Technoarete Transactions on Advances in Computer Applications (TTACA)
Vol-3, Issue-1, March 2024
e-ISSN: 2583-3472

Using of Fuzzy Logic Matching Algorithm for Hindi Dataset

Madhuri Sharma [1], Medhavi Malik [2*]

[1] Department of Computer Science and Engineering, SRM Institute of Science and Technology, Delhi-NCR Campus, Ghaziabad, Uttar Pradesh, India
[2] Department of Computer Science and Engineering, Maharaja Surajmal Institute of Technology, Janakpuri, New Delhi, India
*Corresponding Author Email: medhavimalik28@gmail.com

Abstract

Various new computing methods based on Fuzzy Logic can be used in various systems. Numerous researchers and developers are working with Fuzzy Logic and its Applications. One such application is applied on the dataset which contains the values in Hindi Language; the major problem is that data come from different sources and when do the merge operation it will create a lot of problem; since same names are represented by different ways of writing it. Here, we apply a technique of fuzzy matching algorithm for hindi dataset. We are applying for the project of sharing bike concept which comes from different data sources. Fuzzy Logic is a very promising mathematical approach for training the dataset which being characterized by subjectivity & imprecision. More emphasis is put on the fuzzy string matching and its criteria for the performing the merge operation. This paper presents an analysis of the results achieved using fuzzy logic to model the dataset. Assessing the performance of datasets after performing matching algorithm at different levels using traditional methods.

Keywords

Devanagari Script, Fuzzy Logic, Matching Algorithm, NLP.

INTRODUCTION

Fuzzy systems are an alternative to the traditional methods of dealing with affiliation and logic, which have their origins in ancient Greek philosophy and applications in the field of artificial intelligence. [1] We humans are often unsure about our decisions i.e. by our very nature we are fuzzy. Trip generation defines that how many people want to go out of their homes and what their purposes are. [2] During the past few years, there is a rapid growth in the number of applications in the area of Fuzzy Logic. Fuzzy system has become one of the most successful technologies for reasoning inference systems. The main reason behind this is closely resembles Human Decision making; which is having the ability to solve the problem and generate the solutions which are precise and definite. These models are transparent and easily accommodate to the new findings. Complexity is directly proportional to imprecision or inexactness. For complex systems, cost is proportional to precision: more precision means higher cost. For a given problem, if fuzzy logic systems are applied, then consider about tolerance for imprecision. Optimality is defined as the percent accuracy: 0% means exact answer and accuracies larger than zero representing answers of lesser accuracy. Fuzzy systems are useful in situations when fast solution is guaranteed but approximated solutions are defined. It also involves highly complex systems whose behaviors are not well understood. Many attempts were made to find algorithms to match strings and, in the process, developing new techniques and methodologies. [3]

There are different areas where fuzzy logic can be applied; but having a lot of challenging problems are facing. Data came from different sources, huge data sets; different data sets & uncertainty of datasets is there. Because of this uncertainty & the use of random variables, there is a lead to inaccuracy of data; estimations of approximations, some data are incomparable, incompleteness of expert knowledge. As a result, they conduct in-depth research into client preferences, which are described as "must-haves" in job offerings. [4]

Traditional Boolean Logic

In traditional Boolean logic, the inputted text is given and processed according to the requirement, produced an output in binary form (either 0 or 1). The accessibility of standard benchmarks has invigorated research in Natural Language Processing (NLP). [5]

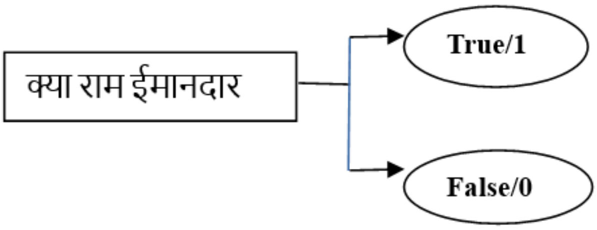


Figure 1. Traditional Boolean Logic

Fuzzy System

While in Fuzzy system, the input text which is given, and processed and output can be in between 0 & 1. There may be not sure about the output; so fuzzy approach is better used.

