Plagiarism Search within One Document

Máté Pataki
Department of Distributed Systems
Computer and Automation Research Institute of the Hungarian Academy of Sciences
Mate.Pataki@sztaki.hu

Abstract. This paper is a usability study of a plagiarism search method proposed by Csernoch Mária at the II. Hungarian Computer Linguistics Conference. The method promises to be able to detect plagiarism is a document based on the change in style between different parts of the document.

Keywords: plagiarism search, similarity, word frequency, sentence length

1 Introduction

Plagiarism, the citation of the work of an other person without giving him the credit, became one of the largest problems in the last decade which universities have to face. This is the reason why more and more plagiarism search services and programmes appear.

There are three different approaches of automated plagiarism detection, the first is to compare documents pairwise to each other and detect the overlapping [1], the second approach is to compare a document to a database of documents and see if there are similar ones among those [2]. Both of these two could be realized as online services or standalone programmes, and both require that the document which was plagiarised is available in the local repository. Because many cases of plagiarism involve documents from the Internet, which is the largest existing repository, a complete comparison with all the pages is not possible.

To search for similar documents on the Internet [3] one could use example sentences or phrases from the suspected documents, but which ones to choose is difficult to determine. The third approach of plagiarism detection could be helpful in this particular case, by sectioning the document based on the change in style between sentences and paragraphs (style markers). After the suspicious parts are extracted, one can search for one or two sentences in each part on the Internet. This method of sectioning was proposed by Csernoch Mária at the II. Hungarian Computer Linguistics Conference [4]. She used it for analysing literary works and discovered that the changes in style signal important places in the work, which on other hand could be used to extract parts written by other authors. The following section is usability study, to see if this method is suitable for plagiarism search or not.

2 2. Style markers

Style markers are statistical parameters, like sentence length or number of new words per sentence, which are calculated for different texts and are compared to see the similarities and differences between them.

The documents, which are uploaded to a plagiarism detection system, could be written in any possible language. Therefore, only language independent style markers are examined in this paper. It could be the scope of another paper to see how much would be gained by using the information embedded in the particular language.

The formatting of the documents contains a lot of information about the document itself, but one can assume that if somebody does adopt the sentences of another person, at least the formatting will be changed according to the current documents style, so this information won't be used. The structure of a document like sentences, clauses and words can be easily identified and are language independent, in the sense that one needs only to know the alphabet of the current language.

3 Calculating the style markers

For the analysis the following five books are used:

- Harry Potter and the Goblet of Fire (by J.K. Rowling)
- Harry Potter and the Sorcerer's Stone (by J.K. Rowling)
- Robinson Crusoe (by Daniel Defoe)
- The Hitch Hiker's Guide to the Galaxy (by Douglas Adams)
- The Lord Of The Rings, The Fellowship Of The Ring (by J. R. R. Tolkien)

Two books of J.K. Rowling are used to see if the style between those two is more similar than the others, which would be a welcome result if one can assume that they are both really written by the same author.

The easiest statistics that can be calculated are average word and sentence length, number of words used, number of distinct words, length of the sentences and subsentences. For all the five books the statistics can be found in the following table.

As can bee seen from the table, three markers can be identified as having significant differences between the works, namely:

1. Number of distinct words
2. Average sentence length
3. Average sub-sentence length

	Robinson Crusoe	Sorcerer's Stone	Goblet of Fire	The Hitch Hiker's Guide	The Fellowship Of The Ring
Number of characters	480538	335007	848630	208392	771146
Number of words	121839	80624	197473	47978	190732
Number of distinct words	6038	5764	10227	5962	8700
Average word lengths	3.94	4.16	4.30	4.34	4.04
Number of sentences	2859	6711	15798	3741	13560
Average sentence lengths (words)	42.62	12.01	12.50	12.82	14.06
Number of subsentences	16787	12476	31290	6438	26440
Average subsentence lengths	7.26	6.46	6.31	7.45	7.21
Average subsentence per sentence	5.87	1.86	1.98	1.72	1.95

Table 1: Statistics calculated for the five books

Average word length is determined by the language of the document (English in this case) and does not differ enough to be significant. Number of distinct words can be considered as being the dictionary from which the writer worked. This could be used to compare the diversity of words between sections. Average sub-sentence per sentence can be calculated from the average sentence length and average sub-sentence length so does not add any extra information to the calculation, and won't be used.

4 Results - Detecting the changes

Even if two authors use the same number of words in their works, the dictionary won't be the same. The same word can be in the working vocabulary of one author while in recognition vocabulary of the other, so each author has his own style, even when writing from the same topic. This information can be used to scam through the document and mark for example the number of new words (which have not been used before) per sentence. The result will be a diagram where the section borders should be visible.

