ACCENTUATING OMAN TRAFFIC USING BIG DATA ANALYTICS BY EXCAVATING WEBLOG & SOCIAL NETWORKING DATA

Title: ACCENTUATING OMAN TRAFFIC USING BIG DATA ANALYTICS BY EXCAVATING WEBLOG & SOCIAL NETWORKING DATA
Proposal ID: Proposal Id not specified
Type of project application: Faculty Mentored Undergraduate Research Award Program (FURAP)

Sector Name

Information Technology and Communication

Team Leader

AL-ABRI, MALLAK (s26s1598@nct.edu.om)

Faculty Mentor

Faisal, Dr. Mohammed (mohammed.faisal@nct.edu.om) (Data Sciences (ENGINEERING AND TECHNOLOGY)) (Software Engineering and Programming (ENGINEERING AND TECHNOLOGY))

Other Team Members

AL YAHYAI, SHUROOQ (s26s1593@nct.edu.om)

Project Description:

Now a days Big Data Analytics and relevant technologies is the biggest buzz in the industry. In connection with Big Data, we cannot ignore the social media and it’s important in daily life, by using social media data, it could be twitter data or web log data, we can perform plethora of analytics, and result of analytics can be used to take corrective measures for the betterment of Oman traffic. In this proposal we are considering Oman traffic to find out the reason and feedback of civilian by mining the data from web log and twitter. Proposed research work will be developed using Hadoop ecosystem/Rapid Miner Studio 8.0/Apache Spark along Flume with data retrieving application. Proposed work will incorporate different verticals and KPI’s to visualize the result. In the final report of the analysis is expected to illustrate the opinion of the Oman civilians by performing natural text processing and expected to find the reasons of accident. Research will be carry forward with CRISP methodology stages on the wheel of Big Data Technologies.

Project Significance:

- **Collect information about traffic:** In this project, we will collect huge data through twitter application about all traffic in Oman.

- **Create a data warehouse about traffic:** In this project, we will collect huge data about traffic with the help of various tools we will first store the data collected from twitter application in HDFS.

- **Search the reasons of traffic problem using big data:** In this project, we will use big data tools live hive and flume on top of Hadoop to analyze and relevant information will be derived to find the possible reasons for unsatisfactory traffics in Oman.

- **Producing the analysis report using big data analytics:** After analyzing the data we will produce the analysis report which will help to take decision in order to reduce traffic hassles in Oman.

Methodology

Proposed Research project will be used **CRISP Methododoly** with the following algorithm:
Step 1: Create Twitter application using the Twitter Application Management

- Username: xyz
- Email: xyz@nct.edu.om

Step 2:Take the necessary keys from Twitter app and obtain the followings:

- Consumer key
- Consumer Secret
- Access Token
- Access Token Secret

Step 3:Fetch the data using Flume using Oman traffic, accident as keywords.
Download supported jars:

- twitter4j-core-4.0.2.jar
- twitter4j-stream-4.0.2.jar
- twitter4j-media-support-4.0.2.jar

Step 4:Create external table for loading the data fetched from twitter as key and value using Hive.

Step 5:Split each tweets into words

Step 6:Create dictionary text file that contain thousands of positive and negative words with their rating.

Step 7:Mapping the words split from tweets with words in the dictionary to count how many positive and negative words.

Step 8:Count the percentage of positive and negative.

Step 9: Find out the reasons of accident by finding and KPI's for instance find out the keywords like alcohol, drink, mobile etc as a reason.

Step 10: Visualize the result using RapidMiner/Excel graph/Tableau application.

CRISP Methodolgy major steps:

1. **Business understanding**
2. **Data understanding**
3. **Data preparation**
4. **Modeling**
5. **Evaluation**
6. **Deployment**

Time Line

1. Literature review:
Collection of primary and secondary data that will help to understand the big data analytics use case and technologies.
1 month
2. Design the architechtural diagram:
Based on the choosen big data technology Hadoop, hive and flume and project requirement design the archtechtural diagram.
1 month
3. Design detailed process diagram:

Based on the proposed architechtural diagram produce the detailed process diagram.
1 month

4. Data collection:

Collect the data using Twitter Application by appropriate keywords in windows environment.
1 month

