Scientific Journal of Impact Factor (SJIF): 4.72

e-ISSN (O): 2348-4470
p-ISSN (P): 2348-6406

International Journal of Advance Engineering and Research Development

Volume 4, Issue 1, January -2017

A System To Measure Text Similarity

Ankush koul[1], Manvi Gupta[2], Pranav Tervankar[3], Prof. Rahul Kadam[4]

[1,2,3,4]*Department of Computer Engineering, Dr D.Y .Patil College of Engineering.Pune*

Abstract — *The aim for proposing a text similarity system is to live the linguistics similarity between 2 documents by victimization the foremost economical formula to live the similarity. The necessity for mensuration text similarity is to boost each program question speed and accuracy, so it's become of nice importance with the advance in web. For humans it's straightforward to search out 2 similar documents supported a selected topic except for a laptop to be trained to live this, stemming and question enlargement techniques is used.The "text similarity tool" is used for mensuration the degree of plagiarism between documents and its originality to an exact extent.*

Keywords: *text similarity tool, Fuzzy algorithm*

I. INTRODUCTION

Textual similarity could be a thought which may be seen as the way of describing the similarity between strings. A string will contain that means, i.e. linguistics are often derived from it. The linguistics similarity of strings could be a special case of linguistics connectedness, that has roots in computing and dates back to 1968. linguistics similarity is employed as a tool to finish similar ideas, whereas linguistics connectedness is employed to finish connected ideas. for example, a automotive and a handwheel square measure additional connected than a automotive and a bicycle, however the latter would be thought of additional similar. linguistics connectedness is employed within the Google programme, wherever it's wont to verify whether or not strings square measure connected in terms of occurrences on a webpage. AN usually used tool for similarity measurements is that the WordNet project developed by Princeton. It consists of words from English people language, wherever every word is expounded to alternative words victimisation relations like synonymity, word and additional advanced ideas like subordination and superordinate word. These relations allows the user of WordNet to find that a automotive and a handwheel square measure connected, dark and lightweight square measure antonyms, bike and bicycle square measure synonyms etc. This project proposes 2 completely different solutions for mensuration matter similarity:
One victimisation the WordNet project to live the linguistics similarity and one victimisation the questionable edit distance between strings. These solutions square measure terribly completely different naturally, therefore a comparison between them are created in addition. This report describes the event of a tool for mensuration matter similarity.
Semantic similarity live could be a central issue in computing, science and science for several years. it's been wide utilized in linguistic communication process, data retrieval, acceptation clarification, text segmentation, question responsive, recommender system, data extraction then on. In recent years the measures supported WordNet have attracted nice concern. They show their skills and create these applications additional intelligent. several linguistics similarity measures are planned. On the entire, all the measures are often classified into four classes: path length based mostly measures, data content based mostly measures, feature based mostly measures, and hybrid measures

II. LITERATURE SURVEY

1. Learning Semantic Similarity for very short texts
Levering knowledge on social media, like Twitter and Facebook , needs data retrieval algorithms to become able to relate terribly short text fragments to every alternative. ancient text similarity strategies like tf-idf cosinesimilarity, supported word overlap, largely fail to supply smart leads to this case, since word overlap is small or nonexistent. Recently, distributed word representations, or word embeddings, are shown to with success permit words to match on the linguistics level. so as to try short text fragments—as a concatenation of separate words—an adequate distributed sentence illustration is required, in existing literature usually obtained by naively combining the individual word representations. we have a tendency to thus investigated many text representations as a mixture of word embeddings within the context of linguistics try matching. This paper investigates the effectiveness of many such naive techniques, also as ancient tf-idf similarity, for fragments of various lengths. Our main contribution may be a start towards a hybrid technique that mixes the strength of dense distributed representations— as hostile thin term matching—with the strength of tf-idf primarily based strategies to mechanically scale back the impact of less informative terms. Our new approach outperforms the prevailing techniques during a toy experimental set-up, resulting in the conclusion that the mixture of word embeddings and tf-idf data would possibly cause a stronger model for linguistics content at intervals terribly short text fragments.

