Comparing text-matching software systems using the document set in Latvian language

Laima KAMZOLA, Alla ANOHINA-NAUMECA
Riga Technical University, Latvia

The formal education especially diploma of higher education often opens the doors to the career opportunities and success in the future life. Not all the students are going a fair path to receive the acknowledgment of education acquired. Moreover, it is seen and discovered that also teachers, employers and employees are plagiarizing their works as well. The more developing are technologies, the more complex and unseen ways how students are cheating are discovered. Therefore there is a need to create more advanced tools to more precisely detect and afterwards make a very detailed report indicating and approving the existence of plagiarism in plagiarized students' or even teachers' work.

Text-matching software systems are usually used for revealing plagiarism. They detect whether equal or similar parts of text can be found in other written sources which are located in databases which is a significant part of any text-matching software. Since there are many different languages in which students and tutors can complete their formal education works, text-matching software can be made for both international and country specific needs. In some countries these tools are united and one software is used for all universities in cooperation with government, for example, in Slovakia such a software system is called "Antiplag". It is used in all Slovakian universities and financially supported by the Ministry of Education, as well as there is a regulation that any student's final thesis need to be evaluated by this software before the student can defend the thesis (Kravjar, 2018).

The paper presents a part of results acquired by an international initiative "Testing of Support Tools for Plagiarism Detection (TeSToP)" in regard to the comparison of different text-matching software systems based on a document set prepared by Latvian participants. The document set included both paraphrased and translated texts from English to Latvian and Russian to Latvian, original texts and a large document in the form of a bachelor thesis.

The first testing document is based on an article in Latvian language from Wikipedia (Wikipedia, 2016) that includes information about robotics history and components (types of muscles, engines etc.). The article is divided into three approximately equal parts - Chapter 1, Chapter 2 and Chapter 3. In Chapter 1, a copy-paste text from Wikipedia article is kept without changing anything excluding text formatting. Both Chapter 2 and Chapter 3 contain a text where one to two words were replaced in each sentence with their synonyms without changing the word order in sentences. Besides word replacing with synonyms, the order of words is changed in each sentence of Chapter 3.

The second testing document uses an article in Latvian which includes an interview about modern technologies in the gambling industry from a well-known web source in Latvia - *Kursors.lv* (Skutelis, 2018). The article is used for the research needs in agreement with the author of the mentioned article. The whole article text is copied to the second document, as well as divided into three similar parts and formatted as Chapter 1, Chapter 2 and Chapter

3 similarly to the first document.

The third document is based on the article on plagiarism detection in English from Wikipedia (Wikipedia, 2018a). The text from the article is copied from its source also including pictures and divided in two large parts. The first part is translated from English to Latvian using “Google Translate” and it is copied to the second document without any changes. The second part of the article is human translated by the author of this paper. The tables which are available in the Wikipedia article are also translated.

Four original and short stories were used for the creation of the fourth document. The main criterion for selecting the stories was a fact that the text was not either published on the Internet or located in “Google Docs” or other online text editor. The document is created with the aim of checking how the text-matching software systems react to the testing documents that do not contain the text from any Internet source.

Usually text-matching software systems have difficulties in revealing if there is translated plagiarisms in the given text. It is more complicated to detect and check if the translations are from a language containing different alphabet with another way of writing the letters, for example, Cyrillic script in Russian. It should be noted that there many people in Latvia who know Russian fluently or Russian is their native language. For these people it can be easier to look for information sources in Russian rather than English, Latvian or other languages. Afterwards this information can be used for their final theses or other works as well.

This is a reason for creating an additional testing document for testing text-matching software systems - the fifth testing document using the text from Wikipedia article in Russian that contains information about the population on the Earth (Wikipedia, 2018b). Similar to the third testing document the selected text from Wikipedia is copied together with tables and pasted into the document. The whole text is divided into two parts: the first part is translated from Russian to Latvian using “Google Translate” without any corrections or changes. The second part of the divided text is human translated by the author of this paper as it was done when preparing the third testing document. The tables which are available in the Wikipedia article are translated as well.

To check if text-matching software systems are able to process large documents, the sixth document was added - a bachelor thesis with a permission of its author. It contains 10064 words in total.

During the research, 16 text-matching software systems were used to check their performance on the set of documents in Latvian language. The paper presents the detailed description of the document set prepared for testing, the research methodology and testing results showing plagiarism coverage. The authors of the paper also make worthwhile and useful conclusions for text-matching software developers, universities, schools and other educational institutions and their representatives about suitability of the known text-matching software systems for Latvian academic environment.

Keywords: text-matching software, plagiarism detection, academic integrity, software testing.

References

Kravjar, J. (2018). Nationwide Barrier to Plagiarism is Bearing Fruit. Presentation at the conference "Building a Culture of Academic Integrity in Education", October 17 of 2018, Riga, Latvia.

Wikipedia. (2016). Robotika. Available at https://lv.wikipedia.org/w/index.php?title=Robotika&oldid=2545152.

Skutelis K. (2018). Tomass Aleksandersons: Modernās tehnoloģijas azartspēļu industrijā (in Latvian). Available at https://kursors.lv/2018/07/31/tomass-aleksandersons-modernas-tehnologijas-azartspelu-industrija/

Wikipedia. (2018a). Plagiarism detection. Available at https://en.wikipedia.org/w/index.php?title=Plagiarism_detection&oldid=860449242.

Wikipedia. (2018b). Население Земли (in Russian). Available at https://ru.wikipedia.org/wiki/%D0%9D%D0%B0%D1%81%D0%B5%D0%BB%D0%B5%D0%BD%D0%B8%D0%B5_%D0%97%D0%B5%D0%BC%D0%BB%D0%B8.