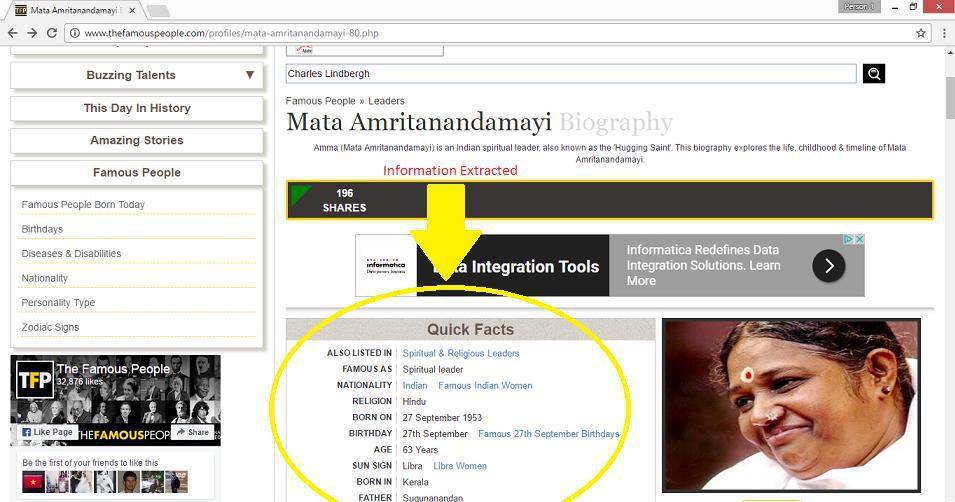
1. **Data Source Description:**

* The data source used in the assignment is a Biography Website named “The Famous People” ([http://www.thefamouspeople.com](http://www.thefamouspeople.com/))
* Thefamouspeople.com chronicles the life history of some of the world's most famous people and achievers. The biographies of these people feature the achievements and works that have influenced the course of history.
* The reason I choose this data source is that it gives good amount of information for the Wrapper to extract which is explained in the next section.

1. **Fields Extracted:**

* The fields extracted are the basic details about the person referred.
* Following is the list of fields and their description used in the extraction

|  |  |
| --- | --- |
| **Field Name** | **Description** |
| **URL** | The page URL referring the Person |
| **Name** | Name of the Person |
| **Famous As** | Famous as (example Writer, Singer) |
| **Nationality** | Country to which he/she belongs |
| **Religion** | Religion of the Person |
| **Born On** | Birth date |
| **Age** | Age of the Person |
| **Sun Sign** | Sun sign of the person |
| **Born In** | City/State where he is born |
| **Father** | Name of Father of the Person |
| **Mother** | Name of Mother of the Person |
| **Spouse** | If married then name of the Person |
| **Founder/Co-Founder** | Institutes founded by the Person |
| **Awards** | Awards the Person achieved |
| **Died on** | Date when he died |
|  | **Table 1 Fields Description** |

****

**Table 2 Webpage Screenshot**

1. **Tools:**

* The tools used for crawling and data extraction are Crawler4j and JSoup Libraries available in Java.
* Crawler4j is an open source web crawler for Java which provides a simple interface for crawling the Web. Using it, you can setup a multi-threaded web crawler in few minutes. Crawler4j JARs are available on the [releases page](https://github.com/yasserg/crawler4j/releases) and at [Maven Central.](http://search.maven.org/#search%7Cgav%7C1%7Cg%3A%22edu.uci.ics%22%20AND%20a%3A%22crawler4j%22)
* JSoup is a Java library for working with real-world HTML. It provides a very convenient API for extracting and manipulating data, using the best of DOM, CSS, and JQuery-like methods. For downloads you can refer [jsoup.org](http://jsoup.org/)
* JSoup implements the [WHATWG HTML](http://whatwg.org/html) specification, and parses HTML to the same DOM as modern browsers do:
  + - * Parse HTML from a URL, file, or string
      * Find and extract data, using DOM traversal or CSS selectors
      * Manipulate the HTML elements, attributes, and text
      * Clean user-submitted content against a safe white-list, to prevent XSS
      * Output tidy HTML

1. **Wrapper Description and Implementation:**

* The website was analysed and the data/fields are extracted using the structure of the HTML page downloaded. Thus it is a DOM Tree Based Manual Wrapper. The advantage of the wrapper is that it is very easy to implement and extract data from the HTML page.
* The disadvantage is that if the website structure changes in future the wrapper may detect false fields and needs to be inducted to accommodate the changes.
* The wrapper model is LR Wrapping as it detected the start and end tags of the DIV elements from the DOM element for the field extraction for each of the k fields.
* The **Wrapper Algorithm** is:
  + 1. Start Crawling with seed URL as “[http://www.thefamouspeople.com](http://www.thefamouspeople.com/)” using Crawler4j
    2. For each page

If the URL starts with “[http://www.thefamouspeople.com](http://www.thefamouspeople.com/)”and ends with “.\*.php” Then

Get the DOM elements using JSoup

Get the container of the fields DIV indentified by class name “quickfacts”

Extract the child DIV elements for the labels and data

Store the data into a JSON object and write into the extractions.json

* Typical JSON object will be as shown below

{"Born on":"16 September 1838 AD”,"Died At Age":"77","education":"Rockwood Academy (1852)","Spouse\/Partner":"Mary Theresa Mehegan (m. 1867)","Nationality":"Canadian, American","Born in":"Eramosa Township, Ontario, Upper Canada","Sun Sign":"Virgo",","Died on":"29 May 1916 AD","Net worth":"$10 million","name":"James J. Hill Biography","place of death":"St. Paul, Minnesota","URL":"http:\/\/www.thefamouspeople.com\/profiles\/james-j-hill-138.php","Famous as":"Canadian-American railroad executive, Businessman"}