2. Using NLP techniques and fuzzy semantic similarity for automatic plagiarism detection
Plagiarism is one amongst the foremost serious crimes in academe and analysis fields. during this era, wherever access to data has become abundant easier, the act of plagiarism is speedily increasing. This paper aligns on external plagiarism detection methodology, wherever the supply assortment of documents is out there against that the suspicious documents area unit compared. Primary focus is to find intelligent plagiarism cases wherever linguistics and linguistic variations play a crucial role. The paper explores the various pre-processing strategies supported tongue process (NLP) techniques. It any explores fuzzy-semantic similarity measures for document comparisons. The system is finally evaluated mistreatment PAN 20121 information set and performances of various strategies area unit compared.

3. Research on Text Similarity computing based on word vector model of neural networks
Text similarity computing plays a vital role in tongue process. during this paper, we tend to build a word vector model supported neural network, and train Chinese corpus from Sohu News, World News, and so on. Meanwhile, a technique of hard the text linguistics similarity exploitation word vector is projected. Finally, through comparison with the standard calculation methodology TF-IDF, the experimental results prove the strategy is effective.

III. PROPOSED SYSTEM

The system of linguistics matter similarity task has 2 main modules: one is lexical module and another one is dependency parsing based mostly grammar module. each these module have some preprocessing tasks like stop word removal, co-reference resolution and dependency parsing etc.
Figure one displays the design of the system.

