Mahak Samim

A Corpus of Persian Academic Texts for Evaluating Plagiarism Detection Systems

مرکز تحقیقات کامپیوتری علوم اسلامی

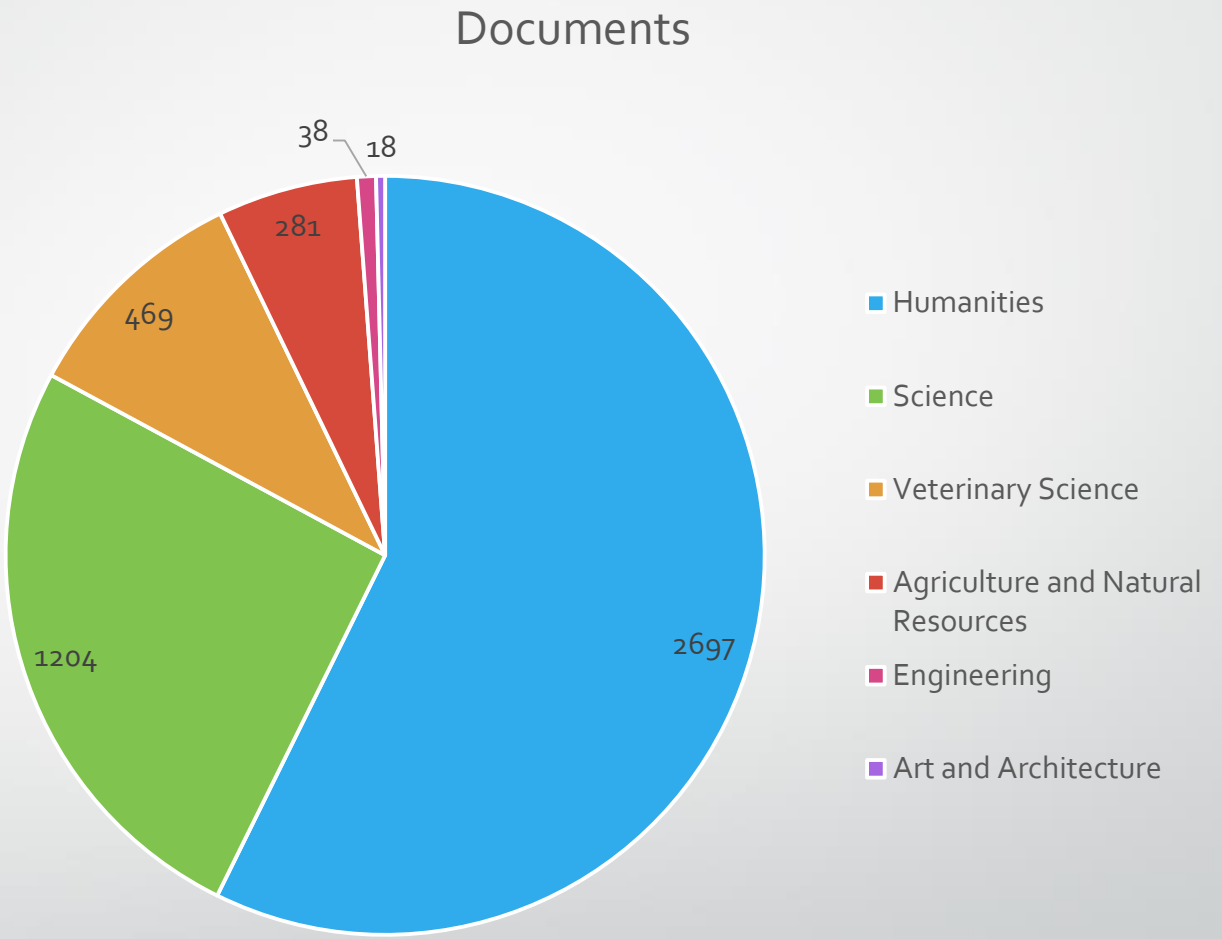
Document Collection

Documents Source

