

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 a researcher wants to use her *own data* for research. Limited by her programing skills, she 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]. Furthermore, concordance search with *Concordancer* is **fast**: simple searches on a large corpus (5 million tokens) could be done in less than a *second*!

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.)

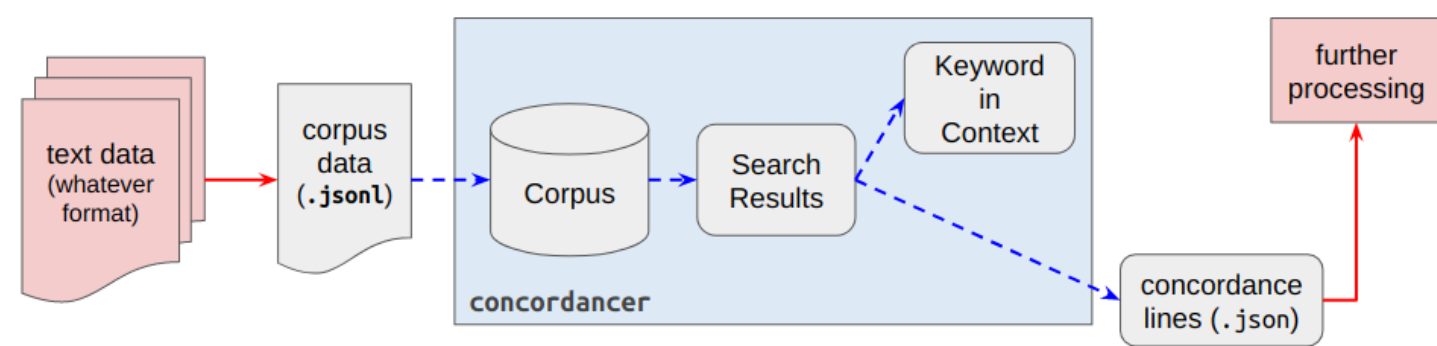


Figure 1: The workflow of using Concordancer

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

最近很想買 Carhartt 的小包可是這	兩/Neu 個/Nf	顏色很難挑。<URL> 土黃 or 黑 —
有刻印 stussy 的 logo <URL> 這支錶有	兩/Neu 個/Nf	錶帶，原本裝的錶帶有點像護腕，
扣子都扣起來了第二套 <URL> <URL> 放	兩/Neu 張/Nf	怕看不清楚這次是百搭
大家好，最近我買了這	兩/Neu 條/Nf	褲子 <URL> <URL> 我本身是胖胖的，小腿
3. 馬卡龍藍 oversize+ 黑褲 + 圓環皮帶（ +	兩/Neu 杯/Nf	飲料 🍹 ） <URL> 🧣 毛衣，都是鏡子自拍沒
的老爹鞋（也是以防一直踩到啦除了	兩/Neu 條/Nf	項鍊是銀色的這回也戴了
書之類的上班上課皆適用已使用大約	兩/Neu 年/Nf	依舊好用雖然現在還好好的但很怕
粗了！但遇到的問題就是這	兩/Neu 個/Nf	顏色搭起衣服來怎麼好像都怪怪的不
earth music 耳夾：gogosing 襪子：tutuanna 頭髮綁法：在	兩/Neu 側/Ncd	耳朵上方綁三股辮，綁好後繞到頭後方
最近給自己買了	兩/Neu 支/Nf	錶當生日禮物（亂花錢的藉口？
耳夾：gladless 頭髮綁法：從耳朵上方抓	兩/Neu 股/Nf	頭髮，用兩股交纏的方式綁到脖子
👉 有事沒事打開一逛就是一	兩/Neu 小時/Na	而且逛的還絕對不能是那種

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 who encounter 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).

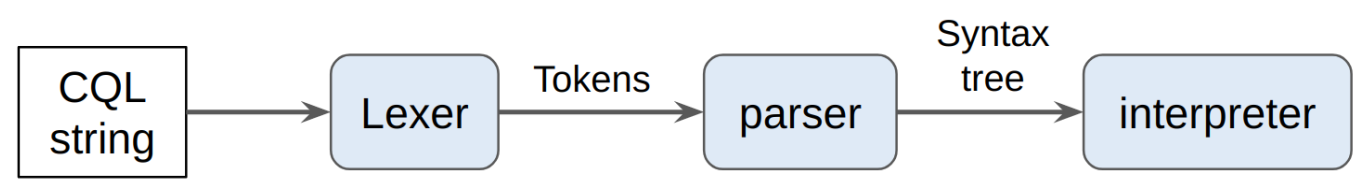


Figure 2: A simple CQL interpreter written in Python

Corpus Indexing

The corpus is indexed (with Python’s *dictionary*) to allow for **efficient searches** [3] in *Concordancer*. A simple search on a corpus of 5 million tokens³ 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 from 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 江琮玉, “小農手作：語料庫索引與建置,” Dec. 12, 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. This is about half the size of the Sinica Corpus 4.0↩

