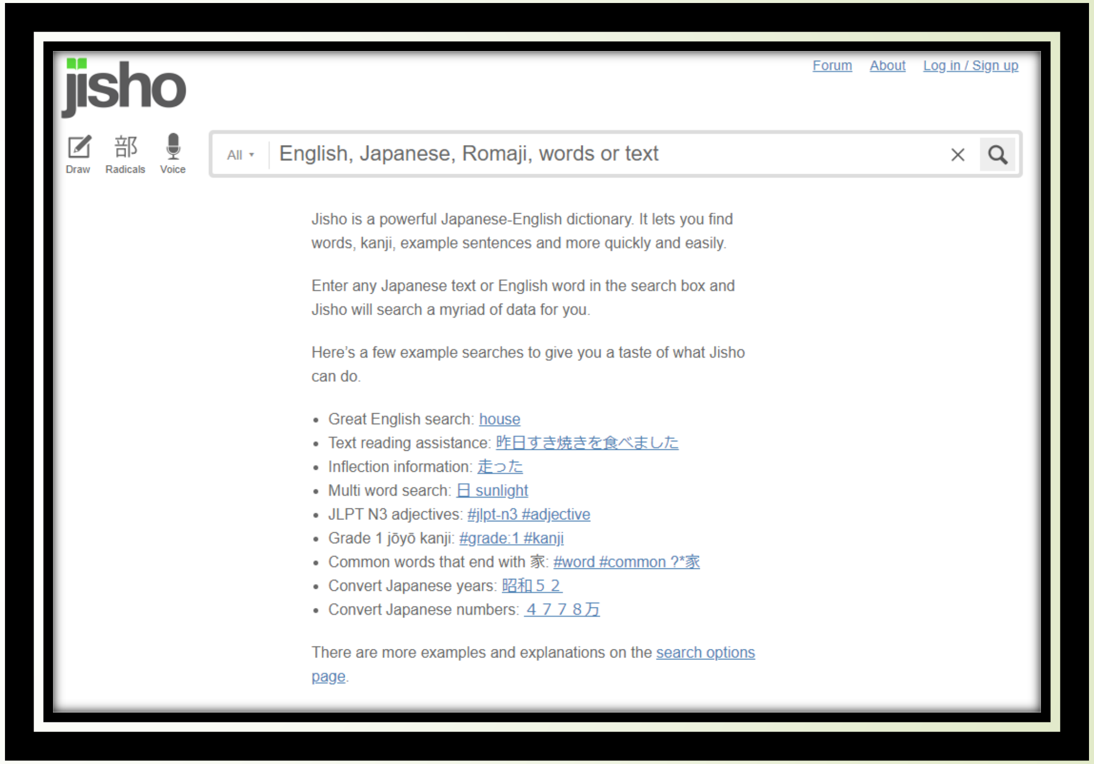
Focus

- Create data from Japanese text that can be used in Machine Learning
- Make several functions from NLTK compatible

Natural Language
Tool Kit (NLTK)
Basic Text Analytics

Methodology – Obtaining Data

- Take data from a Japanese Dictionary (https://jisho.org/)
- Create program to take necessary information from it



Jisho is a powerful Japanese-English dictionary. It lets you find words, kanji, example sentences and more quickly and easily.

Enter any Japanese text or English word in the search box and Jisho will search a myriad of data for you.

Here's a few example searches to give you a taste of what Jisho can do.

