Press Release – Similarity Checker

Tamil Virtual Academy

Tamil has an exceptional literary tradition dating centuries. The amount of treatises and work in Tamil since time immemorial which are preserved in stones, papyruses and even now paper is more than any other language. Tamil was also one of the earlier Indian languages to be computerized. The content in Tamil in the web is far more than of other Indian languages.

There has been a need for similarity checking systems in Tamil that will help the authors identify other ideas that have been published before. These similarity check systems will help prevent repetition of ideas and thoughts between authors. Also, it will help new authors to understand what is already done.

With the above needs in mind the plagiarism detection system for Tamil has been developed by Dr. R. Shriram, Professor, Computer Science and Engineering Department B.S. Abdur Rahman University with funding by Tamil Virtual Academy. The key contributors in the project are Mr. M.Loganathan, Mr. Ahamed Khabeer Bhadusha and Mrs. A. Radhika. The time period of the project is 1 year for development and 1 year of support.

The key highlights of the project are

- Detects exact sentence repetition
- Detects word pattern repetition
- Searches the web repositories for Tamil text
- Focuses on Unicode based content repositories which are accessible from the web
- Developed in open source framework
- Presents the results in a percentage form of similarity
- Opens up a new window for the results
- Highlights the word patterns in the input text in the opened up separate window

The similarity detection system has been developed using the Bing search engine as the base and can now handle a minimum of 3 sentences and a maximum of 250 words. The size limits have been optimized such that the end users get their results quickly and these limits will be optimized based on the usage characteristic.

The similarity detection system will prove a milestone for the authors in Tamil as comparable similarity detection systems for English are costly and thus inaccessible for the common users. The system is hosted in the Tamil Virtual Academy website and has been beta tested and modified based on the results from hundreds of users.

Researchers will benefit as they can find if their ideas have been published already by someone or is new. They can also benefit by reading the existing work and then forming their own work. The authors will benefit as their work copyrights are protected. Others cannot claim ignorance as the tools for checking are available. With the copyright laws being in its infancy now especially for Tamil and Tamil Virtual Academy's push towards Digitization resulting in a large corpus being developed and uploaded into the public domain, this work is a step in the right direction and an absolute need of the hour. The General public will benefit from this tool as they can evaluate the quality of writings they see in social media or blogs and improve their own skills. They can also crosscheck the information received with standard text. The system designed will serve as a proof of concept prototype and spur the further development in the domain.