DEEP LEARNING APPROACH TO DETECT PLAGIARISM IN SINHALA TEXT

T.C. Kasthuriarachchi , E.Y.A.Charles

Department of Computer Science, Faculty of Science, University of Jaffna

tharuka.ckasthuri@gmail.com,charles.ey@univ.jfn.ac.lk

ABSTRACT

Drawbacks and inefficiency of existing language independent plagiarism detection tools and lack of adequate research for Sinhala plagiarism detection this research work was motivated.

Proposed method focuses on developing a deep learning based approach for plagiarism detection in Sinhala documents

To improve the efficiency of the model Natural Language Processing techniques have been applied.

The proposed model was implemented and tested on an in house data set and found to be capable of detecting plagiarism with an accuracy of 97%.

The model is capable to detect direct and sophisticated copying such as replacing the words with their synonyms as well as changing the order of words in a sentence.

OBJECTIVE

The main objective of this research is to address the actual need of automated plagiarism detection tool for Sinhala language and developing a plagiarism detection corpus for future researches.

Corpus for Developing Sinhala Word2vec model

- Used the Sinhala News Corpus published the Language Technology Research Laboratory of University of Colombo School of Computing.
- The documents of the corpus were in utf-8 and utf-16 Unicode formats, so all the documents were converted to utf-8 format using a Linux script.
- Since the documents in the corpus contained lots of noises such as HTML syntax and punctuations, they were cleaned using NLTK python library.

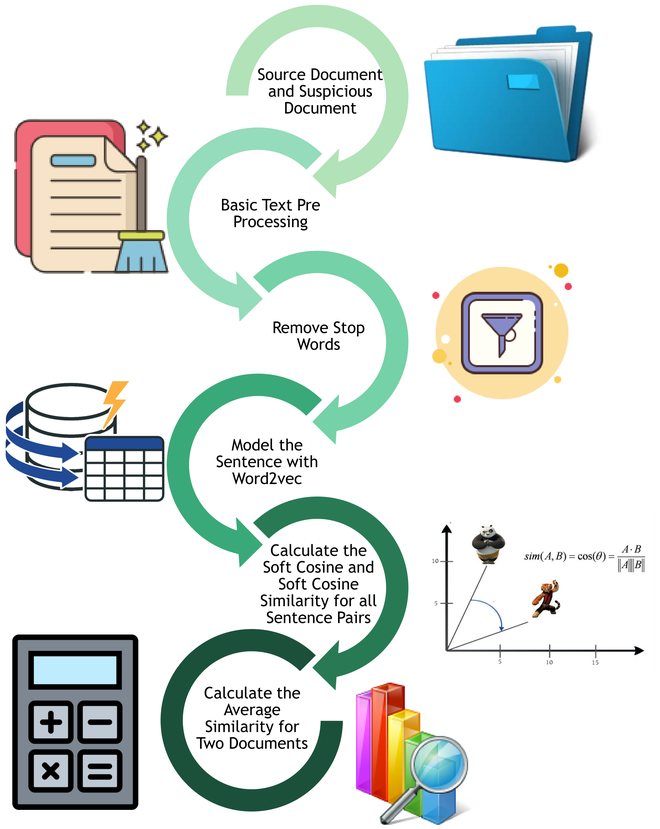
Plagiarism Detection Corpus

- From the UCSC Sinhala News Corpus, 50 documents were randomly selected and the sentences were plagiarised by a selected group of students.
- Plagiarising included changing word order and replacing words with similar words.
- In order to reduce the error probability, the plagiarized text has been rechecked by a Sinhala language experts. In addition the sentences were tagged as whether similar or not manually by the expert.

METHODOLOGY

Phase I

- word2vec model was constructed using a selected Sinhala text corpus.
- Before training the word2vec model, common bigram phrases from the text were extracted. Hence some common phrases like (Sri Lanka) are considered as one word.
- Basic text pre-processing steps such as tokenization, punctuation removal and stop word removal are applied to achieve maximum possible accuracy.
- After experimenting with CBOW and Skip Gram, CBOW method is selected.
- Training loss and perplexity were used as performance measure of the training.