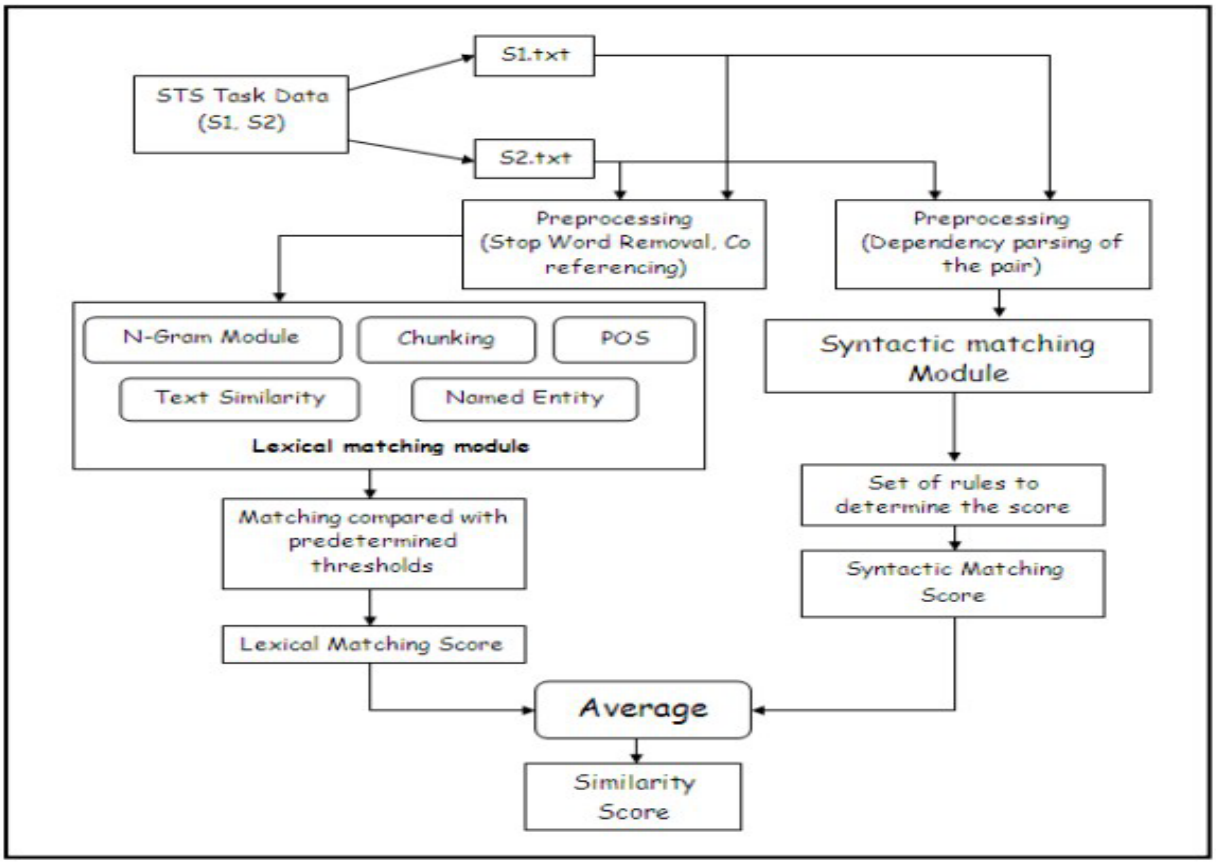

Figure: Planned System Design

ADVANTAGES OF PLANNED SYSTEM:
Computing text similarity could be a foundational technique for a large vary of tasks In tongue process like duplicate detection, question respondent, or automatic essay grading.Text similarity measures play associate more and more necessary role in text connected analysis and applications in tasks like info retrieval, text classification, document cluster, topic detection, topic pursuit, queries generation, question respondent, essay rating, short answer rating, MT, text report et al. Finding similarity between words could be a elementary a part of text similarity that is then used as a primary stage for sentence, paragraph and document similarities.

V. CONCLUSION

The main goal of this project was to implement a tool for measure matter similarity between texts. 2 strategies are projected as solutions to the present problem: measure the linguistics similarity, a special case of matter similarity, wherever the that means of words area unit taking into consideration, exploitation the computer database WordNet and measure the similarity exploitation edit distance. each of the strategies are with success enforced within the final tool. As a minimum demand, it was
specified that it should be doable to perform performance comparisons of the strategies. This practicality has been enforced and envisioned within the graphical interface as 2 bar charts showing the performance of every tool.
We have conducted many tests on the practicality and therefore the code to make sure the tool is functioning properly. These tests was run with success, therefore we have a tendency to conclude that every one identified errors area unit removed and therefore the practicality of every methodology is evidently. This truth is, however, not enough for the ultimate tool to be satisfactory. If the measurements area unit inaccurate in terms of what a person's would take into account similar, the tool isn't terribly helpful. Therefore, many check persons are evaluating the similarity of variety of texts. the typical similarity price for every try of texts has then been compared to the results given by the tool, to validate that the tool is providing helpful information. The results have shown that use of edit distance yield the most effective results, with a correlation of 0:96, whereas use of WordNet gave a constant of 0:76. many of the projected extension are enforced. the look of the similarity engine created it straightforward to implement extensions that permits the user to check quite 2 texts at a time. By exploitation genetic algorithms we've got envisioned the results of a comparison between multiple texts, by making a map wherever similar texts area unit classified along. The tool has been subject to totally different optimizations, all of that have improved the general performance. the foremost effective improvement was a restructuring of text process code, that cause a rise in performance by up to four times. Threads was enforced to permit use of multiple process units. This improvement inflated performance by up to fifteen on a laptop with 2 process units. the ultimate result has been satisfactory. The tool meet the minimum necessities and a number of other extensions has been enforced. By examination the measures performed by the tool with human evaluations, we will conclude that the tool is ready to live matter similarity with very little deviation from what a person's would take into account similar.

ACKNOWLEDGMENT

We might want to thank the analysts and also distributers for making their assets accessible. We additionally appreciative to commentator for their significant recommendations furthermore thank the school powers for giving the obliged base and backing.

REFRENCES

1. Textual Similarity: Comparing texts in order to discover how closely they discuss the same topics Andreas Schmidt Jensen & Niklas Skamriis.
2. Textual Similarity Johan van Beusekom Peter Gammelgaard Poulsen
3. Eneko Agirre, Mona Diab, Daniel Cer, and Aitor Gonzalez-Agirre,"Semeval 2012 task 6: A pilot on semantic textual similarity," In Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation, pp. 385-393.Association for Computational Linguistics, 2012.
4. David Jurgens, Mohammad Taher Pilehvar, and Roberto Navigli, SemEval
5. Ankush Maind, Prof Anil Deorankar, and Dr Prashant Chatur, "Measurement of semantic similarity between words: A survey," International Journal of Computer Science, Engineeringand Information Technology 2, no. 6 (2012): 189-194.
6. www.google.com
7. www.ieeexplore.ieee.org
8. www.wikipedia.org

