International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 04 Issue: 05 | May -2020 ISSN: 2582-3930

Survey on Text Alikenessinnatural Language Processing

Sathish R M.E.,(PhD).[1], Koushika.J[2],Nanthini.G[3],DhivyaBharathi.P[4]

[1,2,,3,4]Department of Information Technology ,KGiSL Institute Of Technology

***Abstract*—** Natural language processing (NLP) is a computer science branch concerned with understanding human language and communication, and translating these into an embedding that is comprehensible to the computer.Their purpose in this paper is to capturemeaningfromthenaturalhumanlanguagethroughNLPandto create similarities between texts .Now a days peoples are connectedwith large amount of data in daily basis .The use of large data set increased problem in social networks of users ,individual person or an organization. In this study we test a new technique based on non- machine learning language using the cosine similarity algorithm The similarity measurement method can be used in text mining to deter mine the correct clustering algorithm for a specific problem. execute and test , so as to figure out the original user and fake user in stock exchange which acquires the best outcomes ininvestments.

I. INTRODUCTION

The internet has led to the increase of online electronic documents, further compelling textual categorization and document classification in various online repositories.Text mining, machine learning, and natural language processing techniques and methodologies have been used to process big data that is constantly overwhelming the internet user at present. Seeking similarity between words is a fundamental part of similarity of text which is then used as a primary stage for similarities be tween sentence, paragraph and paper.

Words are similar lexically if they have a similar character sequence. Words are similar semantically if they have the same thing, are opposite of each other, used in the same way. Term similarity is a concept used to check whether two documents are similar through measuring similarity in terms of theirterm.

Other possible measures could include: document length, number of common terms, usual or unusual term s, and number of times that a termappears.

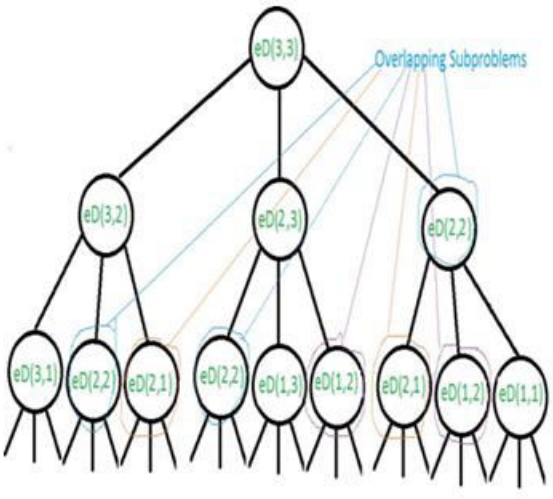
Text similarity measures have been used for a long time in applications in natural language processing and related areas. One of the early applications of text similarity is, perhaps, the vector model in the information retrieval, where the document most important to the input query is calculated by ranking documents in the databasein reverse order of their similarity to the query inquestion. Text similarity was also used for relevance input and text classification (Rocchio 1971), word sense uncertainty (Lesk 1986; Schutze 1998) and, more recently, for extractive summarization (Salton et al. 1997) and methods for automated evaluation of machine translation (Papinenietal.2002)ortextsummarization(Lin&Hovy2003).

Measures of text similarity have also been found to be useful. Measuresoftextsimilaritywerealsofoundto beusefulforassessing text coherence (Lapata&Barzilay 2005). With few exceptions, the typical approach to identifying similarities between two text segments is to use a simple lexical matching method and to generate a similarity score based on the number of lexical units that exist in both input segments. Improvements to this simple method called halting, stopping-word elimination, part-of-speech marking, the longest matching sequence. In this paper, we propose a method for measuring the semantic similarity of texts by making use of informationthatcanbeobtainedfromthesimilarityofthewordpart.

II. *BACKGROUND KNOWLEDGE*

A. Edit DistanceBased

Algorithms falling within this category attempt to measure the numberofoperationsneedtoconvertonestringintoanother.More the number of operations, the resemblance between the two stringsis less. One aspect to notice, each index character of a string is given equal importance in their cases.