Phase II

- Word vectors for the words in a source sentence (S1) and target sentence (S2) are computed using the word2vec model (see below illustration).

$$WV_1 = [w_{11}\ w_{12}\ \dots\ w_{1j}\ w_{1j+1}\ \dots w_{1d}]$$
$$WV_2 = [w_{21}\ w_{22}\ \dots\ w_{2j}\ w_{2j+1}\ \dots w_{2d}]$$
$$\vdots$$
$$WV_i = [w_{i1}\ w_{i2}\ \dots\ w_{ij}\ w_{ij+1}\ \dots w_{id}]$$
$$\vdots$$
$$WV_n = [w_{n1}\ w_{n2}\ \dots w_{nj}\ w_{nj+1}\ \dots w_{nd}]$$
$$SV_k = [sv_{k1}\ sv_{k2}\ \dots sv_{kj}\ sv_{kj+1}\ \dots sv_{kd}]$$

- These vectors are used to generate the word vectors for the sentences using (1).

$$sv_{kj} = \frac{\sum_{i=1}^{n} w_{ij}}{n} \qquad (1)$$

- After modelling the sentence with word vectors, all sentences in a target document are compared with all sentences in the source document using the cosine similarity metric (2) and the soft-cosine (3) similarity metric, (generalization of cosine similarity).

$$cosine(S1, S2) = \frac{\sum_{i=1}^{k} S1_i S2_i}{\sqrt{\sum_{i=1}^{k} S1_i^2}\sqrt{\sum_{i=1}^{k} S2_i^2}} \qquad (2)$$

$$softcosine(S1, S2) = \frac{\sum_{i=1}^{k}\sum_{j=1}^{k} S1_i M_{ij} S2_j}{\sqrt{\sum_{i=1}^{k}\sum_{j=1}^{k} M_{ij} S1_i S1_j}\sqrt{\sum_{i=1}^{k}\sum_{j=1}^{k} M_{ij} S2_i S2_j}} \qquad (3)$$

- If the similarity measure of two sentences is higher than a set threshold value, then the target sentence is considered as plagiarized from the source sentence.
- Finally, the average similarity scores are calculated for the target document.

Sample Results Taken from Word2vec Model

Most Similar words for word දෙවියන්	
දෙවියන්ගේ	0.6743
විශ්ණු	0.6706
යෙහෝවා	0.6517
දෙවිවරුන්	0.6422
දෙවියන්ට	0.6364
දෙවි	0.6158
දෙවියන් වහන්සේ	0. 6025

RESULTS

Suspicious Sentence	Original Sentence	Soft Cosine	Cosine	Expert Judgement
දෙදියගල දකුණු පළාතේ පිහිටි ස්භාවික සෞන්දර්යෙන් අනූන සුන්දර ගම්මානයකි.	දකුණු පළාතේ මාතර දිස්ත්රික්කයේ අකුරැස්ස ප්රාදේශීය ලේකම් කොට්ඨාසය තුළ පිහිටි ස්වාභාවික සෞන්දර්යෙන් අනූන දෙදියගල සුන්දර ගම්මානයකි.	0.6757	0.7741	Similar
දෙදියගල ස්වභාවික සෞන්දර්යයෙන් අනූන දකුණු පළාතේ මාතර දිස්ත්රික්කයේ අකුරැස්ස ප්රාදේශීය ලේකම් කොට්ඨාසය තුළ පිහිටි සුන්දර ගම්මානයකි.	දකුණු පළාතේ මාතර දිස්ත්රික්කයේ අකුරැස්ස ප්රාදේශීය ලේකම් කොට්ඨාසය තුළ පිහිටි ස්වාභාවික සෞන්දර්යෙන් අනූන දෙදියගල සුන්දර ගම්මානයකි.	0.8999	0.9790	Similar
පවුල් 404ක් පමණ ජීවත්වන මෙම ගම්මානය සිංහරාජ අඩවිය පාමුල උඩුගම කල්එළිය ගමට මායිම්ව පිහිටා තිබේ.	දකුණු පළාතේ මාතර දිස්ත්රික්කයේ අකුරැස්ස ප්රාදේශීය ලේකම් කොට්ඨාසය තුළ පිහිටි ස්වාභාවික සෞන්දර්යෙන් අනූන දෙදියගල සුන්දර ගම්මානයකි.	0.1896	0.5664	Not Similar

Threshold value	False Positive	False Negative	Accuracy
Using Cosine Similarity			
0.55	2	115	0.9539
0.69	38	41	0.9689
0.90	386	3	0.8466
0.99	507	0	0.8001
Using Soft Cosine Similarity			
0.30	10	71	0.9681
0.32	11	63	0.9708
0.39	44	45	0.9649
0.50	132	18	0.9409

CONCLUSIONS

Performance of the proposed model is based on a relatively small data set which was created by a group of students.

Annotating a sentence whether it is a plagiarised one or not in comparison with original sentences was performed by a single expert.

Work is still needed to construct a large and well evaluated Sinhala plagiarism corpus, such that the performance of the model can be well evaluated.