http://www.moderntechno.de/index.php/meit/article/view/meit23-02-026
DOI: 10.30890/2567-5273.2022-23-02-026

METHODS OF CHECKING THE AUTHENTICITY OF TEXTS

Voitsekh V.
[1] *Ukrainian American Concordia University,*
8/14, Turhenievska Street, 01054, Kyiv, Office 1-4

Abstract. *A general overview of the available methods for verifying the authenticity of texts was performed, the advantages and disadvantages of each were analyzed. The shingle method has been implemented, as well as a modified string-matching algorithm, which allows you to find even modified and paraphrased plagiarism with high accuracy. The results obtained by various verification methods were analyzed.*

Key words: *Authenticity of texts, Shingles method, Plagiarism, Text analysis, Document Indexing*

Introduction

When checking texts for authenticity, the main indicator is the absence of plagiarism in the text. Plagiarism is the use of another person's work in any form and presenting that work as one's own without citing the original work. Most often, it is used for personal gain. Currently, the following types of borrowed text are distinguished:

1. direct plagiarism, i.e. copying someone else's work without indicating the source
2. self-plagiarism — using one's own previous works
3. accidental plagiarism, i.e. failure to indicate the original source is unintentional
4. outsourcing — hiring other writers, bloggers, friends to write on your behalf.

The problem of detecting plagiarism has become widespread with the advent of the Internet, and each year it becomes more difficult to find the original author. This problem is actively studied from an engineering and scientific point of view. On the other hand, the fight against existing methods of identifying plagiarism is also actively developing, namely:

- deep text rewriting
- retelling the text in your own words
- use of synonyms and epithets
- adding automatic transfers
- translation of foreign language texts
- use of materials not indexed in search engines
- replacing frequently repeated words

However, every year systems for checking the uniqueness of texts improve and not every bypass can work. Today, there are hundreds of different resources for checking texts for authenticity. In addition to information about whether the text is unique, various implementations may additionally indicate the sources in which the text was found, and which are considered original; the percentage of insignificant and stop words in the text, and others.

Most plagiarism detection systems implement one of two general detection approaches, which can be conventionally called external and internal. The external approach is to compare a suspect document with a certain reference set, which is a set

of documents that we believe to be completely authentic. Based on the document similarity criteria, as well as the selected model, the task is to find all documents that have text that is like the text in the suspicious document above the selected degree of similarity. The internal approach is exclusively the use of certain techniques to analyze suspicious text without performing comparisons with external documents.

The Fingerprint method

The fingerprint method is currently the most widely used method for detecting plagiarism. This method represents each document as a sample of a certain number of substrings (n-grams). Sets indicate the imprint of each document. A suspected document is checked for plagiarism by calculating its fingerprint and comparing the resulting fingerprint with previously calculated fingerprint indices for all documents in the reference collection. If the prints match, then the corresponding text segments also match and are potential plagiarism if they exceed the chosen similarity threshold.

The limiting factors of this method are computational resources and time, so this approach is often used to check only a subset of fingerprints, with the aim of speeding up the calculation and being able to check on a large collection of reference documents

A common practice when working with a large set of documents is to create an imprint of each document. In a more general sense, a print is a lower-order representation of each document, and accordingly, a set of documents. To do this, you can use the hash function sha1 or md5. The fingerprint method can be used for different purposes depending on the hashing function.

Another possible approach is to use word filtering based on IDF (Inverted Document Frequency) values, which are calculated before processing the entire document corpus.

String-searching algorithm

When applied to the problem of authenticity of texts, documents are compared for word-for-word matches. A variety of methods have been proposed for this task, some of which have been adapted to an external plagiarism detection approach.

Validating a document using this method requires a lot of computation and storing all the documents in the reference collection in a comparable form in order to compare them pairwise. As a rule, suffix models of documents, such as suffix vectors or suffix trees, are used for this. However, the use of this approach remains quite computationally expensive, which makes it unsuitable for checking large collections of documents.

Citation based plagiarism detection

Citation-based plagiarism detection relies on citation analysis and is therefore the only approach to plagiarism detection that does not rely on textual similarity. This method examines citation and background information in texts to identify similar patterns in citations. Therefore, this approach can be applied to scientific texts or other academic documents that contain citations.

Prototypes with citation-based plagiarism detection systems already exist. Citation proximity and similar order are the main criteria used to calculate document similarity. The absolute or relative share of total citations, the probability that the citations coincide in the document are also considered.

Stylometry

Stylometry uses statistical methods to determine an author's unique writing style and is primarily used to determine authorship. By creating and comparing stylometric models for different segments of the text, it is possible to identify paragraphs that are stylistically different from the rest, and therefore potentially borrowed.

The shingles algorithm

In 1997, Andrii Broder and Udi Manber proposed a "syntactic" method for evaluating the similarity of documents, which is based on the representation of a document in the form of a set of various sequences of a specific length, which consist of neighboring words. Such sequences are called "shingles". Two documents will be considered similar if their shingle sets overlap. The number of shingles roughly corresponds to the length of the number of words in the document, which is usually a large number.

