Nepali Plagiarism Detection

Data set

The training and test dataset will be collected manually from different Nepali articles, books and other resources that are available.

Project idea

Plagiarism is when two sentences are similar if they contain the same sequence of consecutive words. Lancaster and Culwin have stated in their paper, "plagiarism as theft of intellectual property which has been around as long as human has produced work of art and research". Basically it is what you try to present someone else's work as your own work without referencing to the original source.

We will be performing monolingual plagiarism detection (Nepali-Nepali). Under monolingual plagiarism, we will first perform intrinsic plagiarism detection and if possible we will move on to extrinsic plagiarism detection as well. The basic idea of intrinsic plagiarism detection is to construct a program to recognize whether a certain pair of articles are similar by measuring the deviation between the text segments using natural language processing (NLP). It does not require any external sources for detection.

We are planning to apply frequency–inverse document frequency (TF-IDF), and cosine similarity as the comparison methodology. However, we will explore other suitable algorithms like vector based, syntax based, fuzzy based, semantic based methods as well. We will first research on the appropriate algorithm for text processing and plagiarism detection method and then try to develop our own algorithm and apply it. Then we will perform experiments to validate our dataset and also try to recognize any sort of plagiarism in other articles.

Software and tools

Programming Language : Python 3.5
IDE : PyCharm/Jupyter Notebook
Libraries : Natural Language Toolkit (NLTK)

Team members and work division

1. Manasi Kattel (18)
2. Araju Nepal (28)
3. Ayush Kumar Shah (44)

S.No.	Work	Team members
1	Data Collection	All
2	Analysis	All
3	Algorithm Study	Manasi, Ayush
4	Design	Ayush, Araju
5	Coding	All
6	Training	All
7	Testing	Araju, Manasi
8	Documentation	All

References

[1] Garg, U. and Goyal, V. (2019). *Maulik: A Plagiarism Detection Tool for Hindi Documents*. [online] Indjst.org. Available at: http://www.indjst.org/index.php/indjst/article/view/86631/68242?fbclid=IwAR3OWyDtI0MwlKEK9TFEvF1JGG80csLGbI0-1ayQyoyUzDew7oIzJoAfSlY [Accessed 16 Apr. 2019].

[2] Pdfs.semanticscholar.org. (2019). [online] Available at: https://pdfs.semanticscholar.org/c4cf/61c6b5aa96bac544c9e997b934403c6d7e26.pdf [Accessed 14 Apr. 2019].

[3]"Part 1-Sinhala Language based Plagiarism Detection in Natural Language Processing", *Medium*, 2019. [Online]. Available: https://medium.com/@erandiganepola/overview-of-natural-language-processing-based-plagiarism-detection-deede2fbeb85. [Accessed: 16- Apr- 2019].