**Edureka Python Project Documentation**

**Problem Statement**

IMDB provides a list of celebrities born on the current date. Below is the link: [http://m.imdb.com/feature/bornondate](http://m.imdb.com/feature/bornondate" \t "_blank)

Get the list of these celebrities from this webpage using web scraping (the ones that are displayed i.e top 10). You have to extract the below information:

1. Name of the celebrity
2. Celebrity Image
3. Profession
4. Best Work

Once you have this list, run a sentiment analysis on twitter for each celebrity and finally the output should be in the below format

1. Name of the celebrity:
2. Celebrity Image:
3. Profession:
4. Best Work:
5. Overall Sentiment on Twitter: Positive, Negative or Neutral

Hint: Use IMDB scrapping sample example as reference for scraping the mentioned web page. For sentiment analysis use the Twitter sentiment code as reference.

Python version 2.7

**Tools and Packages Used**

• Version: Python 2.7[VERY IMPORTANT]

 Tweepy: An easy-to-use Python library for accessing the Twitter API.

*Text Blob:*

*Text Blob* is a Python (2 and 3) library for processing textual data. It provides a simple API for diving into common natural language processing (NLP) tasks such as part-of-speech tagging, noun phrase extraction, sentiment analysis, classification, translation, and more.

• Beautiful Soup [Symbol] Beautiful Soup provides a few simple methods and Pythonic idioms for navigating, searching, and modifying a parse tree using Python parsers like lxml and html5lib. It automatically converts incoming documents to Unicode and outgoing documents to UTF-8. Here's the documentation.

• Selenium [Symbol] The web driver kit emulates a web-browser (I chose chrome driver) and executes the JS scripts to load the dynamic content.

**Challenges Faced during the project**

Well as being a complete newbie in programming it took more time to understand the programming concept but the time spent for this project was worth every second.

I understand it took little  longer time for me to complete the project .Initially when I started working on IMDB project I used just beautiful soup to scrape the content from IMDB website however I was unable to scrape the required content from the website as there was dynamic script used for images and other contents of the website.

Did thought of using regular expression to scrape the information from IMDB  but still because of JavaScript I couldn’t scrape the required details from the website.

Did took help of “Edureka” team to know what needs to be done in order to scrape the JavaScript content from IMDB website . As per the suggestion from Edureka team  downloaded the selenium driver and also went ahead and  installed the chrome driver (resources used to understand the functionality of selenium driver was YouTube and also Selenium web driver website: <http://www.seleniumhq.org/docs/01_introducing_selenium.jsp>

Other references:

Reference: [http://fruchter.co/post/53164489086/python-headless-web-browser-scraping-on-amazon](http://fruchter.co/post/53164489086/python-headless-web-browser-scraping-on-amazon" \t "_blank)

Used text blob library to evaluate the sentiment analysis of tweets:

Reference: <https://textblob.readthedocs.io/en/dev/>