Tài liệu hướng dẫn sử dụng:

* Đầu vào:
  + 2 file:
    - crawl.cfg: File cấu hình
      * urls = <http://24h.com.vn/> : Url các website để lấy dữ liệu
      * same.domain = true : Chỉ lấy dữ liệu từ các domain chỉ định, không trỏ sang các domain khác
      * output.type = file : 2 giá trị file | folder
      * output.folder = D:/output2/ : output folder
      * output.file = D:/24h.txt : output file
      * max.item.check = 10000 : số lượng item kiểm tra không trùng lặp trước đó
      * id.containers = wrapper\_container,body: chỉ lấy dữ liệu trong ID
      * class.containers = container,content,pContent,main: chỉ lấy dữ liệu trong class
    - url-extracted.vars: Chứa các URL đã quét
* Đầu ra:
  + Xuất ra file hoặc folder đầu ra:
    - Nếu là folder: Mỗi file trong đó là 1 url đã quét. Trong mỗi file, mỗi dòng là 1 câu
    - Nếu là file: Mỗi dòng 1 câu
* Công cụ: Crawler dữ liệu web
* Thư viện sử dụng:
  + stanford-parser-3.5.2.jar
  + jsoup-1.8.3.jar: <http://jsoup.org/>

*“jsoup is a Java library for working with real-world HTML. It provides a very convenient API for extracting and manipulating data, using the best of DOM, CSS, and jquery-like methods.*

*jsoup implements the*[*WHATWG HTML5*](http://whatwg.org/html)*specification, and parses HTML to the same DOM as modern browsers do.*

* *scrape and*[*parse*](http://jsoup.org/cookbook/input/parse-document-from-string)*HTML from a URL, file, or string*
* [*find*](http://jsoup.org/cookbook/extracting-data/selector-syntax)*and extract data, using DOM traversal or CSS selectors*
* [*manipulate*](http://jsoup.org/cookbook/modifying-data/set-html)*the HTML elements, attributes, and text*
* [*clean*](http://jsoup.org/cookbook/cleaning-html/whitelist-sanitizer)*user-submitted content against a safe white-list, to prevent XSS attacks*
* [*output*](http://jsoup.org/apidocs/org/jsoup/select/Elements.html#html--)*tidy HTML*

*jsoup is designed to deal with all varieties of HTML found in the wild; from pristine and validating, to invalid tag-soup; jsoup will create a sensible parse tree.”*

* Thuật toán:
  + Bước 1: Đọc dữ liệu cấu hình. Các URL đã đọc
  + Bước 2: Lần lượt lấy dữ liệu từ các URL thu thập được. Đồng thời thu thập URL mới
  + Bước 3: Tiền xử lý dữ liệu đọc được. (cắt câu, kiểm tra trùng lặp…)
  + Bước 4: Ghi ra file và tiếp tục lặp lại từ bước 2