Seasonal analysis of tourist attractions applying Web scraping and Data Visualization

Srinidhi Hiriyannaiah
Assistant Professor
Dept of CSE
Ramaiah Institute
of Technology
srinidhih@msrit.edu

Koushik A S
Student
Dept of CSE
Ramaiah Institute
of Technology
askoushik4@gmail.com

Akanksh B S
Student
Dept of CSE
Ramaiah Institute
of Technology
bsakanksh@gmail.com

Anirudha V Bharadwaj
Student
Dept of ECE
Ramaiah Institute
of Technology
anirudhavb97@gmail.com

***Abstract*—Travel and Tourism is an important area for any country. With the emergence of strong global middle class there is an increase in demand for tourist places. These tourist places have a lot to offer in terms of trekking, river rafting, sightseeing, cultural festivals and the list goes on. But unfortunately, lot of tourists do not visit these places at the right time. This often deteriorates their experience and thereby making them spread a bad word. This is one of the main reasons why tourist places in India remain unexplored. Tourists are often mislead as different websites suggest different visiting seasons. Our goal is to collect data from various websites using a technique called web scraping and processing the collected data using data integration. The above processed data is visualized and presented in a user friendly manner. This provides an insight to tourists as to when to visit a particular place so that they have a pleasurable experience and enjoy to the fullest possible extent.**

***Index Terms*—IEEEtran, journal, LaTeX, paper, template.**

I. Introduction

A. The significance of Travel and Tourism

Travel and Tourism is an important aspect for a nation.[1] Travel and Tourism plays a vital role in contributing to national GDP and also creates a ton of new jobs and opportunities every year. It is no different for India. Statistics say that the total contribution of Travel and Tourism to GDP was INR14,018.5 billion (USD 208.9 billion). In 2016, the total contribution of Travel and Tourism to employment, including jobs indirectly supported by the industry was 9.3% of total employment (40,343,000 jobs). This is expected to rise by 1.8% in 2017 to 41,074,000 jobs and is speculated to rise exponentially. If a tourist enjoys his time, he will definetely recommend it to others and share his good experience. Therefore, it is of utmost importance to ensure that the tourist enjoys the place to its full extent.

B. Why Web Scraping?

Web Scraping[2] (also termed Screen Scraping, Web Data Extraction, Web Harvesting etc.) is a technique employed to extract large amounts of data from websites whereby the data is procured and saved to a local file in the computer or to a database in popular formats like CSV and Json.

Data displayed by most websites can only be viewed using a web browser. They do not offer the functionality to save a copy of this data for personal use. The naive idea then is to manually copy and paste the data which is a very tedious job that usually takes many hours or sometimes days together to complete. Web Scraping is a method of automating this process which instead of manually copying the data from websites performs the same task within a fraction of time.

C. What is D3.js?

D3.js[3] is a JavaScript library for manipulating documents based on data. D3 is a powerful tool in bringing data to life using HTML, SVG and CSS. D3s emphasis on web standards gives full capabilities of modern browsers without trying too hard to work on a proprietary framework. It allows combining powerful visualization components and facilitates a data-driven approach to DOM manipulation. Finally data can be made interactive through the use of D3.js which provides data-driven transitions and transformations.

II. Obtaining the Dataset and Data Pre-Processing

The first step is to form a data set from the data acquired from various web sites. We started off by gathering information featuring almost all the tourist places in Karnataka. The reason for choosing one state is to start implementing from a small dataset and visualize the same. The same techniques can be extended to other states as well. We have around forty major tourist places in Karnataka which have to be explored at the right time.

We started off by finding out different websites and travel blogs which suggested best time to visit various places. The major hurdle was to gather data which was scattered across different websites. However we found out eight websites which had suggestion for three or more places. Goibibo, makemytrip, holidayiq, yatra were few among the other websites that we looked for. The next step was to access the webpages concerning to all the individual tourist places from these websites in a loop so that we could collect the details. To access different webpages, we split URL into two halves for instance "https://www.yatra.com/india-tourism/" + place + "/best-time-to-visit". The reason for doing this is to access all webpages from the website easily just by changing place in the loop.

