NEWS CLASSIFICATION AND ANALYSIS – PHASE 1

News websites are one of the best sources of information all around the world. It is one of the fastest and appropriate mode of information sharing. Newspapers and news website make all the regions of the world connected in real time. News sites are now a days designed to give a hassle-free experience to users and therefore they are designed by keeping in mind the ease of readability, access to different verticals, crispness of the content, quality. There is a profound team of people working days and nights to bring the crucial information to the end users on daily basis. It requires a well-qualified team of reporters, writers, proof-readers, bench markers, software developers, Analysts and data scientists.

These websites not only live in present day, but keep record of all past events from history in the form of archives to deliver a competent source of information. This historical information is crucial in analysing past behaviours and decision making over different domains. Policy makers use the analysis on the historical data to design new policies and take better decisions.

PROBLEM STATEMENT:

News archives contain several Gigabytes of data, which is not easy to handle. For simplicity they are stored in simplest format possible. But during analysis, analysts require a meta information about each news article. Sometimes it is available but there are cases where meta information is missing. This is where data science comes into play. Use cases are built to analyse the available meta data and hence analyse meta information about the news articles, where it is not present. In this project we have to build such use case where we can generate the metadata around news articles and make it ready for further analysis. In phase 1, we need to gather information from archives of past few years from different news websites. In this phase we will scrape the news websites for news and related information.

The task assigned to you is to scrape archives from **Deccan Herald** website for years 2015 to 2019. You will be allotted tasks to scrape data based on different timeframes. The details to be scraped are given below:

1. **Date**: This column will contain the date on which news was reported.
2. **Author**: This column will contain name of the author of news report.
3. **Vertical**: This column will contain the news verticals (e.g. Entertainment, Sports, Lifestyle, Economy, etc.).
4. **Headline**: This column will contain the news headline.
5. **Description**: This column will contain the main content of the news report.

You have to write a code to scrape the above data from **Deccan Herald** website. The time frames of the archives will be allocated to you individually. The data scraped has to be stored in three formats, viz. **csv**, **json** and **xml.**

Note: Kindly refer the sample_dh.csv which contains the sample data for your reference.

SUBMISSION:

You have to submit the following files:

1. The Jupyter notebook or python code file containing your code.
2. The csv, json and xml files in which you have to store the scrapped data.

You will be provided a link to upload the data you scraped. Kindly upload the data on the provided link.