We crawled the websites of journals

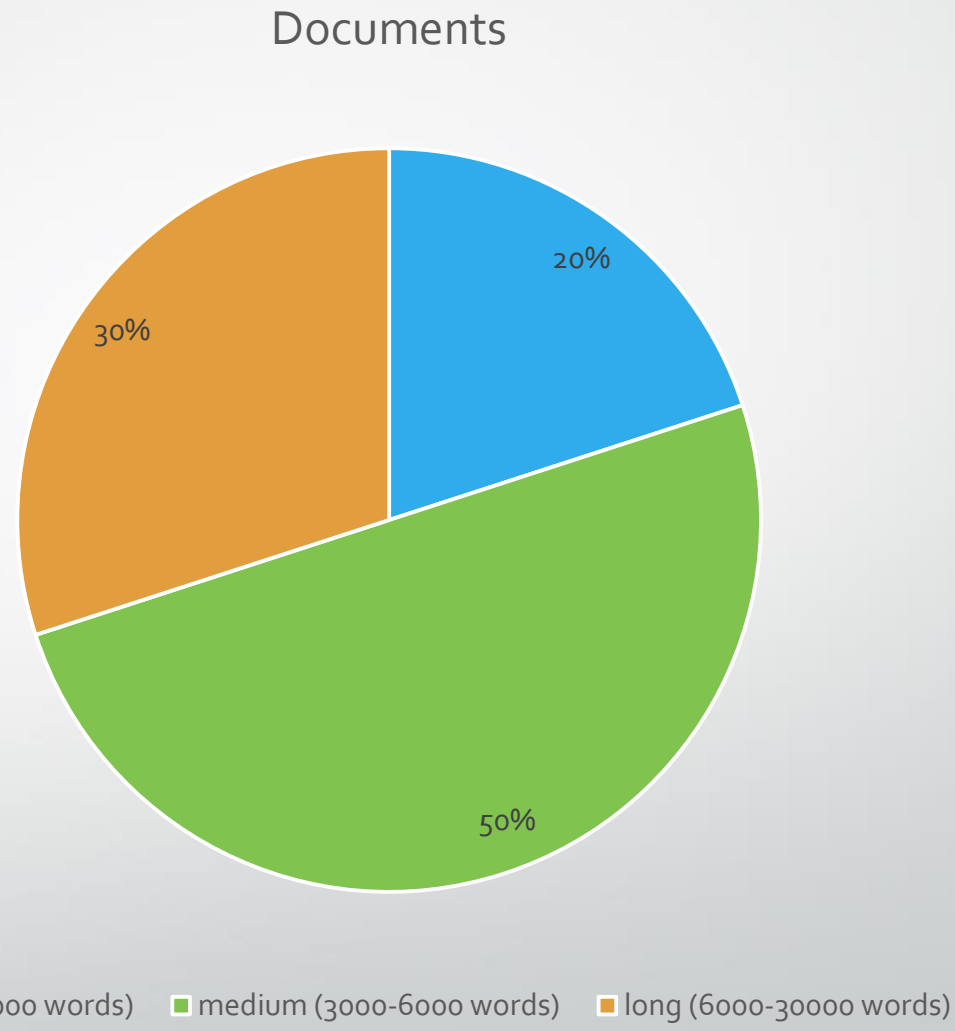
Subject distribution

Statistics of number of documents per subject



Length distribution

Statistics of document lengths



