International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

LOCATION PREDICTION ON TWITTER USING MACHINE LEARNING TECHNIQUES

A.MANASA[1], A.BHAVANI[2], A.PRANAVI[3], MR.VENKATA RAJESH[4]

[1,2,3] UG Students, Department of CSE, ***Malla Reddy Engineering College for Women (UGC-Autonomous) Maisammaguda, Hyderabad***, Telangana, India.

[4] Assistant Professor, Department of CSE, ***Malla Reddy Engineering College for Women (UGC-Autonomous) Maisammaguda, Hyderabad***, India.

ABSTRACT:

Location prediction of users from online social media brings considerable research these days. Automatic recognition of location related with or referenced in records has been investigated for decades. As a standout amongst the online social network organization, Twitter has pulled in an extensive number of users who send a millions of tweets on regular schedule. Because of the worldwide inclusion of its users and continuous tweets, location prediction on Twitter has increased noteworthy consideration in these days. Tweets, the short and noisy and rich natured texts bring many challenges in research area for researchers. In proposed framework, a general picture of location prediction using tweets is studied. In particular, tweet location is predicted from tweet contents. By outlining tweet content and contexts, it is fundamentally featured that how the issues rely upon these text inputs. In this work, we predict the location of user from the tweet text exploiting machine learning techniques namely naïve bayes, Support Vector Machine and Decision Tree.

Keywords: Location, twitter, ML, SVM.

1. INTRODUCTION:

Users may post explicitly their location on the tweet text they post, whereas in certain cases the location may be available implicitly by including certain relevant criteria. Tweets are not a strongly typed language, in which users may post casual with emotion images. Abbreviated form of text, misspellings, and extra characters of emotional words makes tweet texts noisy. The techniques applied for normal documents are not suited for analysing tweets. The character limitations of tweets about 140 characters may make the tweet uneasy to understand, if the tweet context is not studied.

The issue of location prediction related named as geolocation precition is examined for Wikipedia and web page documents. Entity recognition from these formal documents has been researched for years. Different types of content and context handling on these documents are also studied extensively. However, the location prediction

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal www.ijarst.in
ISSN: 2457-0362

problem from twitter depends highly on tweet content. Users living in specific regions, locations may examine neighborhood tourist spots, landmarks and buildings and related events.

Home Location: User's residential address given by user or location given by user on account creation is considered as home location. Home location prediction can be used in various application namely recommendation systems, location based advertisements, health monitoring, and polling etc. Home location can be specified as administrative location, geographical location or co-ordinates. Tweet Location: Tweet location refers to the region from where the tweet is posted by user. By construing tweet location, one can get tweet person's mobility. Usually home location collected from user profile, whereas tweet location can be arrived from user's geo tag. Because of the first perspectives on tweet location, POIs are comprehensively received as representation of tweet regions. Mentioned Location: When composing tweets, user may make reference to the names of a few locations in tweet texts. Referenced location prediction may encourage better understanding of tweet content, and advantage applications like recommendation systems, location based advertisements, health monitoring, and polling etc. In this study, we include two sub-modules of mentioned location: First one is recognizing the mentioned location in tweet text, which can be achieved by extracting text content from a tweet that refers to geography names. Second one is identifying the location from tweet text by solving them toentries in a geographical database.

LITERATURE SURVEY

1) Geolocation Prediction in Social Media Data by Finding Location Indicative Words

AUTHORS: Han, Bo & Cook, Paul & Baldwin, Timothy

Geolocation prediction is vital to geospatial applications like localised search and local event detection. Predominately, social media geolocation models are based on full text data, including common words with no geospatial dimension (e.g. today) and noisy strings (tmrw), potentially hampering prediction and leading to slower/more memory-intensive models. In this paper, we focus on finding location indicative words (LIWs) via feature selection, and establishing whether the reduced feature set boosts geolocation accuracy. Our results show that an information gain ratiobased approach surpasses other methods at LIW selection, outperforming state-of-the-art geolocation prediction methods by 10.6% in accuracy and reducing the mean and median of prediction error distance by 45km and 209km, respectively, on a public dataset. We further formulate notions of prediction confidence, and demonstrate that performance is even higher in cases where our model is more confident, striking a trade-off between accuracy and coverage. Finally, the identified LIWs reveal regional language differences, which could be potentially useful for lexicographers.

