# Cách lấy dữ liệu

* Vì các web có số lượng bài viết lớn nên đều có phân trang, scraping dữ liệu dựa theo từng trang của phân trang này.
* Trước tiên vào từng trang của phân trang lấy các link bài viết. Các trang đều chạy từ 1(trang đầu của phân trang) đến trang cuối. tùy theo đặc điểm của từng trang mà trang cuối có thể tìm bằng cách khác nhau, có 1 số trang ghi trang cuối ngay dưới và 1 số khác không có. Sau khi có các link bài viết ta vào từng link lấy dữ liệu.

# Đặc điểm của các web

* VNExpress
  + Lấy trang lớn nhất bằng cách kiểm tra từng link của trang có hợp lệ không, hợp lệ là số trang trong link nhỏ hơn hoặc bằng số trang lớn nhất. Nếu qua số trang lớn nhất link sẽ bị redirect về trang đầu, đây là điều kiện để xem link có hợp lệ không. Các bước kiểm tra đã được giảm đi đáng kể nhờ thuật toán.
* VietNamPlus
  + Số trang lớn nhất có ghi ở cuối mỗi trang, chỉ cần chạy từ trang đầu đến trang cuối lấy dữ liệu.
* BaoBinhDuong
  + Số trang lớn nhất không có tuy nhiên nếu vượt qua trang lớn nhất thì số trang lớn nhất sẽ hiện ở cuối trang.
* SGGP
  + Giống VNExpress
* ThoiDai
  + Bên tiếng việt nó hiện trang lớn nhất ở cuối.
  + Bên tiếng trung không có hiện, giống VNExpress chỉ khác link không hợp lệ là link mà không có các danh sách bài viết, kiểm tra có empty hay không.
* BaoThaiNguyen
  + hiện số trang lớn nhất ở dưới góc.
  + Tiếng trung chỉ là 1 mục trong menu của nó
* NhanDan, TapChiCongSan
  + Tìm trang lớn nhất bằng thuật toán
* BaoChinhPhu
  + hiện số trang lớn nhất ở dưới góc.
  + Bên tiếng trung chạy như rùa.

n is number date, time similarity 100% between 2 files;

percentSimilarity =

|  |  |
| --- | --- |
| Politic | |
| Percent similiraty | Number Files |
| 8.333334 % | 4 |
| 9.090909 % | 2 |
| 10.0 % | 14 |
| 11.111112 % | 5 |
| 12.5 % | 6 |
| 14.285715 % | 12 |
| 16.666668 % | 48 |
| 20.0 % | 50 |
| 22.222223 % | 3 |
| 25.0 % | 70 |
| 28.57143 % | 1 |
| 33.333336 % | 132 |
| 37.5 % | 3 |
| 40.0 % | 8 |
| 42.857143 % | 4 |
| 50.0% | 326 |
| 66.66667 % | 41 |
| 75.0 % | 7 |
| 85.71429 % | 1 |
| 100.0 % | 462 |
| Total | 1199 |