# Step by Step web scraping Pseudo Algorithm

**Step 1:** define a function to extract all links on a webpage and return a list.

**Step 2:** start with the parent website(layer 0) and extract all child data frame and store it in a child data frame. Each row in the data frame contains the child link, the parent link, and the layer number of the child link.

**Step 3:** Store the website already visited in the visited data frame.

**Step 4:** Iterate as the parent website is appended in the visited data frame, the child list becomes the new parent and a new child is instantiated.

**Step 5:** After each iteration add the parent data frame to the visited data frame, and increment the layer.

**Step 6**: Stop if the child link is outside the domain of the website, or if its already in the visited data frame.

**Step 7:** After the data frame containing all the webpages in the website is accumulated save the data from the websites using pymongo.

**Step 8:** Repeat steps 1 to 7 for rest of the websites

# Challenges

1. Identifying duplicate links with slightly different stinrgs, (Signature of a Page)
2. Creating MonDB data base