2) Where Are You Settling Down: Geo-locating Twitter Users Based on Tweets and Social Networks

AUTHORS: Shaowu ZhangHongfei Lin

In this paper, we investigate the advantages of taking two dimensions of tweet content and social relationships to construct models for predicting where people settle down as their profiles reveal

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal www.ijarst.in
ISSN: 2457-0362

city- and town-level data. Based on the users who voluntarily reveal their locations in their profiles, we propose two local word filters - Inverse Location Frequency (ILF) and Remote Words (RW) filter - to identify local words in tweets content. We also extract separately the place name mentioned in tweets using the Named Entity Recognition application and then filter them by computing the city distance. We consider users' friends and 2-hop of followings. In our experiment, we finally combine these two dimensions to estimate user location and achieve an Accuracy of 56.6% within 100 miles in city-level and 45.2% within 25 miles in town-level of their actual location which outperforms the single dimension prediction and the baseline. Twitter's open and succinct service allows it to gather vast amounts of data and updates by users who come from different places. The user always inadvertently leaks some dialect words and place names of his/her residence in the process of adding updates. Understanding the geographic features of those update statuses enables the system to push better local advertising, highlight points of interest, show local news, create recommendations for friends living in the vicinity, and even help search engines understand users' search intentions better. In this paper we build textual models of local words and place names based on pure tweets to estimate a user's place of residence, even when the user does not explicitly reveal the place name, or his/her geographic coordinates in the profile.

3) Multiple Location Profiling for Users and Relationships from Social Network and Content

AUTHORS : Li, Rui & Wang, Shengjie & Chen-Chuan Chang, Kevin

Users' locations are important for many applications such as personalized search and localized content delivery. In this paper, we study the problem of profiling Twitter users' locations with their following network and tweets. We propose a multiple location profiling model (MLP), which has three key features: 1) it formally models how likely a user ollows another user given their locations and how likely a user tweets a venue given his location, 2) it fundamentally captures that a user has multiple locations and his following relationships and tweeted venues can be related to any of his locations, and some of them are even noisy, and 3) it novelly utilizes the home locations of some users as partial supervision. As a result, MLP not only discovers users' locations accurately and completely, but also "explains" each following relationship by revealing users' true locations in the relationship. Experiments on a large-scale data set demonstrate those advantages. Particularly, 1) for predicting users' home locations, MLP successfully places 62% users and outperforms two state-of-the-art methods by 10% in accuracy, 2) for discovering users' multiple locations, MLP improves the baseline methods by 14% in recall, and 3) for explaining following relationships, MLP achieves 57% accuracy. In the literature, many methods [8, 5, 11] have been proposed to profile users' locations in the context of social network. Specifically, they focus on profiling a user's home location, which is the single "permanent" resident location of the user, by exploring her social network (e.g., friendships) and content (e.g., tweets). Intuitively, both types of data provide valuable signals for

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal www.ijarst.in
ISSN: 2457-0362

profiling users' locations, as a user is likely to 1) connect to others living close to her, and 2) tweet her nearby "venues".

EXISTING SYSTEM:

In the Existing system to the problem of finding location from social media content. The Social Networks from and motivated by Term frequency (TF) and inverse document frequency (IDF), they arrived Inverse City Frequency (ICF) and Inverse Location Frequency (ILF) respectively. They raked the features by using these frequency values and TF then by TF values. From this they arrived that local words spread in document in few places and have high ICF and ILF values. They approached model for identifying local words indicative or used in certain locations only. They aimed to identify automatically by ranking the local words by their location, and they find their degree of association of location words associated to particular location or cities.

DISADVANTAGES OF EXISTING SYSTEM:

- The issue of location prediction related named as geolocation prediction is examined for Wikipedia and web page documents.
- Entity recognition from these formal documents has been researched for years.
- The location prediction problem from twitter depends highly on tweet content.
- **Algorithm**: Term Frequency (TF) and Inverse Document Frequency (IDF)

PROPOSED SYSTEM:

Live stream of twitter data is collected as dataset using authentication keys. The aim of proposed system is to predict the user location from twitter content considering user home location, tweet location and tweet content. To handle this we used three machine learning approaches to make prediction easier and finding the best model amongst them. Live tweet stream from twitter for keyword "apple" is collected and stored in Tweettable. Live twitter data can be collected by registering a consumer_key, consumer_secret, access_token, access_token_secret for authentication and collecting live stream of tweets. We have collected more than 1000 tweets of particular keywords such as Indian city hashtag names. You can also search tweets based on hashtags.

