French Dictionary

Hussem Ben Belgacem

Paris, France
hussem.benbelgacem@gmail.com

Introduction

The goal of this project is to provide a complete and easy to use french dictionary for developers. To do so, we use data extracted from the free multilingual community based site **Wiktionary** [2]. From those extracted data, we create multiple CSV files with header (form, tags) where each files corresponds to one of the following classes of words: *adjectives* (adj.csv), *adverbs* (adv.csv), *conjunctions* (conj.csv), *determinants* (det.csv), *nouns* (noun.csv), *prepositions* (prep.csv), *pronouns* (pron.csv) and *verbs* (verb.csv). In addition to those files, another file called dictionary.csv is created that contains all the words from the different files without the tags nor the header.

Wiktionary extraction

As previously mentioned, we use data extracted from Wiktionary to generate our dictionary. Those data come from the **wiktextract**[4] project. More specifically we use the JSON generated by the wiktextract project containing the scraped Wiktionary data that can be downloaded on **kaikki.org**[1].

CSV files generation

The JSON file from kaikki.org needs to be processed in order to keep only the necessary data. This processing is done through 4 easy steps. The first phase is to extract only the *pos*, *forms* and *word* columns from the loaded JSON data where *pos* is the columns that contains the name of the class of words, *forms* is the columns that contain all the forms for a given word (*form* and *tag* for each form of word) and *word* the concerned word. We then have to denormalize the column *forms* to have direct access to *form* and *tags*. In the third phase, we make sure that we have the infinitive of each words using *retrieveMissingWordForms* function. Finally, we remove the duplicates using *dropDuplicates* function, we sort by alphanumerical order and we save all the files in UTF-8 format with "`\n`" newline character and "," as the separation character between the columns.
An up-to-date version of the generated files alongside the instructions to use the program to generate them can be found in the **French-Dictionary**[3] Github repository.

Note on the data

Example of data:

form	tags
n'importe quels	"['masculine', 'plural']"
nos	
nosdites	"['feminine', 'plural']"
nosdits	"['masculine', 'plural']"
soit ... soit	['canonical']

Most of the data take the form of only one word but some can be composed of multiple words like "*n'importe quels*". Some have tags and others don't. If you need to use the tags you will have to evaluate the string under the *tags* column for the considered word to have access to its content programmatically. This can be easily done using the *eval()* function if your are using Python. Another case is the use of expressions like "*soit ... soit*". The quality of the output data is highly dependent on the quality of the extracted data present in the JSON generated using wiktextract. However, we can add new processing steps to get a cleaner dictionary. Don't hesitate to open issues on the related Github repository if necessary.

References

1. Json file, https://kaikki.org/dictionary/French/kaikki.org-dictionary-French.json
2. Wiktionary, https://www.wiktionary.org
3. BenBelgacem, H.: Dictionary, https://github.com/hbenbel/French-Dictionary
4. Ylonen, T.: Wiktextract, https://github.com/tatuylonen/wiktextract