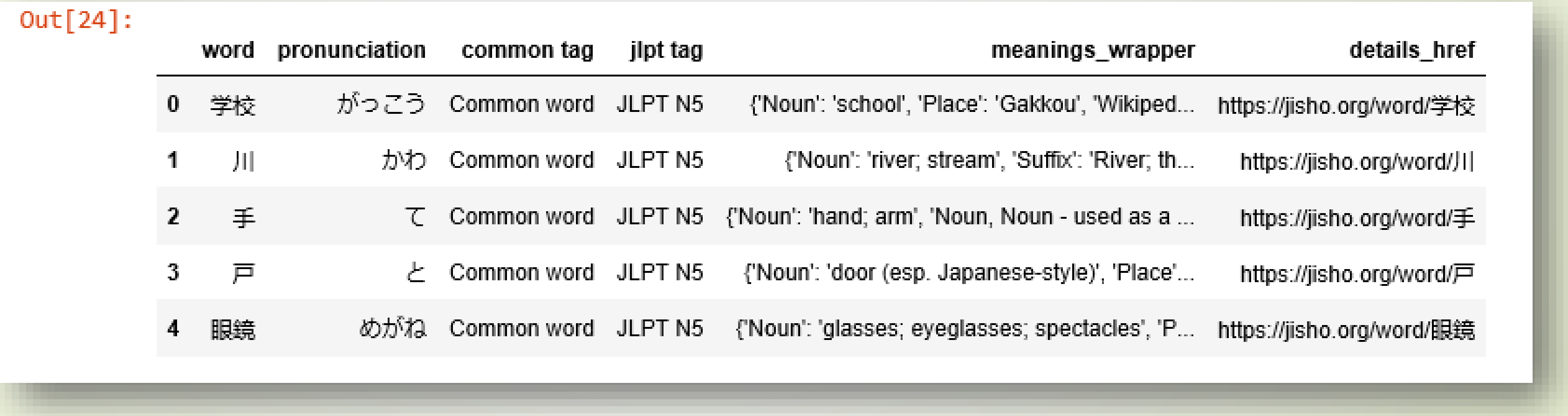
- Great English search: house
- Text reading assistance: 昨日すき焼きを食べました
- Inflection information: 走った
- Multi word search: 日 sunlight
- JLPT N3 adjectives: #jlpt-n3 #adjective
- Grade 1 jōyō kanji: #grade:1 #kanji
- Common words that end with 家: #word #common ?*家
- Convert Japanese years: 昭和５２
- Convert Japanese numbers: ４７７８万

There are more examples and explanations on the search options page.

Methodology – Cleaning Data

- Null, Blank, Incorrect Values

- Removing empty columns



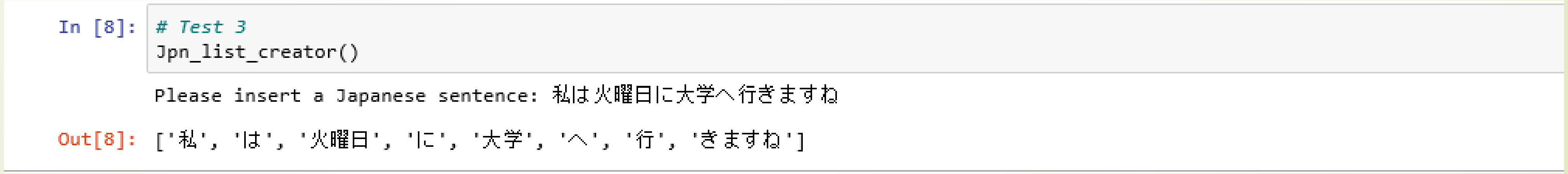
Out[24]:

	word	pronunciation	common tag	jlpt tag	meanings_wrapper	details_href
0	学校	がっこう	Common word	JLPT N5	{'Noun': 'school', 'Place': 'Gakkou', 'Wikiped...	https://jisho.org/word/学校
1	川	かわ	Common word	JLPT N5	{'Noun': 'river; stream', 'Suffix': 'River; th...	https://jisho.org/word/川
2	手	て	Common word	JLPT N5	{'Noun': 'hand; arm', 'Noun, Noun - used as a ...	https://jisho.org/word/手
3	戸	と	Common word	JLPT N5	{'Noun': 'door (esp. Japanese-style)', 'Place'...	https://jisho.org/word/戸
4	眼鏡	めがね	Common word	JLPT N5	{'Noun': 'glasses; eyeglasses; spectacles', 'P...	https://jisho.org/word/眼鏡

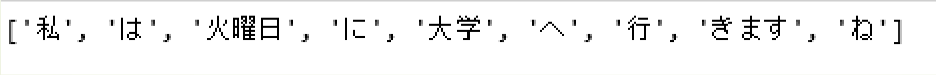
Methodology – Tokenization and Stop Words

- Split characters by script
- Recognize Japanese Stop Words (aka Particles)

