**HTML Parsing using Nokogiri**

**/Gemfile**

***source 'https://rubygems.org'***

***gem 'nokogiri'***

**Sample ruby file to fetch and parse google news response**

**/Sample\_parser.rb**

***require 'open-uri'***

***require 'nokogiri'***

***def parse\_my\_code***

***# Parse the URI and retrieve it to a temporary file***

***tmpfile = open('https://news.google.com')***

***# Parse the contents of the temporary file as HTML***

***doc = Nokogiri::HTML(tmpfile)***

***# Define the css selectors to be used for extractions, most***

***article\_css\_class =".esc-layout-article-cell"***

***article\_header\_css\_class ="span.titletext"***

***article\_summary\_css\_class =".esc-lead-snippet-wrapper"***

***# extract all the articles***

***articles = doc.css(article\_css\_class)***

***#html output***

***html = ""***

***#extract the title from the articles***

***articles.each do |article|***

***title\_nodes = article.css(article\_header\_css\_class)***

***# since there are multiple titles for each entry on google news***

***# for this demo we only want the first (topmost)***

***#***

***# its very easy to do, since title\_nodes is of type NodeSet which implements Enumerable***

***# > title\_nodes.class***

***# => Nokogiri::XML::NodeSet***

***# > title\_nodes.class.ancestors***

***# => [Nokogiri::XML::NodeSet, Enumerable, Object, Kernel, BasicObject]***

***prime\_title = title\_nodes.first***

***# Even when the css selector returns only one element, its type is also Nokogiri::XML::NodeSet***

***summary\_node = article.css(article\_summary\_css\_class)***

***# > summary\_node.class***

***# => Nokogiri::XML::NodeSet***

***# > summary\_node.size***

***# => 1***

***# Create an "---------" line for the title***

***separator = "-" \* prime\_title.text.size***

***# Extracting the text from an Nokogiri::XML::Element is easy by calling the #text method,***

***# notice how we can also do it on the NodeSet,***

***# there it as a different semantic by invoking #text in all the children nodes***

***html += "%s\n%s\n%s\n\n\n" % [prime\_title.text, separator, summary\_node.text]***

***end***

***puts html***

***end***

***parse\_my\_code***

**Run ruby file in console**

***$ ruby Sample\_parser.rb***

**Output:**

***After midnight vote, Trump ahead of Clinton 32-25 in New Hampshire in US Presidential poll***

***------------------------------------------------------------------------------------------***

***NEW DELHI: In a traditional midnight vote+ held in three tiny towns in the state of New Hampshire+ , US Republican candidate Donald Trump came out ahe***

***ad of Democratic opponent Hillary Clinton in the US Presidential elections, reported USA Today.***

***Blast kills child, injures 32 Indian peacekeepers in east Congo: UN***

***-------------------------------------------------------------------***

***KINSHASA: An explosion killed a child and injured 32 Indian peacekeepers in the eastern Democratic Republic of Congo city of Goma on Tuesday, the U.N.***

***People are fuming at this photo of a JNU student being 'manhandled' by Delhi Police***

***-----------------------------------------------------------------------------------***

***Recently, a haunting picture of Najeeb Ahmad's mother being pulled by Delhi Police shook the Internet. By: Trends Desk Published:Nov 8, 2016, 14:28.***