ADVANTAGES OF PROPOSED SYSTEM:

- The information extracted from live includes tweetid, name, screen_name, tweet_text, HomeLocation, TweetLocation, MentionedLocation.
- Tweet text is compared with natural language tool kit package available in python to extract data from Cursor to Pandas Dataframe.
- Python programming, with few libraries used are scikit learn, numpy, pandas and geography.

Algorithm: Naive Bayes, Support Vector Machine, Decision Tree

3. METHODOLOGY

MODULES:

- **User**
- **Admin**
- **Data Preprocess**
- **Machine Learning**

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal www.ijarst.in
ISSN: 2457-0362

MODULES DESCRIPTION:

User:

The User can register the first. While registering he required a valid user email and mobile for further communications. Once the user register then admin can activate the customer. Once admin activated the customer then user can login into our system. User can search tweets based on hashtag. The first 100 tweets will get from twitter database and displayed to the user. At this time we are using geo code to identify the user location and tweet location. Most of the time user will not provide coordinates of his identity in the twitter account. So we are taking that as label class. This all tweets and geo code will stored in the database. Later we can apply the machine learning algorithms to test prediction result. The y_pred and y_test will displayed on the console. By help of sklearn.model_selection we can split the data into trainandtest. here we taken 80% of data for training and remiaing 20% for the testing.

Admin:

Admin can login with his credentials. Once he login he can activate the users. The activated user only login in our applications. The admin can set the training and testing data for the project dynamically to the code. After user operated the algorithms on provded dataset. The admin can view the results of naivebayes, svm and Decision tree results on his screens.

Data Preprocess:

Extra characters are removed from tweet text. Capitalize all words to find for geo location. Here we are using geography python library to get the exact latitude and longitude points of the users. Remove the tweet if user home location not mentioned. Mention home location in tweet location, if user tweet location is null

Removes tweets if no location is mentioned in tweet text. Final extract geodata from tweet text. Last step is to assign float value to the locations by its latitude and longitude values.

Machine Learning:

Naive Bayes Classification

Naive Bayes classifier is the most popular and simple classifier model used commonly. This model finds the posterior probability based on word distribution in the document. Naïve Bayes classifier work with Bag Of Words (BOW) feature extraction model, which do not consider the position of word inside the document. This model used Bayes Theorem for prediction of particular label from the given feature set. The dataset is split into trainset and test set. Upon test set, NB_model is applied to find the location prediction.

Support Vector Machine

Support vector machine is one of most common used supervised learning techniques, which is commonly used for both classification and regression problems. The algorithm works in such a way that each data is plotted as point in ndimensional space with the feature values represents the values of each co-ordinate.

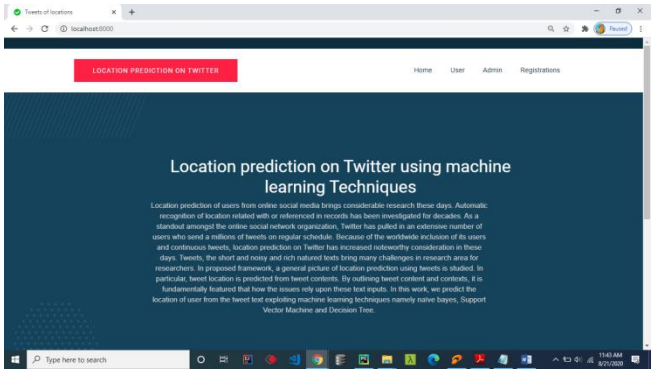
Decision Tree

Decision tree is the learning model, which utilizes classifications problem. Decision tree module works by splitting the dataset into minimum of

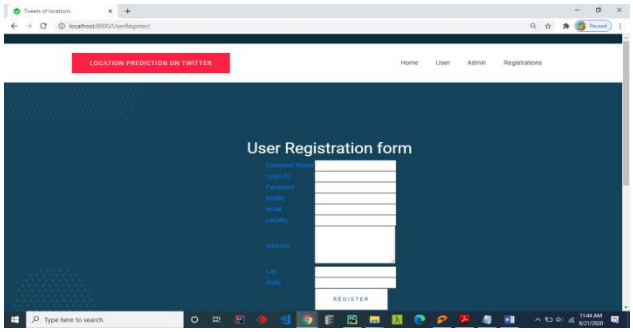
International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

two sets. Decision tree's internal nodes indicates a test on the features, branch depicts the result and leafs are decisions made after succeeding process on training.

