Institute for Systems Analysis
Federal Research Center «Computer Science and Control»
of the Russian Academy of Sciences

The Paraplag: Russian Dataset for Paraphrased Plagiarism Detection

Zubarev D.V. – PhD student
Sochenkov I.V. – PhD

Paraplag

- The text dataset in Russian to evaluate and compare quality metrics of different plagiarism detection approaches

- ParaPlag is open and available on the Web
  https://plagevalrus.github.io

Outline

- The source collection - 5.7 millions texts
- Essays with reused text from some documents in the source collection
- The essays are divided into two groups by a different level of difficulty

Text rewrite techniques

- DEL – delete some words
- ADD – add some words
- LPR (Light Paraphrase) - change word forms (number, case, etc.);
- SHF (shuffling) – change the order of words or clauses
- CCT (concatenation) – concatenate two or more original sentences
- SEP (sentence splitting) – split the original sentence into two or more sentences.
- SYN (synonymizing) – replace some words or phrases of the original sentence with synonyms
- HPR (Heavy Paraphrase) – heavy rewrite of the original sentence