Worst case recursion tree when m = 3, n = 3.
Worst case example str1="abc" str2="xyz"

Fig 1: Edit Based Algorithm

The similarity here is a growing sub-string element between the two strings.Thealgorithms, try to find the longest sequence in both sets, thehigherthesimilarityscoreis,themorethosesequencesarefound. Remember,herethecombinationofsame-lengthcharactersisequally important.

Algorithm: Similarity Search (SS)

Input: A sequence 'A', which is the query sequence and a sequence 'B' obtained from Pre-Search

Output: Similarity score for short-listed sequences

Description: Sequence A slides over sequence B, one character each pass. In each pass, characters are matched from left to right and a score Sn is assigned for each pass. This process is repeated for each B obtained in Pre-Search

Algorithm:
Score Sn=0; k=1;

1. Compare matching pairs of A and B from left to right.
   a. Add k to Sn for each mathc.
   b. Double k if previous pair is matched or reset it to 1, otherwise
   c. Limit the value of k to 10,000
   d. Normalize Sn for the overlapping length of A and B
2. Repeat step 1 till the last character of A is matched with last character of B.
3. Get the maximum value of Sn as the final similarity score.
4. Based on a suitable threshold, such as 150, either accept the sequence as similar, or reject otherwise.

B.Token Based

The expected input in this category is a set of tokens, rather than complete strings. In both sets the idea is to find similar tokens. More thenumberofcommontokens,themorethesetsaresimilar.Through dividing using a delimiter a string can be divided into sets. This way we can turn a sentence into word tokens or characters withn-grams.

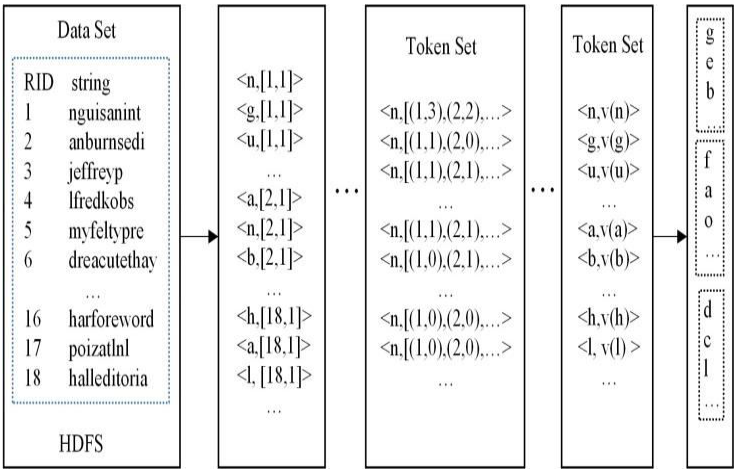


Figure 3 – token based algorithm

B.SequenceBased

The similarity is a factor of common sub-strings between the two strings. The algorithms, try to find the longest sequence which is present in both strings, the more of these sequences found, higher is the similarity score. Note, here combination of characters of same length have equal importance.

Edit Distance Based

1. HammingAlgorithm

This distance is determined by overlaying one string over another and identifying the positions where the strings are different.Classical implementation was intended to handle same-length strings. Some implementations can solve this by adding a prefix or suffixed padding. Nevertheless, the concept is to find the total number of places where one string varies from another.

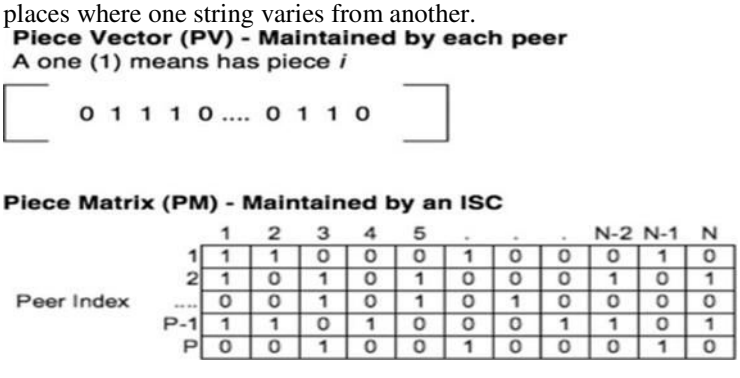


Fig 4 : hamming algorithm

2.Levenshtein distance algorithm

This distance is determined by calculating the number of edits that turn a string into a string. The allowed transformations are insertion-adding a new character, removing-deleting a character and replacing- replacing one character with another. By performingthese three operations, the algorithm attempts to modify the first string corresponding to the second. At the end we get a distance toediting.