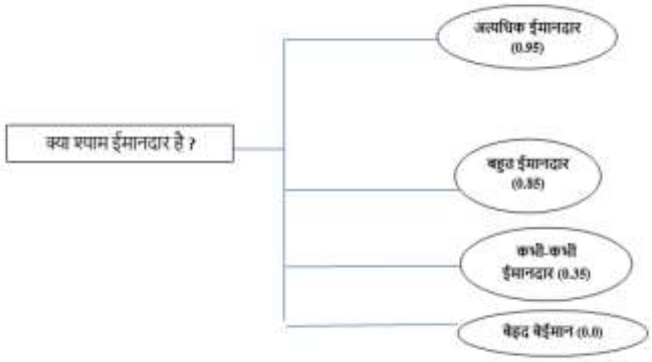


Figure 2. Fuzzy System

Data Ambiguities

Since for Hindi language dataset, there can be a number of ambiguities regarding either the datasets or in language. Many business applications have ambiguities in datasets. Suppose a column from a particular application dataset. Here, one of the values is represented as “चैन”; there is an ambiguity in the word which represents “आराम” in column from one source; “माला” from another source. While integrating, the above record will not be tagged to the word “चैन”. It may also indicate duplications due to incorrect naming conventions or multiple naming conventions. Various impacts are arising: Duplicated Records, Lack of Integrity, Inaccurate Insights, and Untagged Records. To such problems, the Hence, the best solution would be to consolidate similar data automatically through advanced text analytics algorithms. The practical solution for integration is to address the above issues such as Fuzzy Logic in NLP.

Fuzzy Logic used in the domain of Natural Language Processing that that aids in identifying and make similar records. Those misinterpreted or misspelled records have to associate with similar records. The main advantage of Fuzzy Logic is having one single source of truth, eliminates duplicates, identifying similar sounding business names.

Example: बात मत करो |
मत के आधार पर फैसला हुआ |

In first sentence, मत means “stop talking” while in second “The decision was based on votes”.

Data Integration

It is a preprocessing technique that combines data from multiple heterogeneous data sources into a coherent data store and provides a unified view of data. This approach is formally defined as triplets of **<U, H, Q>**; where

U stands for Universal Schema
H stands for Heterogeneous Source of Schema
Q stands for mapping between queries of source and Universal Schema

STANDARDIZING NAMES USING FUZZY LOGIC

If we are trying to integrate data from different sources in multiple formats; we can integrate into one record using the Fuzzy Logic Algorithm.

Table 1. The word ‘GATE’ in hindi from the different sources. All these words sense to a single word.

Data Source 1	Data Source 2	Data Source 3	Data Source 4
दरवाजा	द्वार	**प्रवेशमार्ग**	**कपाट**

As a result, the algorithm will group all versions and assign them to a standardized notation of **दरवाजा.**

Fuzzy Step

The following equation is used to find the most probable distance:

$$\mathbf{argmax}_{x \in \text{datasets}}\, \mathbf{P(x)\ P(w|x)}$$

Where, P(x) indicate how often the word ‘x’ is used in the dataset. It can be computed by counting the number of times a word appears in whole datasets. P(w | x) is the probability that the author typed in the wrong word ‘w’instead of correct word ‘x’.

PROBLEM DESCRIPTION

Sharing of bike is a project which gives the experience on two wheels especially on tourist places. The main aim of this designing is for quick trips and has convenience in mind. This concept is especially for fun, saves time and affordable way to get around. Tourists simply join this; ride the bike and return it. You can join this simply by becoming a member online or do a registration at any station. Take as many short trips as you want while your pass or membership is active. Return your bike at any station. You can simply commute from one place to another.

In the dataset, it contains **duration (in ms), start_date, start_station, end_date, end_station, bike_number, subscription_type**. But the main problem is that datasets are above from different stations. When we apply merge or join operations here, in order to get the datasets in one frame for simplification of cost(or pricing) over it; Major problem behind is of integrating the data from different stations since they are containing the ambiguity in their names of stations.

For this various string matching algorithms are includes. Recognizable proof, elucidation and reaction to client prerequisites are the key achievement factors for organizations, paying little heed to their industry. [6]

HOW DOES FUZZY NAME MATCHING WORK?