Once we have the URL, we make a get request. If the request is successful response = 200 is received, otherwise details about that place is not available in that webpage. In case the request is successful, we make use of the Python library, Beautiful Soup[4] to access the contents of the desired webpage. This library has a feature to access data from a particular class or id in a html document. The data obtained pose some problems because of the nature of the language used which is inherently unstructured. In every webpage, the required data that is the best season to visit was distributed all throughout the article. Therefore, we break this unstructured data into individual sentences, by using split command. The main purpose of splitting is to locate keywords which convey the same meaning as of ideal time such as "great months", "best season", "best time, "best months", "most favourable time" and a few more in order to identify the right sentences which bear the required data.. We carry out the above said procedure on all the websites and tabulate a Json file.

We also require a shape file to represent the data in a visually interactive way. Shape file is a vector data storage format for storing the location, shape and attributes regarding geographic features. It basically stores a set of related files and contains one feature class which can be used to represent the details of a particular location on the map. India map shapefile was downloaded which contained all the states and their respective districts as its attributes.

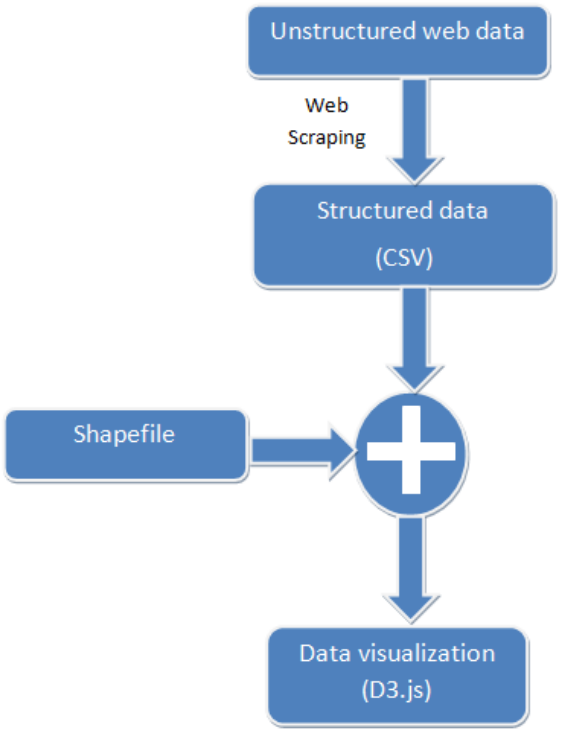

Fig 1 : Flowchart representing the methodology

III. Data Integration

The data that we acquired through various websites indicated the ideal period to visit a particular place by showing the details of start month, end month and the district to which the place belonged to. The information gathered from various websites coincided majority of the times, however there was a fair amount of deviation in certain cases. So one obstacle ahead of us was to integrate the accumulated data in the most appropriate possible manner. The two choices available to us were to either combine the data in a way which suggests a union of all the accumulated time periods or its intersection.

Though the seemingly indigenous idea of a union seems interesting as it equips tourists with the flexibility of visiting the tourist attractions by providing a wider time range, it looses out on data accuracy. In order to address this problem, intersection was considered. The

idea behind implementing intersection was to provide tourists with a clear insight to enjoy their trip though having a downside of a narrow interval. Clever use of the basic concept of a one dimensional array did the trick. While iterating through the months, count of the particular month was incremented which was represented by the array index as shown Fig 2. The months which received the maximum increments were considered and others were discarded. This implementation facilitated to have a CSV file which could easily be ported.