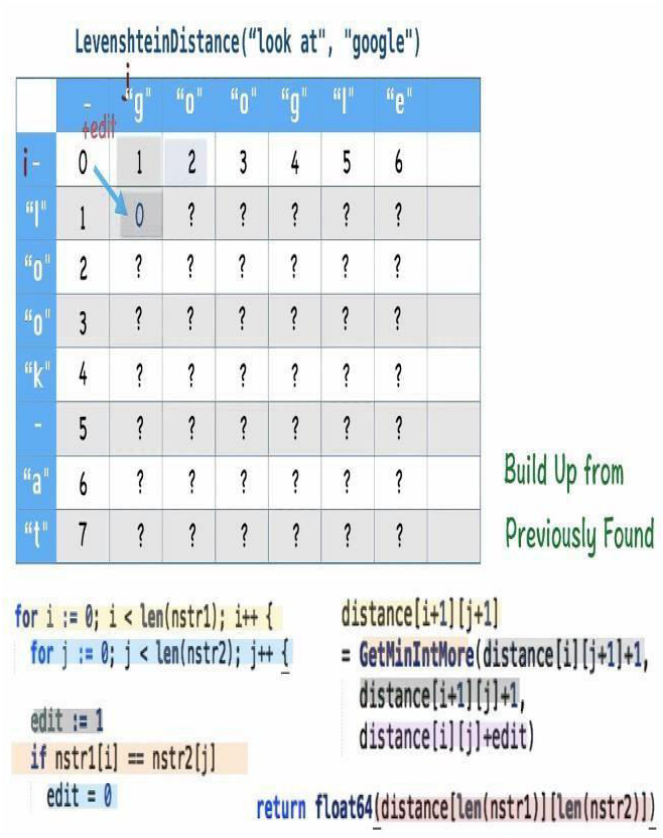


Figure 5 -Levenshtein Distance Algorithm

3.Jaro-Winkler

(1) They contain the same characters, but within a certain distance from each other

(2) The order of the matching characters isidentical.

Tobeaccurate,thedistanceoffindingsimilarcharacteris1 lessthan half of longest string length. The algorithm is directional and gives a high score if matching the strings is from the beginning.

$$d_j = \frac{1}{3}\left(\frac{m}{|s_1|} + \frac{m}{|s_2|} + \frac{m-t}{m}\right)$$

Distance Formula

$$d_w = d_j + (\ell p(1 - d_j))$$

Token Based Algorithm

1.Jaccard index

The formulae, that falls under the set similarity domain, is to find the number of common tokens and divide them by the total number of unique tokens.

Formula

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

2.Sorensen-Dice

The concept is to identify and divide the specific tokens by the total number of tokens present by combining the two sets.

Formula

$$DSC = \frac{2|X \cap Y|}{|X| + |Y|}$$

C.Sequence Based

1.Ratcliff-Obershelp similarity

The idea is intuitive yet quite simple. Find the two strings with the longest common sub string.Delete the part from both strings, and split it at the same spot. It breaks down the strings into two sections, one left and one to the right of the typical substring found.Take now both strings to the left and call the function to find the longest common substring again. .

2.Cosine Similarity

The method used to calculate the degree of similarity is cosine similarity, this approach is a traditional method often used and combined with the TF-IDF method.Cosine Similarity is a measureof similarity between two vectors obtained from the multiplication of the cosine angle of two vectors to becompared.