The most important use cases of fuzzy matching come about when we want to join tables using the name field. The company stores the responses at different stations. To merge the collected information, the company would want to join tables using the start_station attribute as the primary key. The problem arises there is not always a guarantee that a person will fill out exactly the same name defined in another table. When a string-matching algorithm is applied, there is not always a guarantee that result is more accurate since there is a slight variation in writing the names; thereby the lower in prediction accuracy. So, a Fuzzy Matching String Algorithms can be applied; which is itself a set of rules. The problem of identifying the Proper Noun and Common Noun is very difficult. [7]

Fuzzy Matching

Fuzzy algorithm identifies the pairs of words for every combination of account from multiple sources. Technique [8] can be used to search or match strings in special cases when some pairs of symbols are more similar to each other than the others. The algorithm looks for similar sounding words based on various parameters and identifies those accounts from different sources to cluster them together and form a single name.

Fuzzy Matching algorithm will identify Name reversal, Misspellings, Name variations, Phonetic Spellings, Different Spellings of names, Shortened names, Abbreviations, Insertion or Removal of Punctuation, Special Characters, Spaces. Matching strings is a problem that occupied many researchers over the past decades. [3]

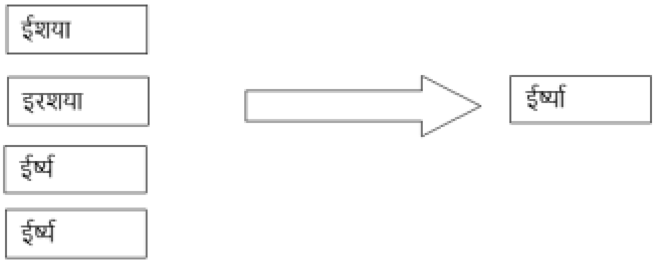


Figure 3. Misspelled words all mapped to one word

Levenshtein Distance (LD)

LD is used to calculate the minimum number of operations/ edits required to change a particular string into some other string. Operations can be Addition, Substitution & Deletion. For any operation it costs to +1.

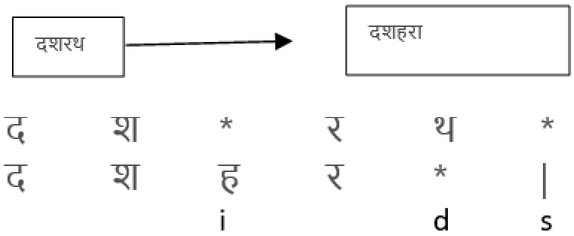


Figure 4. Minimum edit distance between two strings as an alignment. The operation list for converting the top row to bottom row: i for insertion, d for deletion, s for substitution.

Token_Sort_Ratio

When the percentage lies is 100 percentage then it is exact match. When is lies between 75-100 then it is approximate match. Moderate match is 60-75 percentages. Slight match lies between 40-60 percentages.

Ratio (Simple Ratio)

This ratio is used when words are same but the order of the words matters while calculating the ratio.

Sentence 1 = "मेरा नाम राम है"
Sentence 2 = "राम मेरा नाम है"

Above when executing the code, indicates about the 50 % of the word similarity.

Partial Ratio

It first takes the shortest length string and then compares it with all the substrings of the same length. It also helps to perform the substring matching. This dataset is used when we are trying the recognize the same person with two names in different databases especially when first name, last name, middle name is there.

Sentence 1 = "मेरा नाम राम है"
Sentence 2 = "मेरा नाम राम शर्मा है"

Above code when executing, gives 100 % of the word similarity.

Token Sort Ratio

First the input strings are converted into tokens. For pre-processing steps, remove the punctuation marks and stop words; then strings are arranged in lexicographical order. Using the distance similarity ratio is calculated between the strings.