5. Big Data environment setup:

Hadoop ecosystem required a dedicated setup on unix/linux environment. Need to install virtual machine and H adoop ecosystem including: Hive , Pig and Flume.
1 month

6. Execute the model:

Based on the methodology, architechtural and detailed design execute the model for the analytics.
1 month

7. Justify and conclude the proposed model:

Based on the analytics result, the suggestion will be given about the necessity of changing the traffic system in oman.
15 days

8. Visualize the result:

Based on the findings visualize the result using appropriate tool.
15 days

9: Document preparation:

The completed work will be documented in standard format.
1 month

Bibliography

1. Chandrasekhar Rangu, Shuvojit Chatterjee, Srinivasa Rao Valluru, “Text Mining Approach for Product Quality Enhancement” in IEEE 2017.
2. Mr. Peiman Barnaghi and John G. Breslin , “, Opinion Mining and sentiment polarity on Twitter and correlation between Events & Sentiment”, International Conference on Big Data Computing and Application , IEEE 2016.
3. Judith Sherin Tilsha S, Shobha M.S.,” A Survey on Twitter Data Analysis Techniques to Extract Public Opinion.”, IJARCSE , Vol. 5 , Issue 11 , Nov 2015 , 2277128X.
4. Lokmanyathilak Govindan Sankar Selvan,” A Framework for Fast-Feedback Opinion Mining on Twitter Data Streams”, IEEE 2015.
5. T. K. Das , D.P. Acharjya & M. R. Patra, “ Opinion Mining about a product by Analyzing Public Tweets in Twitter “, ICCCI- 2014, Jan 03-05, 2014.
6. Porter M.F, Snowball: A language for stemming algorithms. 2001.
7. Ning Zhong, Yuefeng Li, and Sheng-Tang Wu, “Effective Pattern Discovery for Text Mining”, IEEE Transactions on Knowledge And Data Engineering, Vol. 24, No.1, January 2012.
8. Andrei Sechelea, Tien Do Huu, Evangelos Zimos, and Nikos Deligiannis, “Twitter Data Clustering and Visualization”, in 2016 23rd International Conference on Telecommunications (ICT), 2016 IEEE.
9. Shruti Kohli, Himani Singal, “Data Analysis with R” in 2014 IEEE/ACM 7th International Conference on Utility and Cloud Computing.
10. Arun Jalanila, Nirmal Subramanian, “Comparing SAS® Text Miner, Python, R” in 2016 IEEE International Conference on Healthcare Informatics.
11. Dr. S. Vijayarani, Ms. J. Ilamathi, Ms. Nithya, “ Preprocessing Techniques for Text Mining - An Overview” in International Journal of Computer Science & Communication Networks,Vol 5(1),7-16.
12. Pearanalitycs, “Twitter study — august 2009,” 2009. [Online]. Available: http://pearanalytics.com/wp-content/uploads/2009/08/ Twitter-Study-August-2009.pdf
13. S. Kumar, F. Morstatter, and H. Liu, Twitter Data Analytics. Springer, Aug. 2013.
14. Shilpy Gandharv , Vivek Richhariya , Vineet Richhariya, Nov 2017 “Real Time Text Mining on Twitter Data” in International Journal of Computer Applications (0975 – 8887)

Budget Summary:

Budget Breakup - Year 1

Data collection	Downloading paid research journal , Weblog data charges, Book/Journal for literature review	200.00
SOFTWARE	Software purchse	150.00
STATIONERY	Book/Journal for literature review, To create the poster / model for the demonstration of the proposed model.	100.00
TRAINING	Training on Big Data tools & Technology.	300.00
TRAVEL AND CONSULTATION	Meeting with concern offices to get the feedback and questioners	100.00

Project duration (Days)

3

Current Status: Approved For Funding

Collaborative Partners:

Total number of Collaborative Partners:

Expected Outcomes:

	Omani	Non Omani
Post-Doc		
Technician (Bachelor Holder)		
Technician (Master Holder)		
Technician(PhD Holder)		
Postgraduate(Master)		
Postgraduate(PhD)		
Undergraduate		

No.Expected Publications:

No.Expected Patents:

Additional KPIs:

Submission ID: