API Specification

Introduction

As developers, we all hate paying for APIs. The moment we see the e-payment page, we tend to put our developer hats on, and start coding up a scraper. While this is great for short time, it has two major disadvantages.

a. For every page we want to scrape content off, we would end up creating a scraper. Urghh!!
b. Even the websites you designed your scraper for *(Read : complicated string sorcery of html tags and regular expressions $@*$><||{}@)*, are prone to change, and if someone decides to append an extra semicolon somewhere in the page, cuz, say his Salmon at lunch was overcooked, you can put your hat on again! ☺

What does this API do?

Given a page URL, and a comma-separated list of keywords that you know, for sure, have a pattern, this API first discovers the pattern, and then extracts all keywords that follow this pattern.

How to use this API

`scrapeAnyPageContentWithoutKnowingPattern` ("http://www.alexa.com/topsites/global;", "Google.com,Youtube.com,Facebook.com,Baidu.com,Yahoo.com");

Algorithm

def scrapeAnyPageContentWithoutKnowingPattern(base URL, int limit)

1. String longestcommonPattern = *`discoverPattern`* (baseURL, keywords);
2. output = *`scrapeBasedOnPattern`*(baseURL, longestcommonPattern, 24);

def discoverPattern(baseURL, keywords)

1. for(i in 0 : |keywords|-1)
   sets[i] = all lines from the HTML of page, containing keywords[i]
2. candidatePatterns = set of common patterns b/w sets[0] and sets[1] sorted in decreasing order of length.
3. i = 0
   if(candidatePatterns [i] exist in sets[2: |keywords|-1])
   return candidatePatterns [i]; //found Longest pattern across all keywords
   else
   i++;
   repeat;

def scrapeBasedOnPattern (base URL, String pattern, int limit)

1. scrape pages using longestcommonPattern
2. If HTML contains 'longestcommonPattern', print everything from end of longestcommonPattern
   to the next reserved character for domain names as per RFC 3986 specification.

Example API invocations for finding the longest common pattern, and their output

```
String longestcommonPattern =
discoverPattern("http://www.alexa.com/topsites/global;0",
"Google.com,Youtube.com,Facebook.com,Baidu.com,Yahoo.com");
Output : </div><div class="desc-container"><p class="desc-paragraph"><a
href="/siteinfo/
```

```
String longestcommonPattern = discoverPattern("https://moz.com/top500",
"Google.com,Youtube.com,Facebook.com,Yahoo.com");
Output : " target="_blank">
```

```
String longestcommonPattern = discoverPattern("https://www.quantcast.com/top-
sites", "Google.com,Youtube.com,Facebook.com,Yahoo.com");
Output : <img class="favicon" name="
```

```
String longestcommonPattern =
discoverPattern("https://www.similarweb.com/global",
"Google.com,Youtube.com,Facebook.com,Yahoo.com");
Output : <img class="lazy lazy-icon" itemprop="image" data-
original="https://site-images.similarcdn.com/image?url=
```

Clarification, What this API does not do

1. Let me clarify that **def discoverPattern** is the procedure that is general across all pages and would work without knowing the pattern.
2. Depending on the structure of the page, you would still have to write some custom code. For example, alexa.com displays only 25 results per page and the pages look like global;0, global;1 and so on, whereas, https://moz.com/top500 displays all 500 links on the same page. This is highly dependent on site architecture and needs developer intervention.