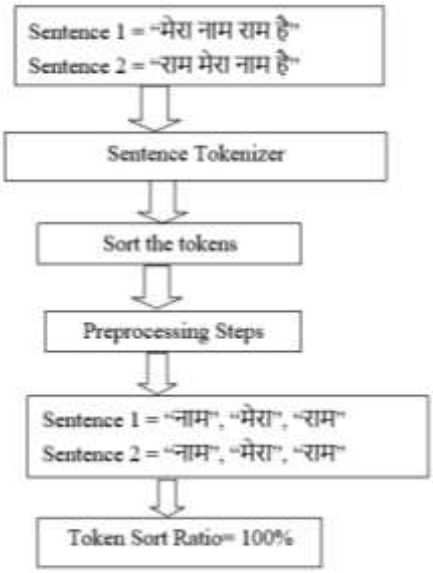

Sentence 1 = "मेरा नाम राम है"
Sentence 2 = "राम मेरा नाम है"
Sentence Tokenizer
Sort the tokens
Preprocessing Steps
Sentence 1 = "नाम", "मेरा", "राम"
Sentence 2 = "नाम", "मेरा", "राम"
Token Sort Ratio= 100%

Figure 5. Token Sort Ratio

Token Sort Ratio used when the order doesn't matter then this is the best way to match the similarity.

Token Set Ratio

In token set ratio, strings are first converted into tokens. Remove the punctuation marks and stop words for pre-processing steps. Instead of doing the sorting alphabetically took out the common words between them. This ratio used when you do not care about the number of times a word in the string is repeated.

Sentence 1 = "मेरा नाम राम है"
Sentence 2 = "राम मेरा नाम नाम है"

Token set ratio comes out to100 % since repetition does not matter.

APPROACH

Pseduo code

```
1. Initialize network weights (often small random values)
2. do
3. for each training example eg
4. do the string matching algorithm
5. perform Levenshtein Distance operation
```

$$\operatorname{lev}_{a,b}(i,j)=\begin{cases}\max(i,j) & \text{if } \min(i,j)=0,\\ \min\begin{cases}\operatorname{lev}_{a,b}(i-1,j)+1\\ \operatorname{lev}_{a,b}(i,j-1)+1\\ \operatorname{lev}_{a,b}(i-1,j-1)+1_{(a_i\neq b_j)}\end{cases} & \text{otherwise.}\end{cases}$$

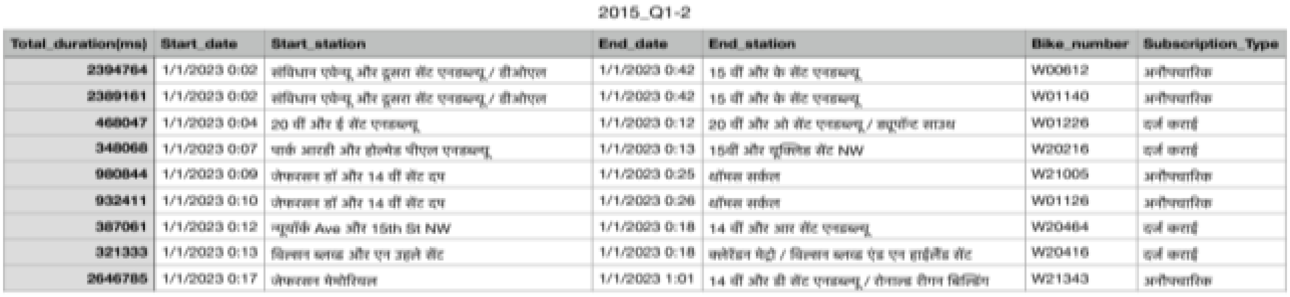
```
6. Consider from dataset1
7. if (lev_{a, b} (dataset1) < (lev_{a, b} (dataset2 ) then
      consider from dataset2
8. merge (dataset1, dataset2)
9. until all training examples are properly merged or until
      criteria is fulfilled
10. return the merged dataset
```

DATASET

At first, we need to preprocess the data from the dataset to clean the unwanted observation. There might have duplicate data because of the maximum amount of data is in hard copy; it may be obvious that one data may be inserted twice. There are also some irrelevant observations which actually don't fit with the specific problem that we are trying to solve.

For the dataset1 & dataset2, start_station, end_station & subscription_type are in hindi formats.