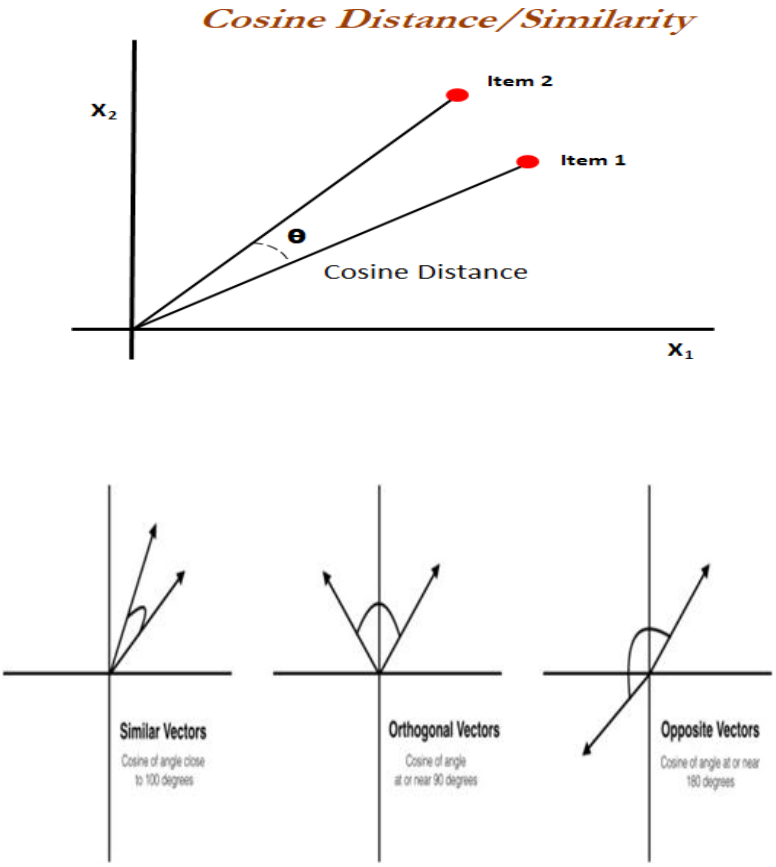


Figure 6 – Cosine Distance Similarity

III. CONCLUSION

Creating NLP-based module & statistical parameter-based module in Data Mining Asit turns out, the inclusion of semantic information in text similarity measurements significantly increases the likelihood of recognition over the random baseline and vector-basedbaselinecosinesimilarityasmeasuredintherecognition paraphrasetask.Thebestperformanceisobtainedusingamethodthat incorporatesmanysimilaritymeasuresintoone,withatotalaccuracy of70.3%,Representingasubstantial13.8%reductionintheerrorrate with respect to the vector-based cosine similarity baselineBy using this Result which give consumer better solution for where to invest their valuable money.

IV. REFERENCE

1] Lu, Jiaheng ; et al. (2013). "String similarity measures and joins with synonyms". Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data: 373–384.doi:10.1145/2463676.2465313. ISBN 9781450320375.Explicit use of et al. in: |first= (help)

2]Navarro, Gonzalo (2001). "A guided tour to approximate string matching". ACM Computing Surveys. **33** (1): 31–88. doi:10.1145/375360.375365.

3]Cohen, William; Ravikumar, Pradeep; Fienberg, Stephen (2003-08-01). "A Comparison of String Distance Metrics for Name-Matching Tasks":73–78.

4]Wael H. Gomaa Computer Science Department Modern Academy for Computer Science & Management Technology Cairo, Egypt" A Survey of Text SimilarityApproaches"

5]SvenKosubDepartmentofComputer&InformationScience, University of Konstanz Box 67, D-78457 Konstanz, Germany "A note on the triangle inequality for the Jaccarddistance"

6] "Cosine Similarity to Determine Similarity Measure: Study Case in Online Essay Assessment" AlfirnaRizqi Lahitani1), Adhistya Erna Permanasari2), Noor Akhmad Setiawan3) (123Department of Electrical Engineering and Information Technology, Faculty of Engineering UniversitasGadjahMada.

7] "A Comparative Analysis of Text Similarity Measures and Algorithms in Research PaperRecommender Systems" MaakeBenard Magara Computer Systems Engineering ,Tshwane University of Technology ,Pretoria, SouthAfrica

8]C. De Boom, S. Van Canneyt, S. Bohez, T. Demeester, and B.Dhoedt, "Learning Semantic Similarity for Very Short Texts,"

9] "Corpus-based and Knowledge-based Measures of Text Semantic Similarity" RadaMihalcea and Carlo Strapparava