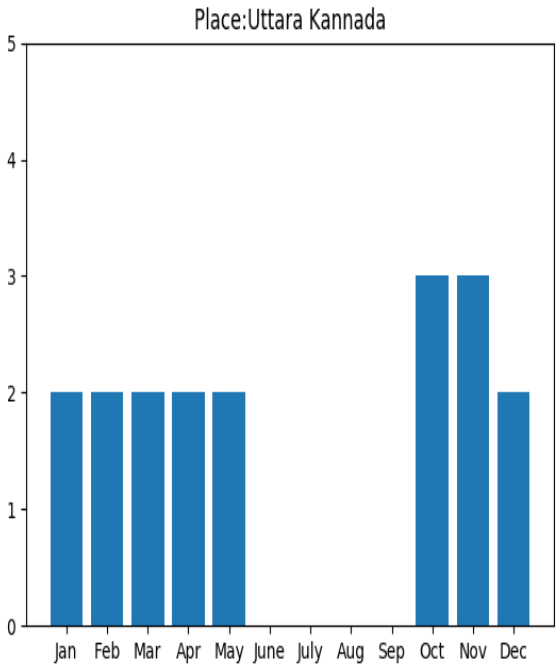

Fig 2 : Bar graph suggesting the best season of a particular place according to different websites

IV. Combining integrated data with shape file

To combine the obtained csv file with the shapefile we used QGIS.[5] QGIS is a free to use open source geographic Information System, which can be used to work with a shape file. We unified the two documents with the help of Join Tab. We merged the columns containing districts which were in common in both the csv and the existing shapefile in order to obtain the modified shapefile. This shapefile consists the additional attributes of tourist places, their respective districts, starting and ending months of the favourable season.

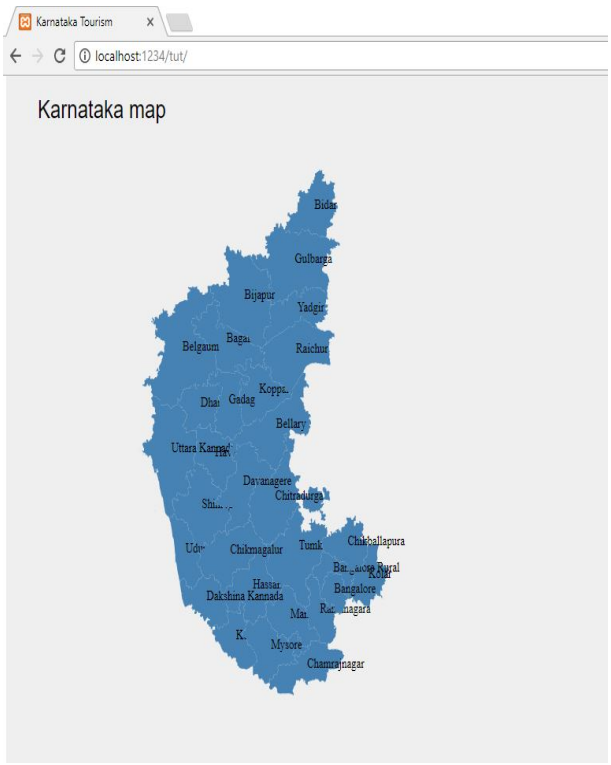

Fig 3 : Visual representation of Shapefile prior to data integration

V. Visual representation of given data

To analyse the collected data, visualisation is an interactive option. Data visualisation is representation of given data in the form of charts, diagrams, pictures and maps. Data visualisation is highly appealing and aesthetically pleasing. The whole purpose of employing data visualisation was to reduce complexity and put across information regarding the seasons in a user friendly way which in turn makes the user experience hassle free. Geographic locations are best represented in a map. This provided the motivation to make the user more comfortable just by hovering around the map rather than experiencing the overhead of searching the best season for different places individually. D3.js a data visualisation library in java script which is a perfect tool for this form of visualisation.[6] D3.js has many functions to work with maps and shape files. This enables to create interactive webpages with maps .

First step is to create an SVG element in the webpage. SVG (Scalable Vector Graphics) is an XML based vector image format for two-dimensional graphics. SVG is needed to represent the given shape file in a webpage. Map in shape file needs to be represented through a projection. D3.js supports different types of Geo projections such as Albers, AzimuthalEqualArea, Mercator etc. We chose Mercator as projection since it is commonly used for navigation. When a particular district in a map is selected, this would select a particular object. This object is going to return the names of tourist places, starting month and ending month highlighting the ideal season of that particular district. These details are going to be appended to the webpage.

