The unified anti-plagiarism system in Poland

Andrzej KURKIEWICZ
Ministry of Science and Higher Education, Poland

The Unified Antiplagiarism System (JSA) is a plagiarism-protection tool all Polish higher education institutions (HEIs) and research institutes are obliged to use. JSA checks theses, before they are defended, against the National Repository of Written Theses, Polish-language Internet resources (NEKST), Wikipedias in the most popular languages and databases of legal acts. The total volume of data used by the System to determine the originality of a thesis is now over 30 terabytes.

The System, built in 2017–2018 at the National Information Processing Institute, encompasses all areas of teaching and is free of charge for all HEIs and research entities. It has been available for checking theses since January 2019, while verifying doctoral dissertations will commence in October 2019.

Poland has opted for a publicly devised and funded antiplagiarism system so that universities and research institutions have free access to a high-quality tool whose performance is consistent across the country. Its development involved an open competition for algorithms and source codes of different solutions so that its architecture reflects the cutting edge in plagiarism detection technology.

We tested our methodology and tools during the afore-mentioned competition in 2017. In addition, using some reference theses, we compared JSA with such commercial systems as Plagiat.pl or Genuino in order to check the detection quality differences. We did not evaluate the System against Turnitin or other international plagiarism detection tools.

The System's operation can be roughly divided into four stages:

1. statistics
2. stylometry
3. identification of documents (say, on the Internet or in the Repository) from which text fragments may have been lifted into the checked thesis or dissertation
4. for selected documents found in stage 3 – more detailed identification of shared or similar text passages between them and the writing under investigation.

Re 1: basic statistical data are collected (regarding the number of words, characters, unrecognised words, special characters or characters from another language, distribution of word length) and compared with averages and distributions found in the Repository.

Re 2: internal analysis – detection, within a single document, of fragments possibly written by another author. Based on the entirety of the text being analysed, the System highlights passages attributable to another person or persons (assuming that the text has a principal author). The objective here is to determine the stylistic profile of the principal author and highlight blocs of text exhibiting stylistic features inconsistent with that profile.

Re 3: external analysis – singling out from the reference databases source documents, that is texts whose fragments appear to have been used in the thesis under investigation. Given that the set of source documents (or a reference corpus) is very large and thus searching it would be too time-consuming, we have developed a number of indices allowing us to build computing clusters.

Re 4: detection of passages (phrases / sentences / paragraphs) shared by two documents, the reference one and the one being checked for originality. JSA does these paired comparisons looking for four different kinds of plagiarism, in different variants:

- copy-paste,
- copy-paste + word order alteration,
- copy-paste + synonym substitution,
- copy-paste + synonym substitution + word order alteration.

During the first 4 months of stable-release operation, JSA has analysed 55 thousand theses, the median time of a single thesis analysis is about 5 minutes, it has almost 50 thousand active users, and 271 active universities. Infrastructure-wise, it has used almost 100 middle-class servers.

Keywords: JSA, anti-plagiarism, system, public, free, high-quality.