(sentence meaning: "On Tuesday I go to college“)



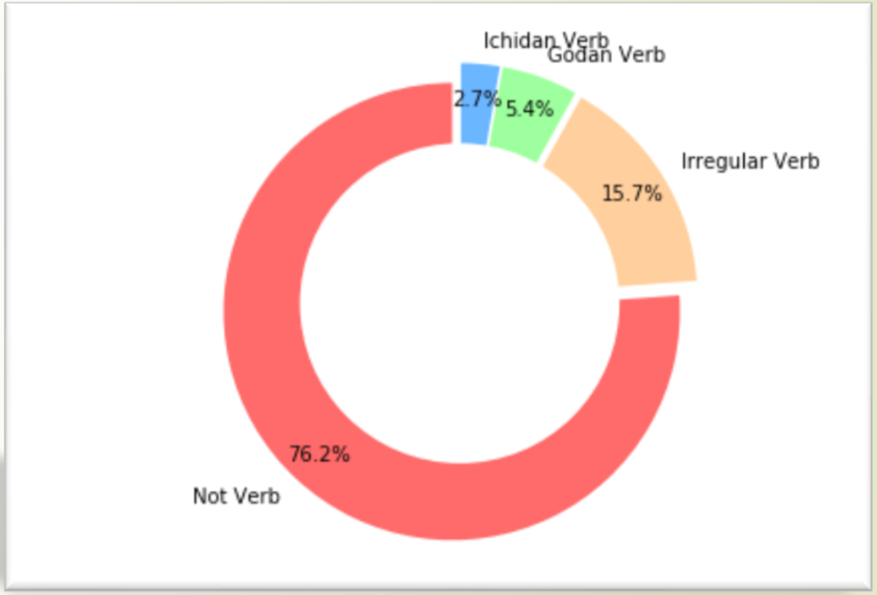
```
In [8]: # Test 3
        Jpn_list_creator()

        Please insert a Japanese sentence: 私は火曜日に大学へ行きますね

Out[8]: ['私', 'は', '火曜日', 'に', '大学', 'へ', '行', 'きますね']
```



```
['私', 'は', '火曜日', 'に', '大学', 'へ', '行', 'きます', 'ね']
```

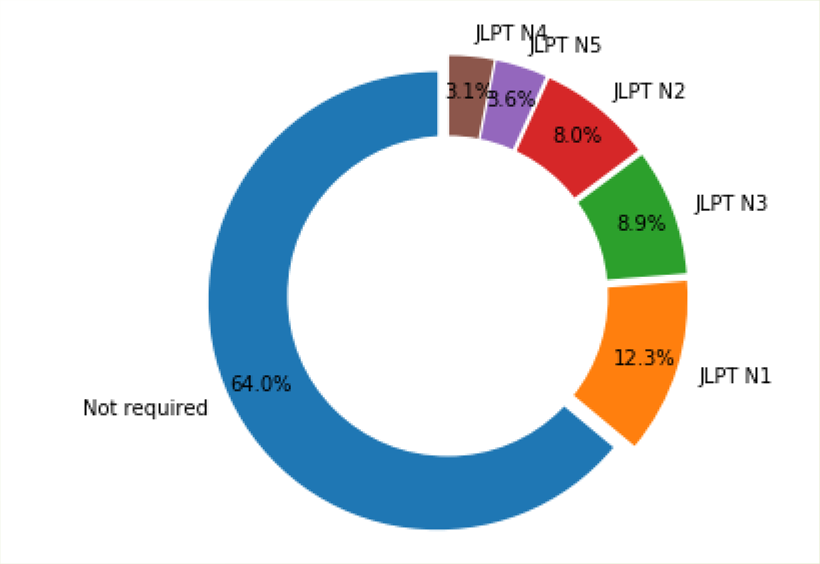
Methodology – Stemming and Lemmatization

- Lemmatization is normal Dictionary form
- Stemming was tricky
  - 20+ conjugations x 14+ verb types
  - Majority of issues come from 5.4% of words!