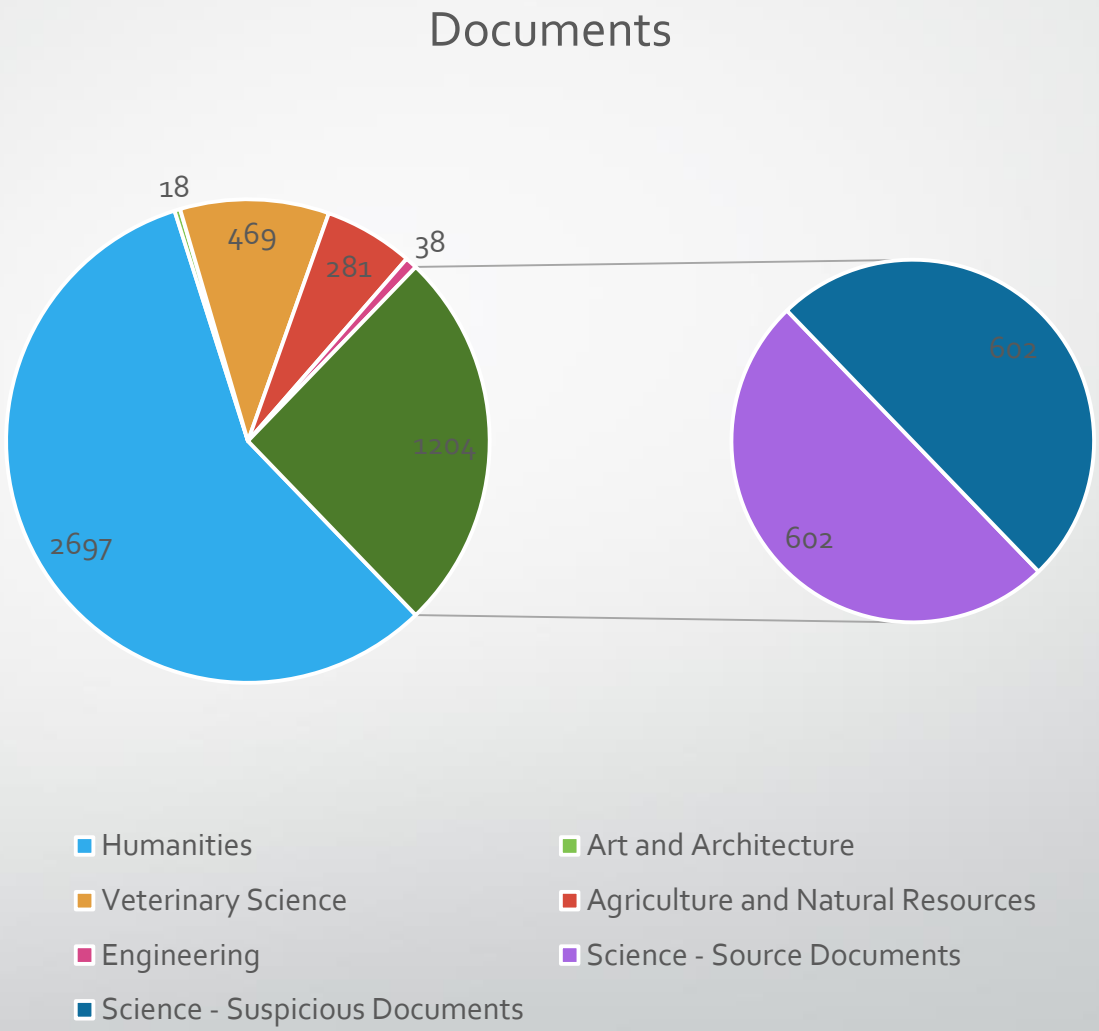
Source / suspicious documents

Source / suspicious distribution

We follow PANs tradition of using half of the documents as source documents and half as suspicious documents.