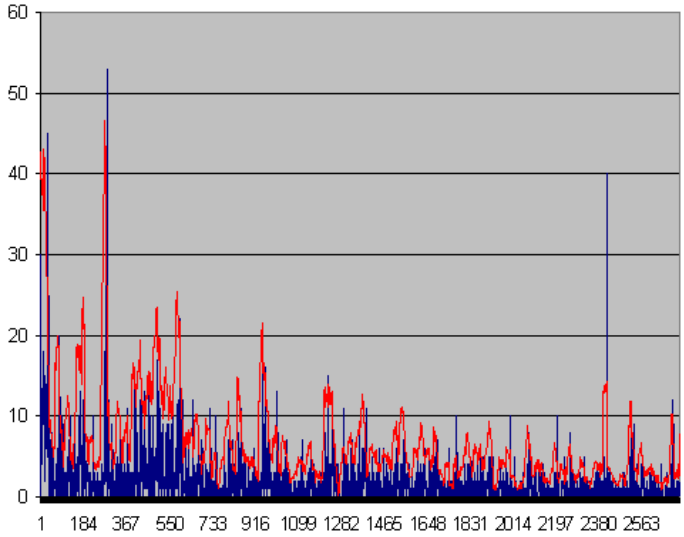

Figure 1: Number of new words per sentence (red is a moving average)

This figure shows a document where the first half is a 100k part of The Hitchhikers Guide and the second half is The Sorcerers Stone. As can be seen the there are many places where a lot of new words come in, so the place where the change occurs at the middle of the document, cannot be distinguished from the others. This means that this information cannot be used alone for identification of the borders of the different parts.

The same diagram for Robinson Crusoe and The Fellowship of the Rings looks really different and the border can be clearly detected at 473.

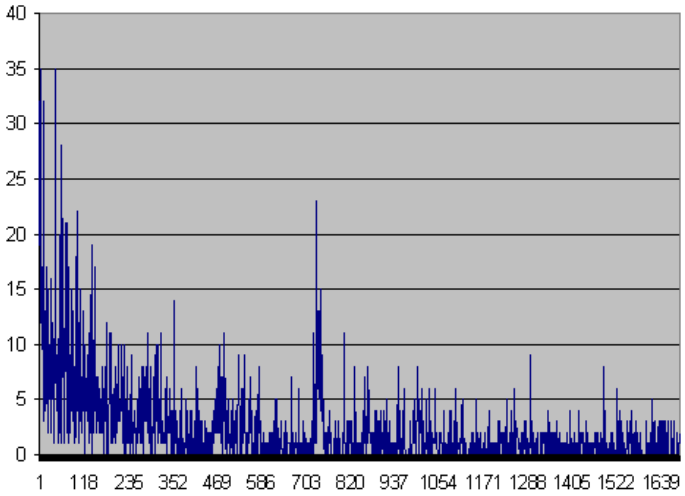

Figure 2: Number of new words per sentence (Robinson Crusoe and The Fellowship of the Rings)

When normalising the number of new words with the number of words in the sentence, the diagram looks clearer, and the border is more visible.

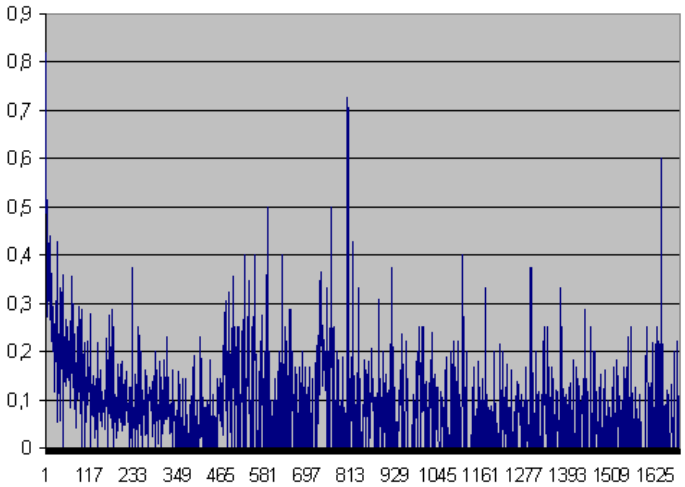

Figure 3: Percentage of new words per sentence (Robinson Crusoe and The Fellowship of the Rings)

The number of distinct words used in by Daniel Defoe is less than the number of words used by J. R. R. Tolkien, so this could explain the peak, but with the same two 100k parts but the other way around this looks only a bit different.

