Hello!

I feel I need to leave at least a small comment on all of this.

Well, there is nothing special about crawler. Small and extremely simple.

I thought that using MySQL would be an overkill, so I`ve chosen SQLite. It has only one table with id, url, raw_page and datetime fields. There schema is not suitable for big data, but is fine enough for this task.

Since there are only 14 urls, it is not feasible to use statistics or ml: (But still, I tried to keep my algorithm general and not customize it too much.

It`s basic assumptions are:

1) Texts are grammatically correct;

2) Pictures are possibly logos, so their “alt” attribute matters;

3) Titles matter;

4) If there is company website, it would resemble company name;

5) Company names consist only of letters.

So I just used frequency dictionary of words starting with an upper letter and frequency dictionary of websites found on this page. Than I computed the levenstein measure between urls and words to find the closest.

The problem I didn`t solve was that I can`t tell if a job is expired. I think I could create a dictionary of stop phrases like “page not found” or “job is expired” or “no longer availavle”, but I doubt that would work with other pages, so I didn`t do it.

Thank for attention!