Data Mining & Warehousing

Lab Assignment – 3 (Mini-project on Visualization and Decision Support)

In this assignment, you will need to programmatically access any e-travel website (such as www.booking.com, agoda.com, yatra.com etc.) to mine data about hotel reviews, statistically analyze the data and visually present your results.

The goal of this assignment is to ensure that you are trained in handling real-world data. Of course, when you deal with real-world data, you will need to work on peripheral activities such as data cleaning, pre-processing the data etc., before you do data mining. This assignment will also train you to deal with unstructured text data. You will also be learning Weka while doing this assignment. (Materials about Weka will be provided and discussed in class.) You will also receive exposure to some important technologies such as Nominatum / MapBox, NLTK etc.

Visit any ONE of these websites and mine the listed data points:

1. The rating for a hotel (Overall Rating)
2. List of comments for a given hotel + rating provided (In reverse chronological order. i.e. newest first)
3. Be sure to mine at least 100 comments for each hotel.
4. Mine the information for 20 hotels in Bangkok, Singapore & Kuala Lumpur **each.**

(Thus in total, you will need to mine: 20 (hotels) x100 (comments each) x3 (cities) = 6k comments)

Save all your mined data in an appropriate format.

After you've completed mining your data, complete the following analysis:

1. Cluster the hotels that have similar ratings (overall rating)
    a. You can use WEKA (or any free software) for performing as well as visualizing the clustering
2. Perform sentiment analysis on the comments.
    a. You can use NLTK's inbuilt polarity score to assign a score to each comment.
    b. Check the co-relation between the Polarity score of each comment and the actual rating provided by the user (the rating linked with that comment).
    c. Aggregate (mean/median) the polarity score of all the comments and calculate an overall polarity score for each hotel. Then check for any co-relation between the overall polarity score and the overall rating of the hotel.
3. Visualize the above results. Be sure to highlight any particular anomalies in your visuals.
4. Use any mapping resource (like Nominatim or MapBox) to geospatially visualize the top 10 hotels in each of the 3 cities. Make 2 maps based on the two scoring mechanisms available:
    a. Top 10 hotels based on overall rating of the e-travel website
    b. Top 10 hotels based on overall calculated polarity score

For this assignment, submit:

- Your final mined dataset.

- A two-page report that contains your visualizations with interesting insights. PDF only. Times New Roman 11 point font size.

You will need to demo your code & display your visualizations in a code review session.

This assignment is worth 20 marks.

Deadline for submission: **11:59 PM IST on 17th April, 2021.**