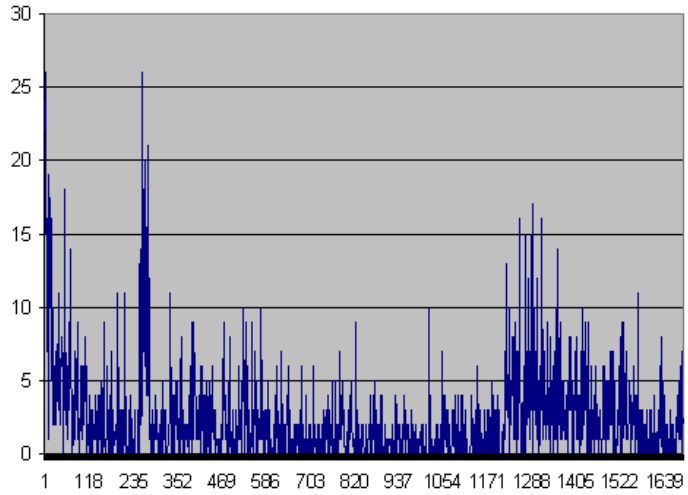

Figure 4: Number of new words per sentence (The Fellowship of the Rings and Robinson Crusoe)

The border is at sentence number 1217, and can clearly be identified. The sentence length in this case looks like follows:

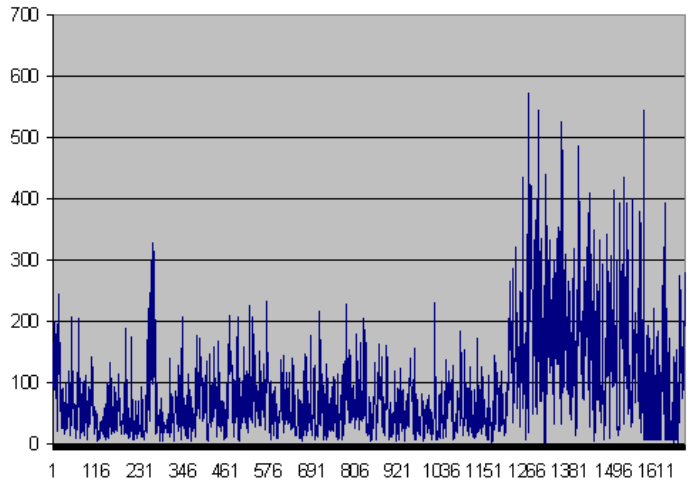

Figure 5: Sentence length (The Fellowship of the Rings and Robinson Crusoe)

This last result was to be expected after knowing that in the one book the average sentence lengths is almost four times as long as in the other. Unfortunately, this is not always the case: In many cases the sentence length is not a good marker. In the first mixed document (The Hitch Hiker's Guide and The

Sorcerer's Stone), where the average sentence lengths are almost the same, the border between the two vanishes.

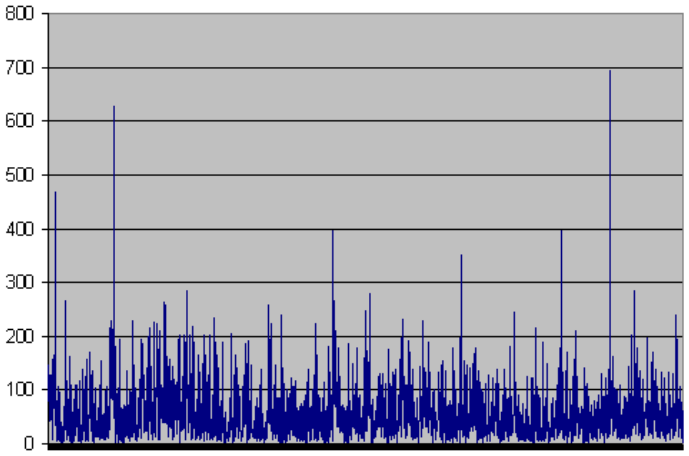

Figure 6: Sentence length (The Hitch Hiker's Guide and The Sorcerer's Stone)

5 Summary

With the above results one can say that this approach could work with some documents, but in many cases, if the copied part is small or it is written in about the same style, this approach would show up a lot of false positive section borders, and sometimes wouldn't show up the real ones.

While analysing the above results an other approach seems to be more effective. The sentences which are sticking out by having a lot of new words or being long are really unique ones, and can be found in almost every part of the document. It would be interesting to see if one could find more plagiarism by searching for those ones, or by searching for the same number of sentences but evenly distributed. This could be the scope of a later paper.

Acknowledgement

The author would like to express his thanks to László Kovács for his support as a scientific advisor.

References

[1] "Eve plagiarism detection system," http://www.canexus.com.

[2] "Kopi online plagiarism search and information portal," http://kopi.sztaki.hu.

[3] "Plagiarism search v 1.0.0," http://baltic.cse.msu.edu/heynige1/Search.

[4] C. Mária, "A szavak véletlenszerű megjelenésén alapuló modellek és az irodalmi művek közötti eltérések magyarázata," II. Magyar Számítógépes Nyelvészeti Konferencia, pp. 211–218, 9-10. December 2004.