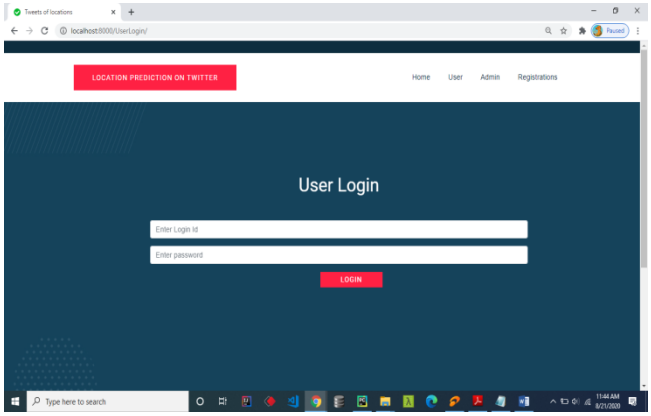
Home page:



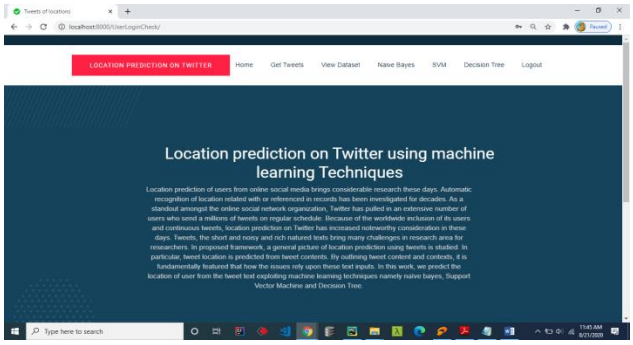
User Register Form:



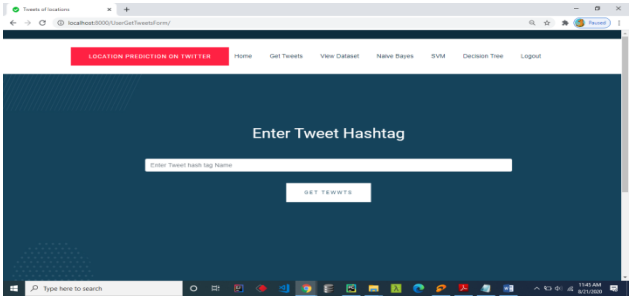
User login page:



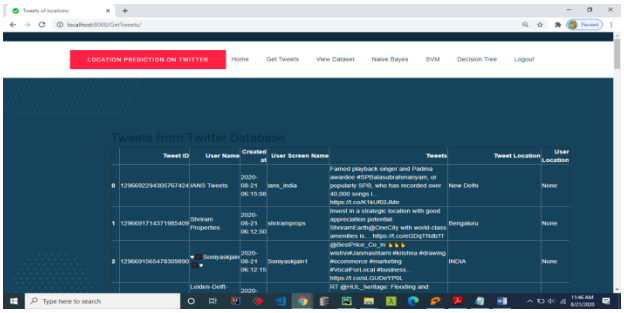
User Home page:



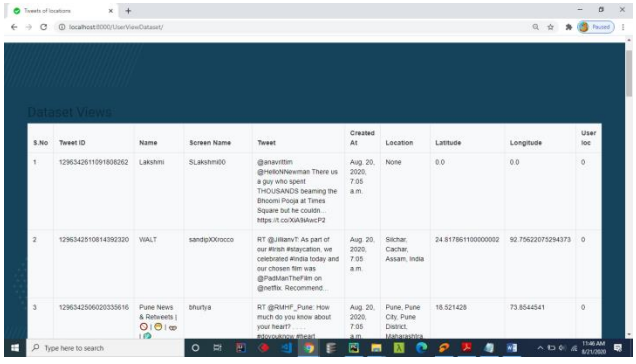
Tweet Search:



Tweet Results:

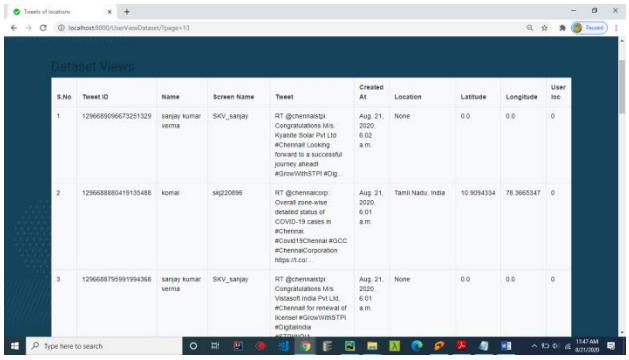


Dataset View:

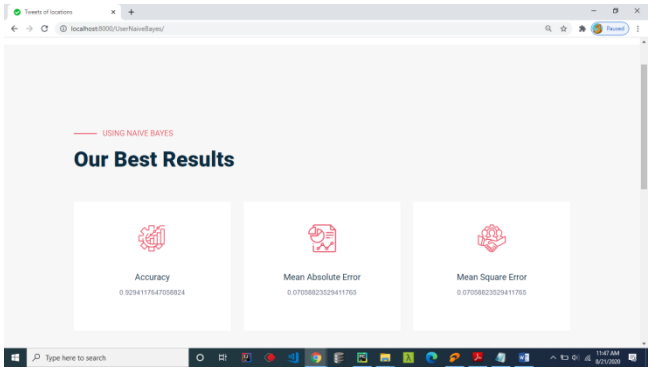


International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

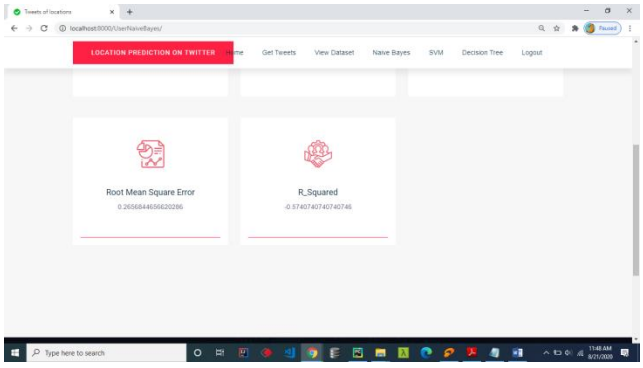
Dataset View:



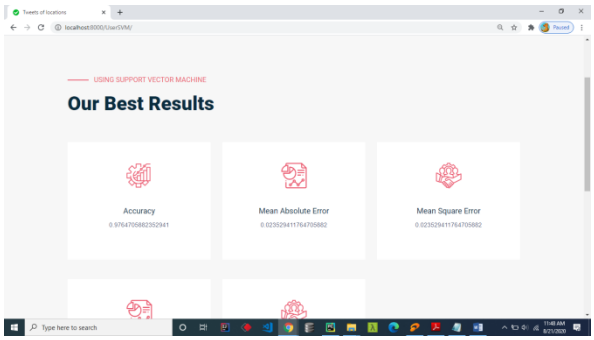
Naïve bayes Results:



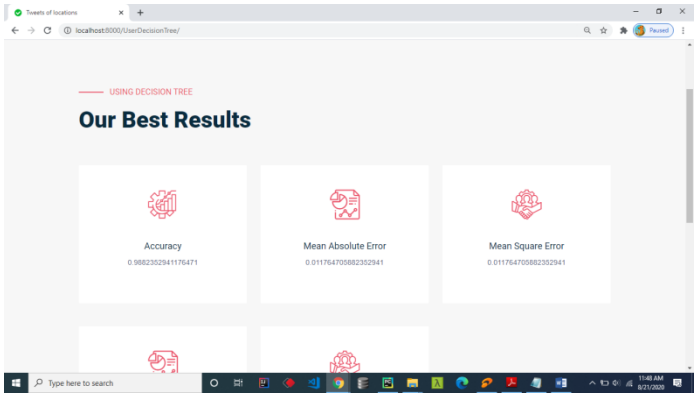
Naïve bayes Results:



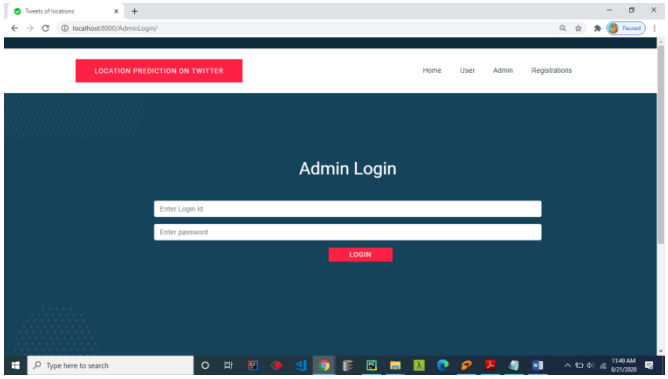
Support Vector Machine Results:



Decision Tree Results:

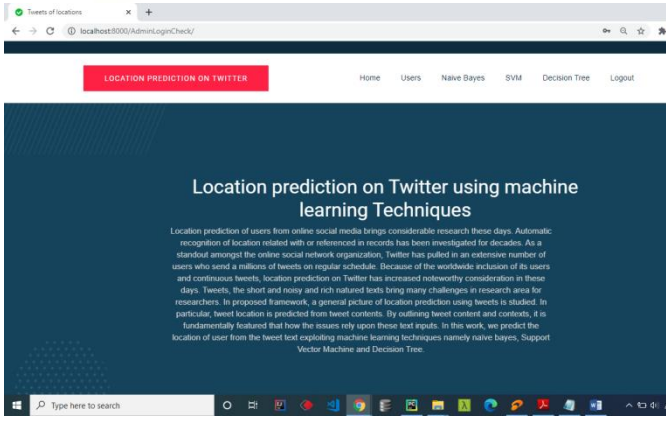


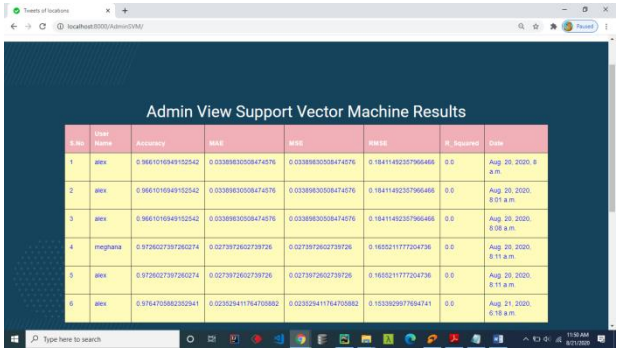
Admin Login:



Admin Home Page:

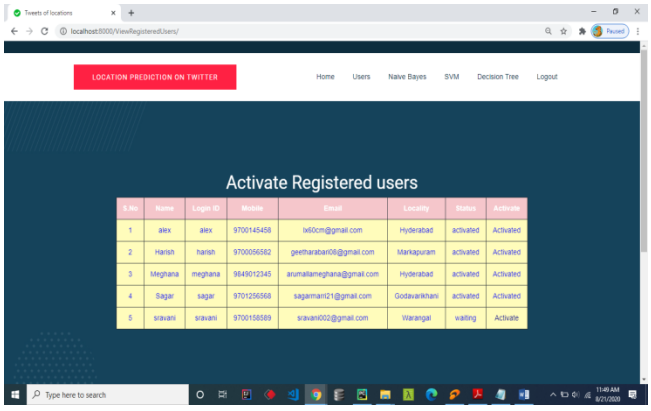
International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
IJARST
ISSN: 2457-0362

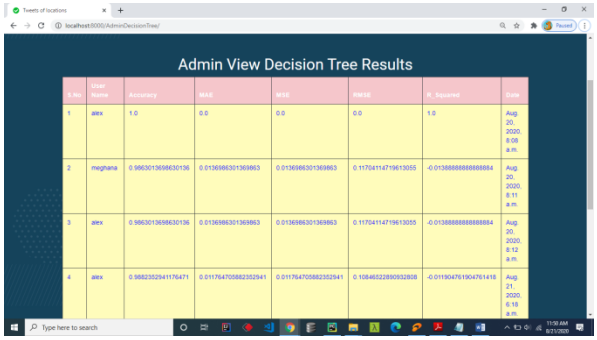




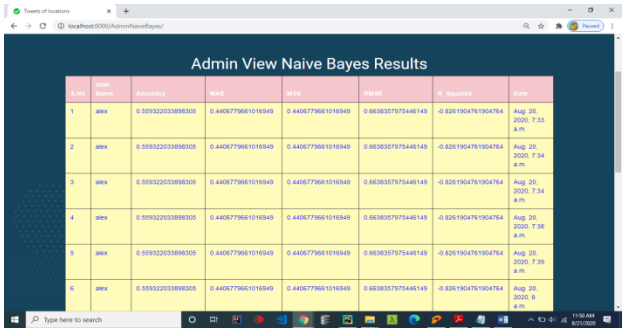
Admin View Decision Tree Results:

