**WebCrawler for Restaurant Menus:**

From a given list of URLs, this crawler looks for pages that are relevant to restaurants, pages that contain restaurant menu and downloads the menus and drops the remaining pages.

The high-level process can be summarized as:

1. From a Base URL, crawl the home page.
2. Using title, head, page tags identify whether that site will contain Restaurant menu or not?
3. Only for Restaurant menu related sites, find all links
4. Classify the pages that contain the menu and keep the menu pages.

**What is completed:**

* Architecture
* Design
* Finalized tools and technologies.
* Design of base classes

**Key Components:**

Classifier:

Classifier uses title, head and page tags to classify relevance of the page to find out whether a page is a restaurant page or not. If it is a restaurants page, then further crawling is performed.

WebCrawler:

WebCrawler is used to crawl main pages, links under that page etc. It stops crawling page if relevance score goes to zero or lesser (i.e., that site doesn’t contain menu).

Web crawler takes inputs seed data and classifier instance which has to inherit from web classifier.  Using seed URLs, it starts extracting title, head, page tags identify whether that category is restaurant, or the page contains menu etc.,

**Which tasks are pending?**

Training – manually create training data set by using existing known restaurant sites, sites containing menu information, non-restaurant pages and converting them to metapy format.

Code, Integration testing – Developing the scraping code, classifier code and integrate them and test it end2end.

**Are you facing any challenges?**

* Integrating classifier with Scrapy.
* Identifying pages that contains menus from other pages in a site.