Stages of text comparison for similarity:

1. canonization of the text.
2. breaking the text into shingles.
3. calculation of shingle hashes.
4. preparation of a sample of control amounts.
5. checking texts for plagiarism.

Implementation of the prototype

The pan13-text-alignment set of documents was used to develop the prototype. This corpus contains more than 3,000 original documents and about 2,000 documents that are suspected of plagiarism.

The first step was to create a basic version based on the shingle method. In most cases, this algorithm is used to find copies and duplicates, however, using various modifications, it can also be used to find individual paragraphs containing plagiarism. For this purpose, the algorithm for checking for plagiarism of a new document can be conditionally divided into two separate stages:

1. finding the most suspicious original documents
2. detailed comparison with each of the found documents.
3. Modified algorithm

We have a set of original unique documents.

1) The first stage of their processing will be the separation of words and parts of speech for each word. For this purpose, you need to use MaxentTagger, which belongs to the Stanford POS Tagger library. The document is divided into sentences, then each sentence is analyzed. Words are selected from the received list of sentence elements. Next, the words are filtered by removing stop words, and words whose length is less than three letters are additionally screened out. The remaining words are entered in the word list of the current document.

2) The next step is word normalization. The word is cleaned from unnecessary characters and with the help of the library for working with WordNet, its basic form is returned for each word. This process is performed for the following parts of speech: noun, verb, adjective, adverb.

3) Document indexing. Using Apache Lucene, we add the document and additional information about it to the directory where they will be stored. All the text

from the document, which after the previous steps is presented as separated words in the basic form, is stored in one field.

4) All words are extracted from the document, regardless of the number, and entered into a separate set, which, after processing the entire corpus, will be used to find the IDF of each word. In this way, a list of words is formed, which can be used as a dictionary, as well as a correspondence between the word and the number of documents in which this word is written is created.

5) Repeat previous steps for each document from the existing set of documents.

6) Calculate the IDF. Words whose IDF is less than a certain threshold will not be considered during the check in addition to stop words.

7) Additional algorithm is used to do the comparison of the data after it was normalized using the steps above.

Summary and conclusions.

The paper considered such an algorithm as the shingles method, as well as its modification for searching for plagiarism that was rewritten. A set of PAN-13 documents was used for testing, consisting of about five thousand documents, including more than three thousand original ones and more than one thousand suspected of plagiarism.

It can be concluded that the modified algorithm works much better in complex cases. It is based on the search for sequences of words and their synonyms in the indexed documents of the collection. To improve the results, used document indexing and full-text search in Apache Lucene, Stanford POS-tagger for sentence analysis and finding parts of speech for each word, and WordNet dictionary for finding synonyms of words.

The modified algorithm shows a low speed of work when checking large volumes of text, or in the case when the library of original documents is very large. In this case, the following modification can be applied: break the process of checking a document for plagiarism into two separate stages. The first step will be to use the shingle method to find the most similar to the suspect original documents from the library. At the second stage, a modified algorithm will be checked, but the document will be compared only with those documents that were found in the previous stage, and not with the entire library. This approach will make it possible to speed up the inspection.

References

1. Jaccard Similarity and Shingling, https://www.cs.utah.edu/~jeffp/teaching/cs5955/L4-Jaccard+Shingle.pdf
2. The Shingles algorithm, https://en.ryte.com/wiki/Shingle_Algorithm
3. WordNet. George A. Miller (1995). WordNet: A Lexical Database for English. Communications of the ACM Vol. 38, No. 11: 39-41. Christiane Fellbaum (1998, ed.) WordNet: An Electronic Lexical Database. Cambridge, MA: MIT Press, https://wordnet.princeton.edu/
4. Apache Lucene, https://lucene.apache.org/
5. Stanford Log-linear Part-Of-Speech Tagger, https://nlp.stanford.edu/software/tagger.shtml

6. Stein, Benno; Koppel, Moshe; Stamatatos, Efstathios (Dec 2007), "Plagiarism Analysis, Authorship Identification, and Near-Duplicate Detection PAN'07" (PDF), SIGIR Forum, 41, https://www.uni-weimar.de/medien/webis/publications/papers/stein_2007o.pdf

7. Dreher, Heinz (2007), "Automatic Conceptual Analysis for Plagiarism Detection" (PDF), Information and Beyond: The Journal of Issues in Informing Science and Information Technology, 4: 601–614, http://proceedings.informingscience.org/InSITE2007/IISITv4p601-614Dreh383.pdf

8. Stylometry-based Fraud and Plagiarism Detection for Learning at Scale, https://www.researchgate.net/publication/271836873_Stylometry-based_Fraud_and_Plagiarism_Detection_for_Learning_at_Scale

9. Top 10 Free Plagiarism Detection Tools, https://elearningindustry.com/top-10-free-plagiarism-detection-tools-for-teachers

10. The Winning Approach to Text Alignment for Text Reuse Detection at PAN 2014, http://ceur-ws.org/Vol-1180/CLEF2014wn-Pan-SanchezPerezEt2014.pdf

11. Finding near-duplicate documents, http://www.cs.princeton.edu/courses/archive/spr08/cos435/Class_notes/duplicateDocs_corrected.pdf

Article sent 23.10.2022