KỸ THUẬT LẬP TRÌNH C#

**HtmlAgilityPack và Ứng dụng phân tích dữ liệu web**

Phan Minh Tài | *Founder of CafeT.vn|ChuyenTin.vn|…*

**MỞ ĐẦU**

Cách đây ít lâu tôi thử xây dựng một WebCrawler nho nhỏ để chọn lọc tin tức từ một số website yêu thích và trích xuất những bài viết/tin tức hữu dụng cho bản thân. Thông qua việc này tôi có sử dụng thư viện mã nguồn mở HtmlAgilityPack và thấy rằng nó khá hữu dụng, do vậy tôi dành chút thời gian để chia sẻ một số vấn đề liên quan đến chủ đề Phân tích và Xử lý dữ liệu Web cho các bạn đang quan tâm.

**GIỚI THIỆU HtmlAgilityPack**

**Cài đặt (Nuget)**

Install-Package HtmlAgilityPack -Version 1.6.0

**SỬ DỤNG HtmlAgilityPack**

**MỘT SỐ BÀI TOÁN THỰC TẾ**

**Bài toán 01:** Tìm HtmlNode có chứa nội dung bài viết của một Url

Thông thường một địa chỉ url đã tồn tại (chẳng hạn các website chuyên về tin tức) sẽ có cấu trúc như sau:

* HtmlDocument
  + HtmlNodes