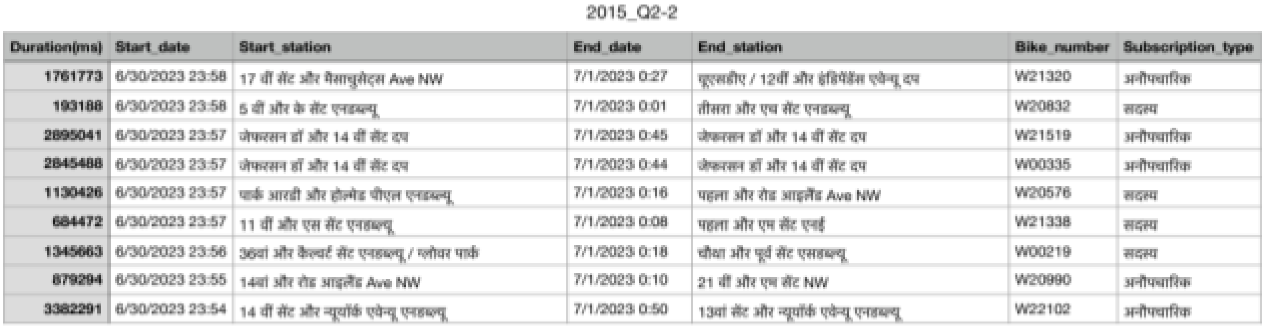
Table 2. Dataset sample of few features (From one Data Source1)



2015_Q1-2

Total_duration(ms)	Start_date	Start_station	End_date	End_station	Bike_number	Subscription_Type
2394764	1/1/2023 0:02	संविधान एवेन्यू और दूसरा सेंट एनडब्ल्यू / डीओएल	1/1/2023 0:42	15 वीं और के सेंट एनडब्ल्यू	W00612	अनौपचारिक
2389161	1/1/2023 0:02	संविधान एवेन्यू और दूसरा सेंट एनडब्ल्यू / डीओएल	1/1/2023 0:42	15 वीं और के सेंट एनडब्ल्यू	W01140	अनौपचारिक
468047	1/1/2023 0:04	20 वीं और ई सेंट एनडब्ल्यू	1/1/2023 0:12	20 वीं और ओ सेंट एनडब्ल्यू / ड्यूपॉन्ट साउथ	W01226	दर्ज कराई
348068	1/1/2023 0:07	पार्क आरडी और होल्मेड पीएल एनडब्ल्यू	1/1/2023 0:13	15वीं और यूक्लिड सेंट NW	W20216	दर्ज कराई
980844	1/1/2023 0:09	जेफरसन डॉ और 14 वीं सेंट दप	1/1/2023 0:25	ओहियो सर्कल	W21005	अनौपचारिक
932411	1/1/2023 0:10	जेफरसन डॉ और 14 वीं सेंट दप	1/1/2023 0:26	ओहियो सर्कल	W01126	अनौपचारिक
387061	1/1/2023 0:12	न्यूयॉर्क Ave और 15th St NW	1/1/2023 0:18	14 वीं और आर सेंट एनडब्ल्यू	W20464	दर्ज कराई
321333	1/1/2023 0:13	विल्सन ब्लव्ड और एन उहले सेंट	1/1/2023 0:18	क्लेरेंडन मेट्रो / विल्सन ब्लव्ड एंड एन हाईलैंड सेंट	W20416	दर्ज कराई
2646785	1/1/2023 0:17	जेफरसन मेमोरियल	1/1/2023 1:01	14 वीं और डी सेंट एनडब्ल्यू / रोनाल्ड रीगन बिल्डिंग	W21343	अनौपचारिक

Table 3. Dataset sample of few features (From one Data Source2)



