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.)

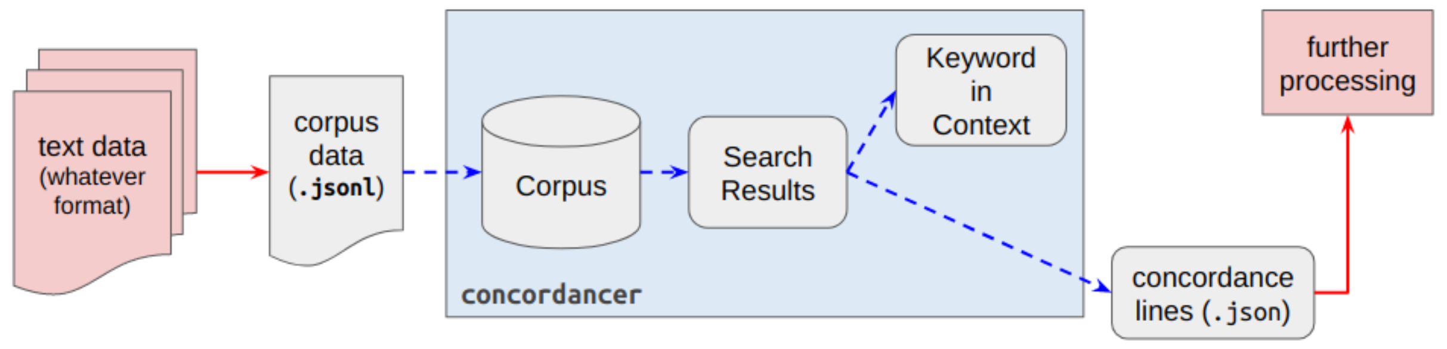
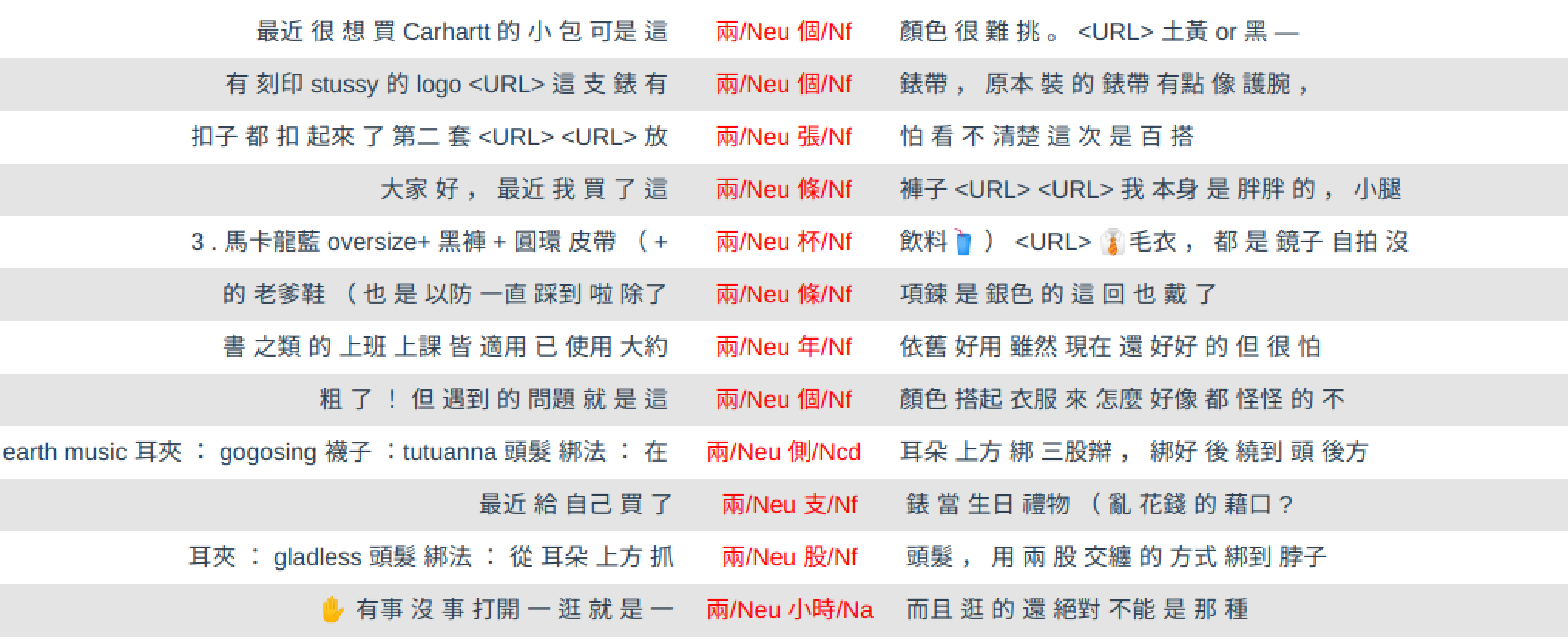


Figure 1: The workflow of using Concordancer

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

bit.ly/pyccd

國立臺灣大學語言學研究所
Graduate Institute of Linguistics, National Taiwan University

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).

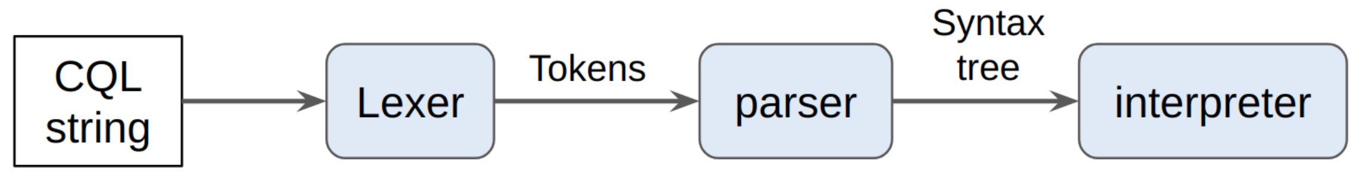


Figure 2: A simple CQL interpreter written in Python

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 tokens[3] 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**[4] (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,"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↩
4. The front-end app originated from a Vue.js project (https://github.com/liao961120/kwic)↩