ANTON SUMMARY

Jiří Janoušek

Abstract: The presentation is aimed at AntOn solution which was created as a key output of IPPHAEA project. The AntOn stands for "Antiplagiator Online" and the solution was developed by IS4U company. AntOn is extending the portfolio of tools and solutions offered by IS4U company's main product UIS—the complex information system. The UIS system provides administrative support to schools and there assists in many key activities—e.g. administrative, study and research processes support. One of the key activities is the research and working with student texts, and that is why the company has decided to create AntOn when asked to participate at IPPHAEA project. The AntOn solution was successfully developed and launched during the project and it was tested together with UIS system in production.

The AntOn solution is a special software component executed at a standalone server. It is intended to accept any texts (documents) in various formats (e.g. DOC, DOCX, PDF, HTML, XML, ODT) to analyze them and to report back about texts' level of similarity to the other texts in the AntOn internal document database. The REST interface is implemented to connect AntOn to any system as a service—e.g. REST connection is used for cooperation with UIS, too.

The AntOn itself works in 4 steps. Firstly, the plain text is extracted from the input document. Secondly, language-dependent stop lists and special language settings are applied. Stop list is a group of extra words, fillings or frequently used words, numbers, special characters etc. These words have no importance for the text processing and they can be removed. Special language settings ensures the most appropriate output for certain language. Next step includes creating special five-words pieces called chunks. A chunk is context-dependent surroundings of the word, so it needs to be created for almost every word (exactly it is $n - 4$ chunks for n words) in the processed text. Finally, the set of chunks for processed document is compared to other chunk sets in the AntOn document database. This is computationally very demanding operation, so there is a special approximation algorithm implemented. It is vital to have good hardware capacity to achieve acceptable speed and results.

The output of the AntOn system is level of similarity of processed document to other documents in AntOn's database. High precisious SDIFF algorithm can be used to show differences between two similar documents when the level goes over the given limit (now the limit is set to 50%). The final decision about plagiarism, however, needs to be made by human inspector, the AntOn system provides only the level of similarity.

To conclude, it is necessary for the school to have any antiplagiator solution because just the existence of the antiplagiator prevents students from cheating. When some survey in existing antiplagiators has been done it was found out that the quality of solution is not crucial. The school is satisfied and is not forced to increase the accuracy of the system. Therefore, the IS4U Company implemented AntOn system at full stable version and due to this no further development is considered now.

More information can be given at `info@is4u.cz`.

Author

Jiří Janoušek, `jiri.janousek@is4u.cz`, IS4U, s. r. o., U vodárny 2a, Brno, Czech Republic