Activate Registerd users:





Admin View naïve bayes Results



Admin View SVM Results:

CONCLUSION

Three locations are considered from twitter data, namely home location, mentioned location and tweet location. When the twitter data is considered, geolocation prediction becomes a challenging problem. The tweet text nature and number of characters limitation make it hard to understand and analyze. In this work, we have predicted the geolocation of user from their tweet text using machine learning algorithms. We have implemented three algorithms to show the better performed one, which is suitable for geolocation prediction problem. Our experiment analysis concluded that decision tree is suitable for tweet text analysis and location prediction problem.

Further Enhancement

In future, we will try to infer the locations from other techniques such as users' friend networks, and other social networks such as Facebook,

Tumblr etc. Another limitation of the current system is that it works for flood related disaster, as the system is trained with flood related corpus. For any other disaster, the system has to be trained with that corpus. The current research opens up several new directions for other researchers to explore. The classification accuracy of the system is 81%, which can be further enhanced by considering more parameters. The inclusion of other languages will further enhance the system, as more users are expressing their views in their native languages. Human experts can study the misclassification cases to find the reasons for such misclassifications. This research can also be used to categorize users based on their movement patterns, which can be used by other businesses such as tourism.

REFERANCES

[1] Han, Bo & Cook, Paul & Baldwin, Timothy. (2012). Geolocation Prediction in Social Media Data by Finding Location Indicative Words. 24th International Conference on Computational Linguistics - Proceedings of COLING 2012: Technical Papers. 1045-1062.

[2] Ren K., Zhang S., Lin H. (2012) Where Are You Settling Down: Geo-locating Twitter Users Based on Tweets and Social Networks. In: Hou Y., Nie JY., Sun L., Wang B., Zhang P. (eds) Information Retrieval Technology. AIRS 2012. Lecture Notes in Computer Science, vol 7675. Springer, Berlin, Heidelberg.

[3] Han, Bo & Cook, Paul & Baldwin, Timothy. (2014). Text-Based Twitter User Geolocation Prediction. The Journal of Artificial Intelligence Research (JAIR). 49. 10.1613/jair.4200.

[4] Li, Rui & Wang, Shengjie & Chen-Chuan Chang, Kevin. (2012). Multiple Location Profiling for Users and Relationships from Social Network and Content. Proceedings of the VLDB Endowment. 5. 10.14778/2350229.2350273.

[5] Jalal Mahmud, Jeffrey Nichols, and Clemens Drews. 2014. Home Location Identification of Twitter Users. ACM Trans. Intell. Syst. Technol. 5, 3, Article 47 (July 2014), 21 pages. DOI: http://dx.doi.org/10.1145/2528548

[6] Miura, Yasuhide, Motoki Taniguchi, Tomoki Taniguchi and Tomoko Ohkuma. "A Simple Scalable Neural Networks based Model for Geolocation Prediction in Twitter." NUT@COLING (2016).

[7] A. Schulz, A. Hadjakos, H. Paulheim, J. Nachtwey, and M. M¨ uhlh¨auser, "A multi-indicator approach for geolocalization of tweets," in Proc. 7th Int. Conf. on Weblogs and Social Media, 2013.

[8] R. Li, S. Wang, H. Deng, R. Wang, and K. C.-C. Chang, "Towards social user profiling: unified and discriminative influence model for inferring home locations," in Proc. 18th ACM Int. Conf. on Knowledge Discovery and Data Mining, 2012, pp. 1023–1031.

[9] B. Han, P. Cook, and T. Baldwin, "A stacking-based approach to twitter user geolocation prediction," in Proc. 51st Annual Meeting of the Association for Computational Linguistics System Demonstrations, 2013, pp. 7–12.

[10] D. Flatow, M. Naaman, K. E. Xie, Y. Volkovich, and Y. Kanza, "On the accuracy of hyper-local geotagging of social media content,"

in Proc. 8th ACM Int. Conf. on Web Search and Data Mining, 2015, pp. 127–136.