Subject distribution in source / suspicious

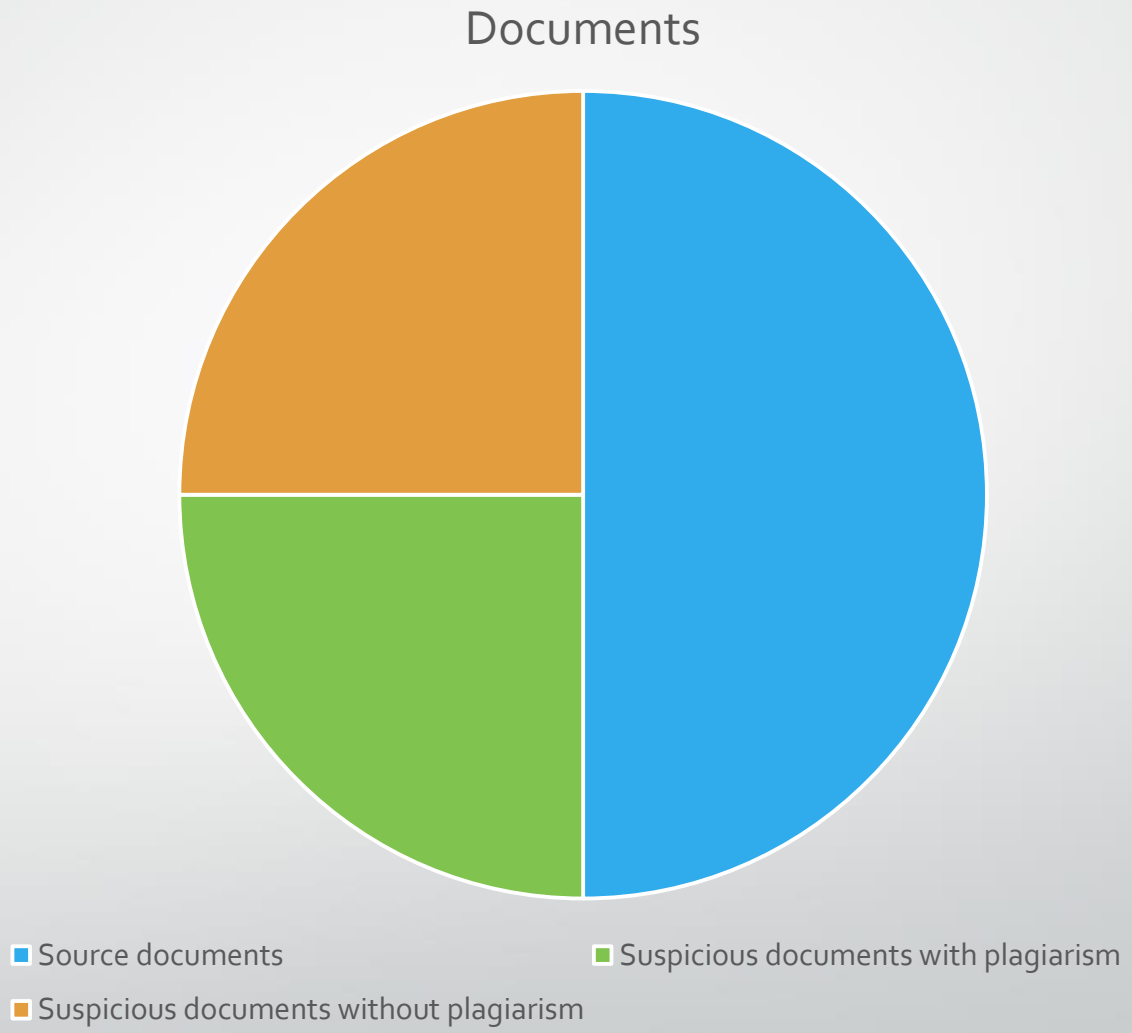
Subjects of the papers were taken into consideration while dividing the collection into halves.