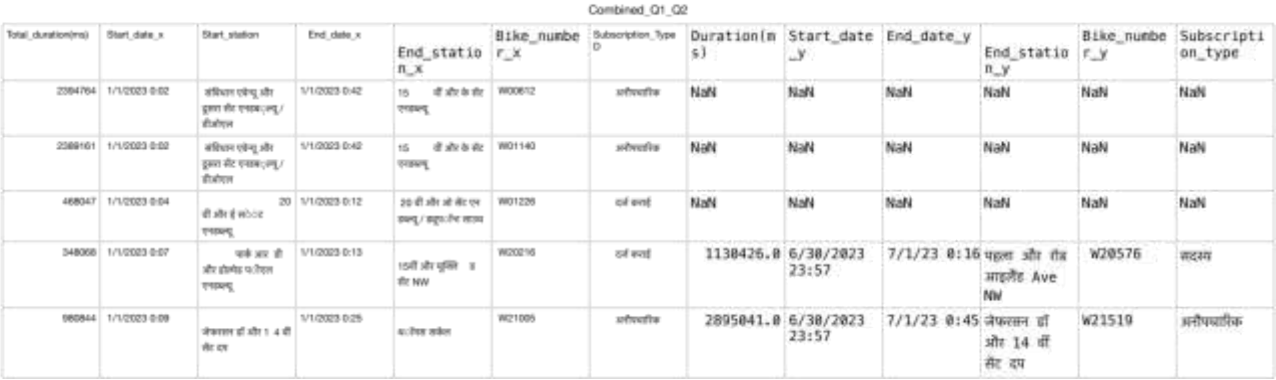
2015_Q2-2

Duration(ms)	Start_date	Start_station	End_date	End_station	Bike_number	Subscription_type
1761773	6/30/2023 23:58	17 वीं सेंट और मैसाचुसेट्स Ave NW	7/1/2023 0:27	यूएसडीए / 12वीं और इंडिपेंडेंस एवेन्यू दप	W21320	अनौपचारिक
193188	6/30/2023 23:58	5 वीं और के सेंट एनडब्ल्यू	7/1/2023 0:01	तीसरा और एच सेंट एनडब्ल्यू	W20832	सदस्य
2895041	6/30/2023 23:57	जेफरसन डॉ और 14 वीं सेंट दप	7/1/2023 0:45	जेफरसन डॉ और 14 वीं सेंट दप	W21519	अनौपचारिक
2845488	6/30/2023 23:57	जेफरसन डॉ और 14 वीं सेंट दप	7/1/2023 0:44	जेफरसन डॉ और 14 वीं सेंट दप	W00335	अनौपचारिक
1130426	6/30/2023 23:57	पार्क आरडी और होल्मेड पीएल एनडब्ल्यू	7/1/2023 0:16	पहला और रोड आइलैंड Ave NW	W20576	सदस्य
684472	6/30/2023 23:57	11 वीं और एस सेंट एनडब्ल्यू	7/1/2023 0:08	पहला और एम सेंट एनई	W21338	सदस्य
1345663	6/30/2023 23:56	36वां और कैल्वर्ट सेंट एनडब्ल्यू / ग्लोवर पार्क	7/1/2023 0:18	चौथा और पूर्व सेंट एसडब्ल्यू	W00219	सदस्य
879294	6/30/2023 23:55	14वां और रोड आइलैंड Ave NW	7/1/2023 0:10	21 वीं और एम सेंट NW	W20990	अनौपचारिक
3382291	6/30/2023 23:54	14 वीं सेंट और न्यूयॉर्क एवेन्यू एनडब्ल्यू	7/1/2023 0:50	13वां सेंट और न्यूयॉर्क एवेन्यू एनडब्ल्यू	W22102	अनौपचारिक

RESULT ANALYSIS

Similar dataset from different sources is merged. Merged datasets are shown below in figure.

Table 4. Combined from Data Source1 & Data Source2



Combined_Q1_Q2

Total_duration(ms)	Start_date_x	Start_station	End_date_x	End_station_x	Bike_number_x	Subscription_Type D	Duration(ms)	Start_date_y	End_date_y	End_station_y	Bike_number_y	Subscription_type
2394764	1/1/2023 0:02	संविधान एवेन्यू और दूसरा सेंट एनडब्ल्यू / डीओएल	1/1/2023 0:42	15 वीं और के सेंट एनडब्ल्यू	W00612	अनौपचारिक	NaN	NaN	NaN	NaN	NaN	NaN
2389161	1/1/2023 0:02	संविधान एवेन्यू और दूसरा सेंट एनडब्ल्यू / डीओएल	1/1/2023 0:42	15 वीं और के सेंट एनडब्ल्यू	W01140	अनौपचारिक	NaN	NaN	NaN	NaN	NaN	NaN
468047	1/1/2023 0:04	20 वीं और ई सेंट एनडब्ल्यू	1/1/2023 0:12	20 वीं और ओ सेंट एनडब्ल्यू / ड्यूपॉन्ट साउथ	W01226	दर्ज कराई	NaN	NaN	NaN	NaN	NaN	NaN
348068	1/1/2023 0:07	पार्क आर डी और होल्मेड प्लेस एनडब्ल्यू	1/1/2023 0:13	15वीं और यूक्लिड सेंट NW	W20216	दर्ज कराई	1130426.0	6/30/2023 23:57	7/1/23 0:16	पहला और रोड आइलैंड Ave NW	W20576	सदस्य
980844	1/1/2023 0:09	जेफरसन डॉ और 1 4 वीं सेंट दप	1/1/2023 0:25	ओहियो सर्कल	W21005	अनौपचारिक	2895041.0	6/30/2023 23:57	7/1/23 0:45	जेफरसन डॉ और 14 वीं सेंट दप	W21519	अनौपचारिक