Methodology – Transferring Data

- Take individual words and cross validate meaning
- Easily readable by NLTK library
  - Barring more complex examples
- Words covered make up majority of words ‘needed’



Capabilities

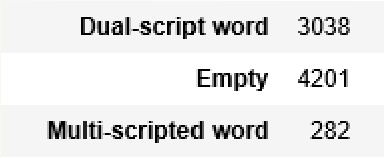
- Split a Japanese sentence into words
- Separate particles
- Search what remains
- Cross reference with English

Hello!

今日は！

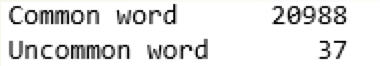
Reservations and Concerns

- Many of the complex portions make up very few words



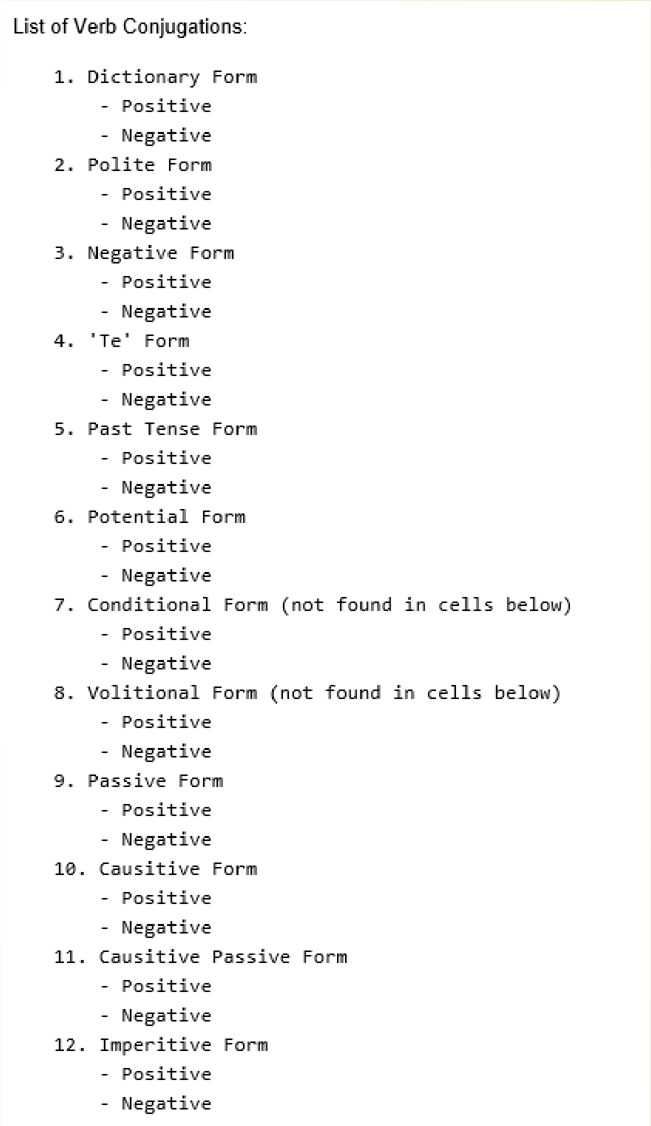
Dual-script word	3038
Empty	4201
Multi-scripted word	282

- Problematic data source



```
Common word        20988
Uncommon word         37
```

- Overbearing conjugation lists



List of Verb Conjugations:

```
   1. Dictionary Form
       - Positive
       - Negative
   2. Polite Form
       - Positive
       - Negative
   3. Negative Form
       - Positive
       - Negative
   4. 'Te' Form
       - Positive
       - Negative
   5. Past Tense Form
       - Positive
       - Negative
   6. Potential Form
       - Positive
       - Negative
   7. Conditional Form (not found in cells below)
       - Positive
       - Negative
   8. Volitional Form (not found in cells below)
       - Positive
       - Negative
   9. Passive Form
       - Positive
       - Negative
   10. Causitive Form
       - Positive
       - Negative
   11. Causitive Passive Form
       - Positive
       - Negative
   12. Imperitive Form
       - Positive
       - Negative
```

Future Endeavors

- Cross reference data with new dataset
- Obtain more accurate and various data
- Create reasonably sized conjugation converter

Thank you for your time!