Testing of plagiarism detection tools for Czech environment

Jan MUDRA, Dita DLABOLOVÁ
Mendel University in Brno, Czech Republic

Introduction

There exist many text matching systems being used as plagiarism detection tools in the current global market. Naturally, the main focus of the systems is on the most world-widely used languages. Even if many systems claim they work for any language, an even if they do, we miss answer on a question – how efficient they are on "minor" languages? Hence many smaller countries, such as Slovakia rather develop and implement their own national systems (Kravjar, Noge, 2013). This brings further questions – is it necessary? Do the national systems perform better on the languages to which they are tailored to? These are few of many questions which should be answered within an international project "Testing of Support Tools for Plagiarism Detection (TeSToP)" which aims to perform a complex testing of so-called plagiarism detection systems using documents in multiple languages, including Czech. Results for this language and for Czech environment are presented in this contribution.

Objectives

The goal of the testing for the Czech settings are the same as for the entire TeSToP project, i.e. testing and evaluating almost 20 support tools for plagiarism detection. In addition, text matching system Odevzdej.cz is evaluated for selected documents. This review includes coverage evaluation and usability evaluation.

Note

This paper is based on a bachelor thesis of the author, which is supervised by the co-author. The thesis covers the topic more widely and it is written in Czech language, this contribution sums up the most important points of the thesis and presents them in English, as they might be interesting to readers interested in the performance of systems for plagiarism detection for languages related to the Czech language.

Testing of Support Tools for Plagiarism Detection (TeSToP)

With the above-mentioned motivation, an international team of volunteers was established in 2018, consisting of professors and students of several universities. The team member is also prof. Debora Weber-Wulff, who led plagiarism testing with her team in 2013 (Weber-Wulf et al, 2013). Since then, no comprehensive testing has been repeated. The team was tasked with testing and evaluating plagiarism detection systems that were selected based on previous knowledge and experience. System representatives were contacted to obtain their agreement with their participation in the testing.

Situation in the Czech Republic

Almost all universities in the Czech Republic use the plagiarism system Theses.cz, which is also used as a database of theses. This system is developed and operated by the Faculty of Informatics of Masaryk University in Brno (FI MUNI, n.d. b). The system is not an official national system and universities are not obliged to use it, but due to its wide used, it can be considered as de facto national system. Theses.cz has a sister project named Odevzdej.cz ("odevzdej" in Czech which means "submit"), which is an e-learning tool performing also text matching using the same database as Theses.cz. The similarity detection function is publicly available, anybody can register and immediately upload documents for verification of similarities (FI MUNI, n.d. a). After uploading of a document, the user receives a confirmation email within few days, it contains the result of the evaluation in the form of a number representing how many percent of the text is considered to be a plagiarism. If the system did not find any text matches and showed 0%, the document test is finished. If the system reported some plagiarism, the detailed report is made available after a payment (27 CZK, which is approximately 1 EUR). The paid report contains highlighted text marked as plagiarism and on the right side there are sources from which it was plagiarized.

Methodology

Testing documents

The methodology overlaps with the methodology of the TeSToP project in general. The set of testing documents mainly overlaps with the TeSToP testing set for other languages. It is composed of following documents:

- Wikipedia article in Czech, Slovak and English - divided into multiple documents using different types of plagiarism: copy&paste plagiarism, the same text with copy&paste plagiarism with white characters replaced by a letter (Czech only), the same text with copy&paste plagiarism containing image instead of the text (Czech only), text with replaced synonyms, and paraphrased text.
- Part of a master theses in Czech submitted in 2010, publicly available online - divided into three documents using different types of plagiarism: copy&paste plagiarism, text with replaced synonyms, and paraphrased text.
- Open access article in Slovak and in English – both divided into three documents using different types of plagiarism: copy&paste plagiarism, text with replaced synonyms, and paraphrased text.
- Translation of an English Wikipedia article to Czech, Slovak and English – always with one part translated by Google Translator only, the other part translated manually.
- Original document - in Czech, Slovak and English.
- A translation of a document in Slovak language to Czech.

The Slovak and English language were added to the Czech test as many Czech universities enables submission of theses in these languages (mainly by the international students, and Slovak students who form quite significant percent of students at Czech universities).

The specific documents for the Czech language are two obfuscating methods - white characters and text as an image and translation from Slovak language. This specific translation

was selected due to a big similarity of Czech and Slovak language, any native Czech speaker can understand Slovak language without a problem (and vice versa).
The tested systems

The 17 tested systems overlap with the systems in the TeSToP project, with one exception - the Czech set was also tested on the system Odevzdej.cz.

The course of testing and evaluation

The course of testing was identical with the TeSToP project to ensure the comparability of results among different languages, hence for more details please see that conference contribution.

Preliminary Results

We will combine complete results from coverage evaluation and evaluation of usability.
Coverage evaluation:

- whether the system marked plagiarized text as plagiarism,
- to what extent has the plagiarized text been marked as plagiarism by the system
- whether the system found the right source from which it was plagiarized,
- whether the system has found other resources.

Preliminary results show that StrikePlagiarism and Urkund are the best for the Czech language. On the other hand, the Slovenian system DVP was the worst.

This rating is different from all other languages. In the preliminary results of all language packs together, the Urkund system, PlagiarismCheck and Turnitin are the best plagiarism detection systems. The worst hit was iPlagiarism.net.

Nevertheless, the final results can significantly differ, as the usability evaluation and the evaluation of the obfuscating methods has not been performed yet.

Keywords: plagiarism detection, text matching software, testing, Czechia, Czech language.

References

FI MUNI (n.d.). Odevzdej.cz - Odhalování plagiátů v seminárních pracích. Retrieved from http://odevzdej.cz [cit. 04-16-2019]

FI MUNI (n.d.). Vysokoškolské kvalifikační práce. Retrieved from http://theses.cz [cit. 04-16-2019]

Kravjar, J., Noge. (2013). Strategies and Responses to Plagiarism in Slovakia. In: *Plagiarism Across Europe and Beyond*. Brno: MENDELU Publishing Centre, p. 201-215. ISBN 978-80-7375-765-6.

Weber-Wulff, D., Möller, C., Touras, J., & Zincke, E. (2013). Plagiarism detection software test 2013. Retrieved from http://plagiat.htw-berlin.de/wp-content/uploads/Testbericht-2013-color.pdf