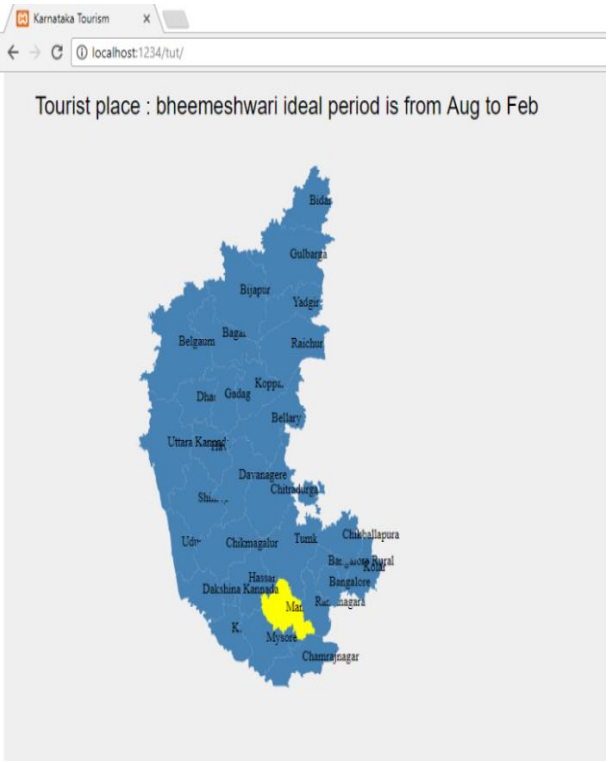

Fig 4 : Map showing best time to visit Bheemeshwari

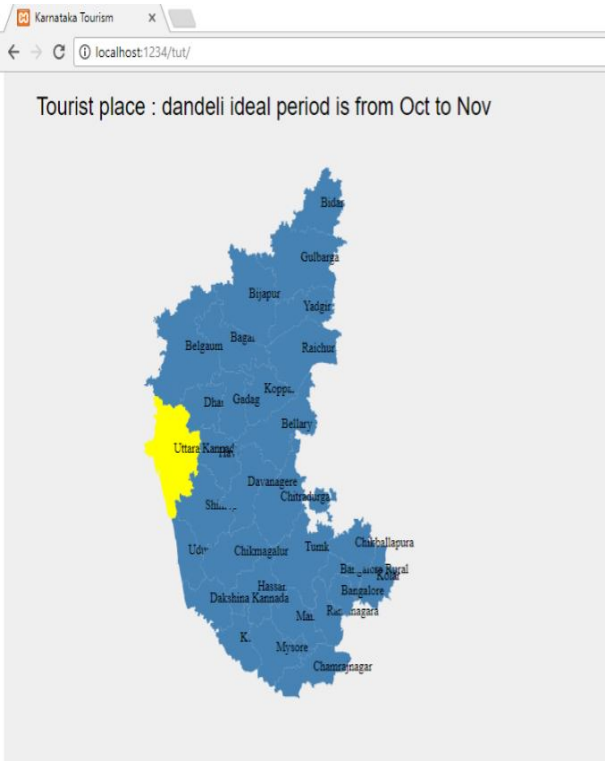

Fig 5 : Map showing best time to visit Dandeli

VI. Conclusion

Tourism has and will have a significant contribution towards the economic growth of the nation. However many of our tourist places have remained unexplored because of their seasonal nature. The main goal of our project was to guide travellers through their vacation by providing them with an insight regarding the ideal time to visit the tourist attractions. Through our work we aim to provide facts and figures to tourists in a consolidated manner. Fluctuating suggestions available across various websites have been combined to provide data which is concise and consistent.

Our application was designed keeping in mind to provide ease of use to the end user without compromising on functionality. We have employed certain visualization techniques which are not only straightforward and intuitive to use but also very effective. It enables people to just stay in one page and hover around the map and get the required information.

References

[1] Economic impact of tourism in India
[2] Web scraping
[3] Data visualization with D3.js
[4] Web Scraping with Beautiful Soup library
[5] Data Integration with shapefile using Qgis
[6] Geo mapping with D3.js