Plagiarism per document

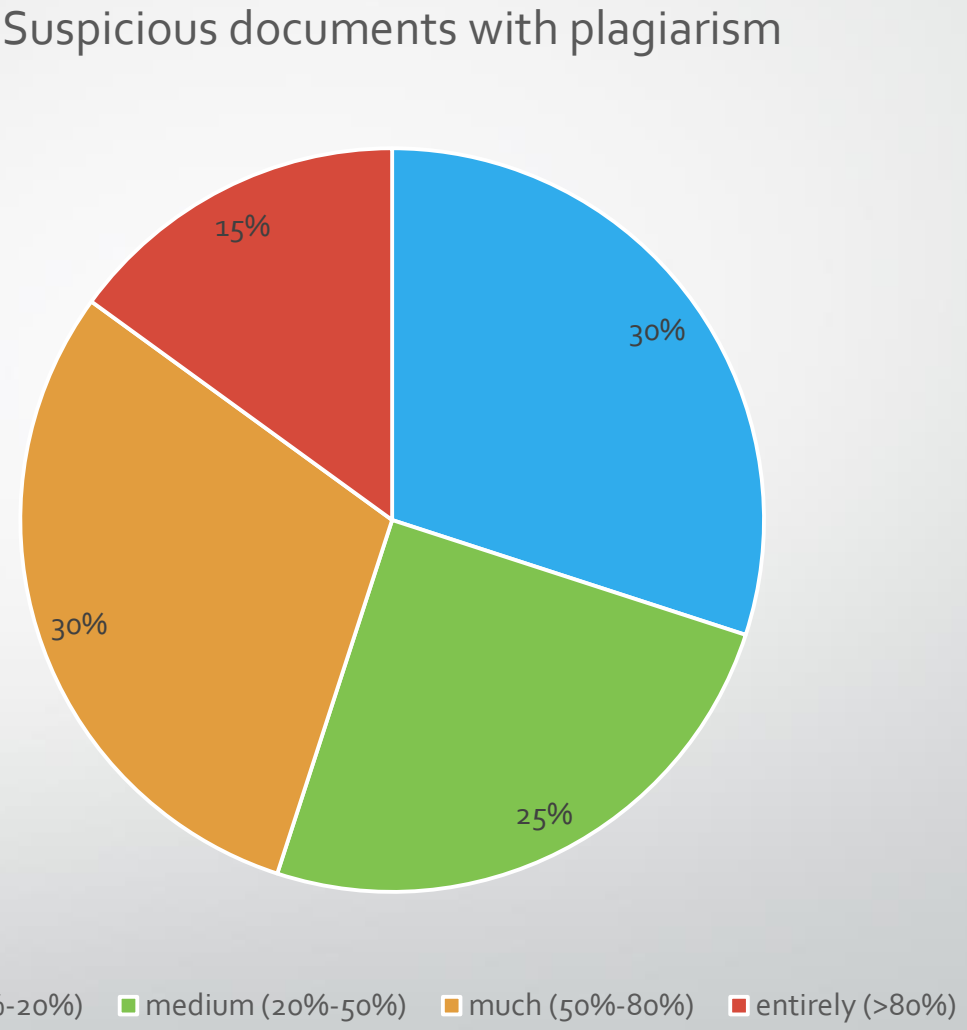
Suspicious documents without plagiarism

The documents without plagiarism allow to determine whether or not a detector can distinguish plagiarism cases from overlaps that occur naturally between random documents.



Suspicious documents with plagiarism

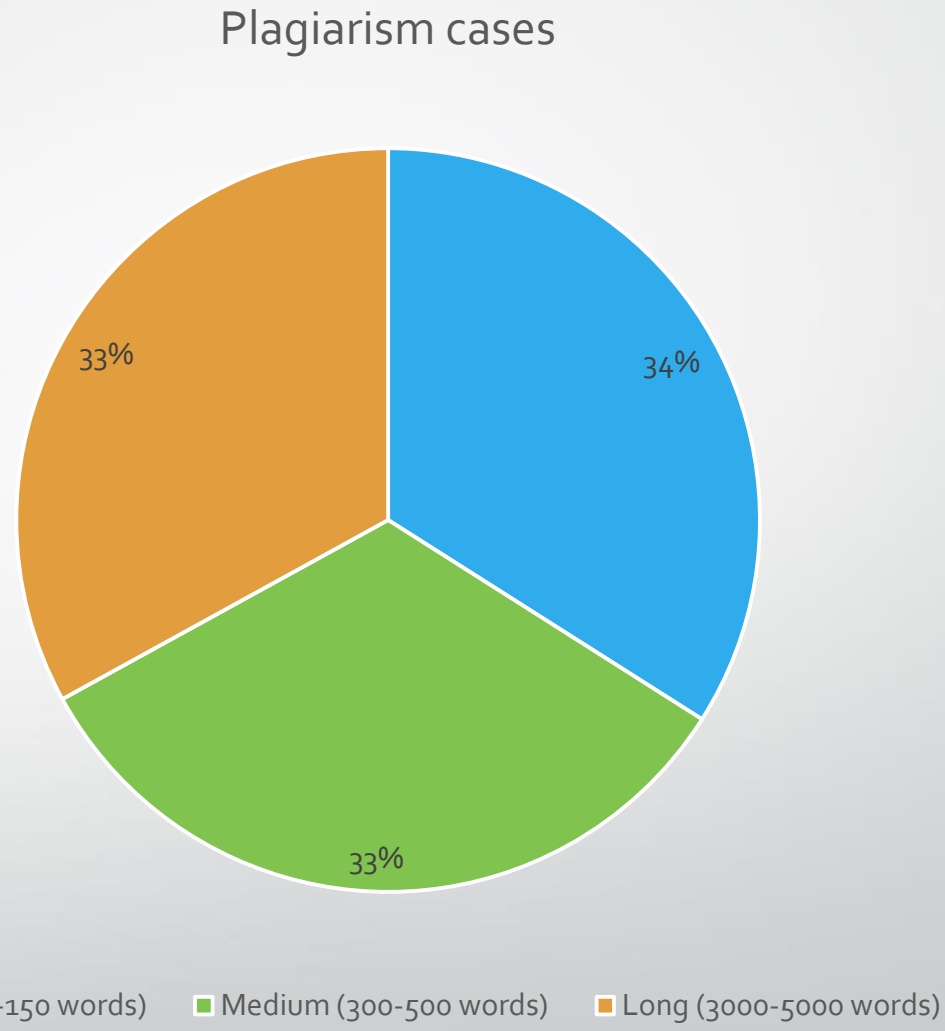
Statistics of plagiarism per document in the suspicious documents with plagiarism, i.e. 25 percent of the whole corpus.