After then, metrics is evaluated by Jaccard Coefficient that measures the similarity between finite sample sets. The Jaccard coefficient can be a value between 0 and 1, with 0 indicating no overlap and 1 complete overlap between the sets. [8]

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}.$$

Sharing of Bike datasets from different sources; when combined produced an accuracy of 89%. Since, datasets are written in native language (hindi language); there may be a problem for a different writing language. This approach somewhat better produced result rather than other approach.

CONCLUSION

With the help of Fuzzy Logic Algorithms, it can solve the real-world problems for normalizing of datasets. This saves a considerable amount of time; when you are integrating the data. This paper described a new method that uses of naïve based technologies to understand string matching algorithms. First, the problem was defined very accurately, and then the algorithm for solving the problem of combing the strings was presented. Finally, the result can be shown how the words can be matched and organized. In future, we will work about NULL values on the datasets.

REFERENCES

[1] Popescu, V. F. & Pistol, M. S. (2021). Fuzzy logic expert system for evaluating the activity of university teachers. International Journal of Assessment Tools in Educations, 8 (4), 991-1008. DOI: 10.21449/ijate.1025690

[2] Sarkar, Amrita. (2012). Application of Fuzzy Logic in Transport Planning. International Journal on Soft Computing. 3. 1-21. 10.5121/ijsc.2012.3201.

[3] Schneider, Moti & Bunke, Horst & Kandel, Abraham. (2001). Using fuzzy logic to match strings in documents. Int. J. Intell. Syst.. 16. 609-619. 10.1002/int.1026.

[4] A. Shankar and M. Malik, "Predicting Person's Pay using Machine Learning," 2022 4th International Conference on Advances in Computing, Communication Control and Networking (ICAC3N), Greater Noida, India, 2022, pp. 205-208, doi: 10.1109/ICAC3N56670.2022.10074407.

[5] Sharma, M., Malik, M. (2021). Use of RNN in Devangari Script. In: Dash, S.S., Das, S., Panigrahi, B.K. (eds) Intelligent Computing and Applications. Advances in Intelligent Systems and Computing, vol 1172. Springer, Singapore. https://doi.org/10.1007/978-981-15-5566-4_53

[6] Kushwah, Virendra & Bajpai, Aruna. (2019). Importance of Fuzzy Logic and Application Areas in Engineering Research.

[7] M. Sharma, M. Malik, N. Gupta, An efficient & learning approach of POS tagging using rule-based for Devanagari script. Int. J. Innov. Technol. Exploring Eng. (IJITEE) 8(7C2). ISSN 2278-3075 (2019)

[8] Snášel, V., Keprt, A., Abraham, A., Hassanien, A.E. (2009). Approximate String Matching by Fuzzy Automata. In: Cyran, K.A., Kozielski, S., Peters, J.F., Stańczyk, U., Wakulicz-Deja, A. (eds) Man-Machine Interactions. Advances in Intelligent and Soft Computing, vol 59. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-00563-3_29

[9] Kosub S (April 2019). "A note on the triangle inequality for the Jaccard distance". Pattern Recognition Letters. 120: 36–8. arXiv:1612.02696. Bibcode:2019PaReL.120...36K. doi:10.1016/j.patrec.2018.12.007. S2CID 564831