Concordancer: A Python Library for Concordance Search with CQL

Yongfu Liao

✉ liao961120@gmail.com

⌨ @liao961120

Graduate Institute of Linguistics, National Taiwan University

Introduction

Concordancer is a Python (≥3.7) library for corpus building and search. It aims at **alleviating the pain of building and searching corpora from self-collected text data**. Most existing large corpora (BNC, COCA, ASBC, etc.) provide web-based search interfaces and support searching with powerful query languages designed for locating complex linguistic patterns in a corpus. Though powerful, these corpora become useless when researchers wants to **use their own data for research**. Limited by the programing skills they have, researchers may not be able to retrieve the necessary data from the corpus for a study. *Concordancer* is here to fill this gap by providing a useful subset of the mighty Corpus Query Language (CQL) [1], enabling users to locate complex linguistic patterns in their own corpora!

Usage

Basically, you just need to...

1. Process raw data into the **required input format**[1]
2. Search the corpus with **Corpus Query Language**
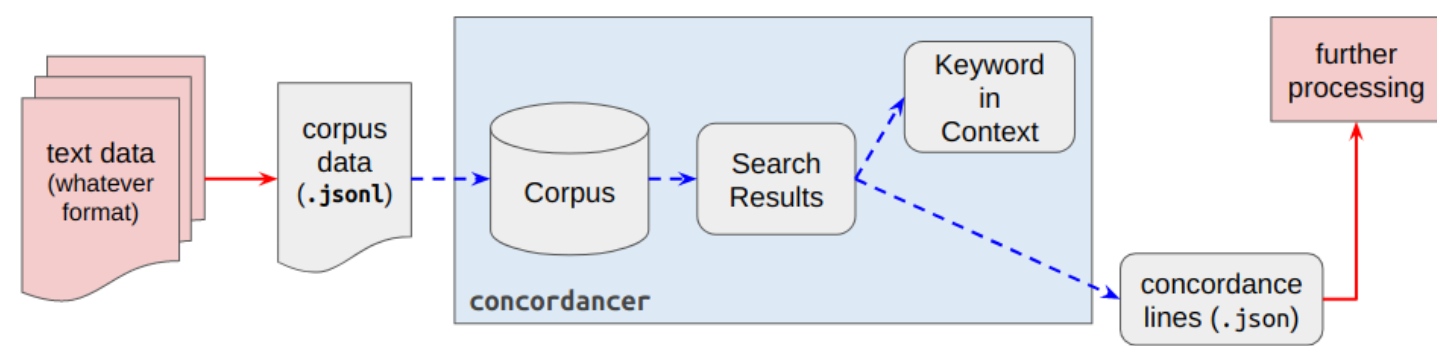3. Do anything you want with the **search results** (convert to csv, analyze, annotate, etc.)

Figure 1: The workflow for using Concordancer

Turn **text data** into a **Searchable Corpus** with **5 lines** of code in **Python**

bit.ly/pyccd

Corpus Query Language

Corpus Query Language (CQL) [1] is a powerful query language designed for searching linguistic corpora and is used by many corpus systems such as the BNCweb and the Sketch Engine.

The full set of CQL is large and often confuses users encountering it for the first time. A great place to start learning CQL is the documentation of CQL provided by BlackLab [2]. A more complete tutorial on CQL can be found in [1].

Program Design

CQL Interpreter

CQL search is supported through cqls[2], a Python library that converts a CQL string into a list of queries represented in JSON (see Fig. 2).
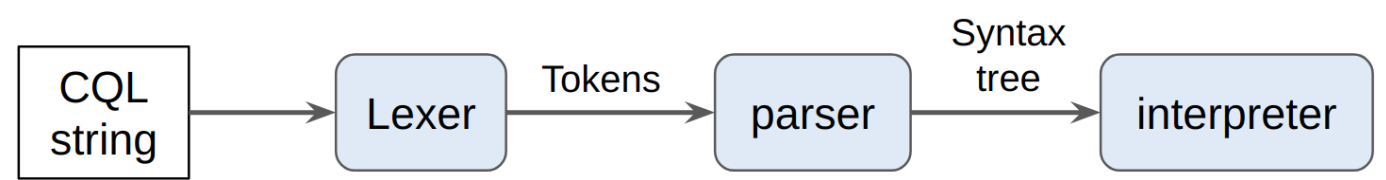

Figure 2: A simple CQL interpreter written in Python

Corpus Indexing

The corpus loaded into *Concordancer* is indexed[3] to allow for **efficient searches** [3]. A simple search on a corpus with 5 million tokens[4] could be done in less than a second!

Front-end Application

Concordancer could be run as a web application, where users can interact with the corpus through a **web browser** (instead of the command line).

References

[1] S. Evert, "The CQP query language tutorial." 2009, [Online]. Available: http://cwb.sourceforge.net/temp/CQPTutorial.pdf.

[2] BlackLab, "BlackLab Corpus Query Language." http://inl.github.io/BlackLab/corpus-query-language.

[3] 洪漢唐 and 江琼玉, "小農手作：語料庫索引與建置," *HOCOR 2020*.

[4] D. Callanan, "Simple Math Interpreter in Python," Feb. 19, 2020.

Footnotes

1. Refer to the library's doc: https://yongfu.name/concordancer↵
2. https://github.com/liao961120/cqls↵
3. Index is constructed with Python's *dictionary*↵
4. This is about half the size of the Sinica Corpus 4.0↵