Plagiarism case length

Statistics of lengths of plagiarism cases

Our corpus consists of a total of 5862 plagiarism cases with lengths between 50 and 5000 words. Long plagiarism cases may include more than one sentence.



Topic match

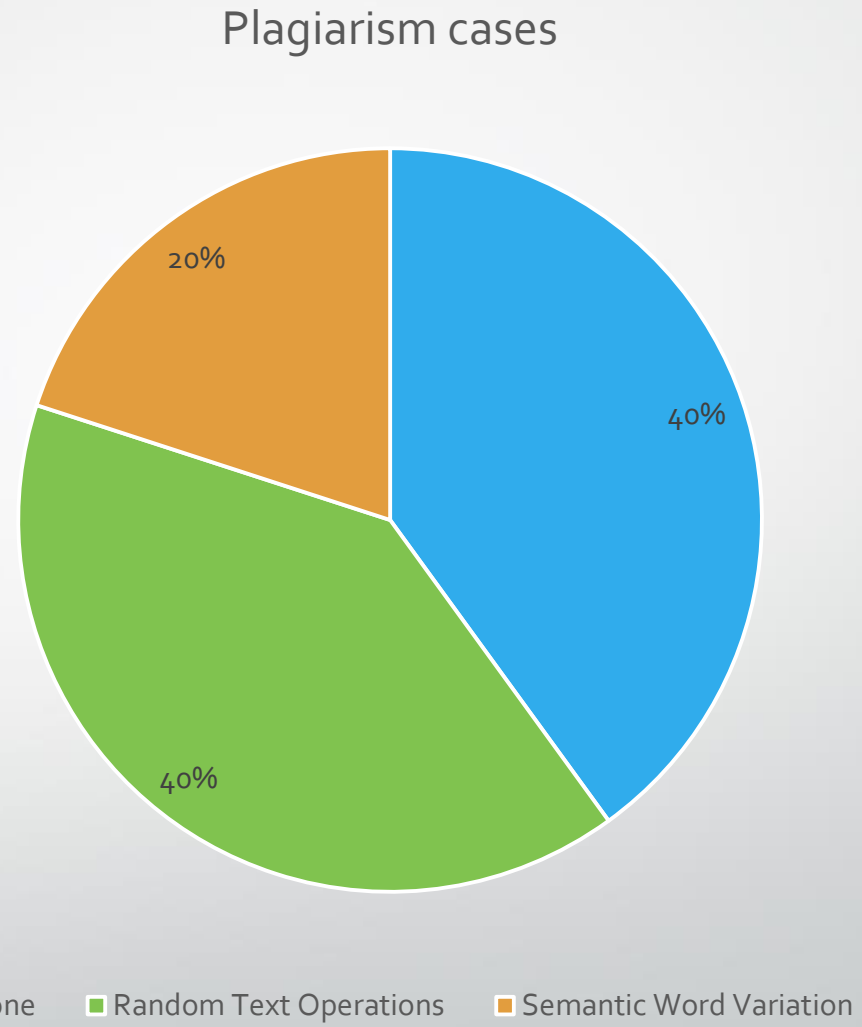
Intra-topic & inter-topic cases

Fifty percent of the plagiarism cases were made between papers with same topics (intra-topic cases) and fifty percent between papers with different topics (inter-topic cases).

Obfuscation types

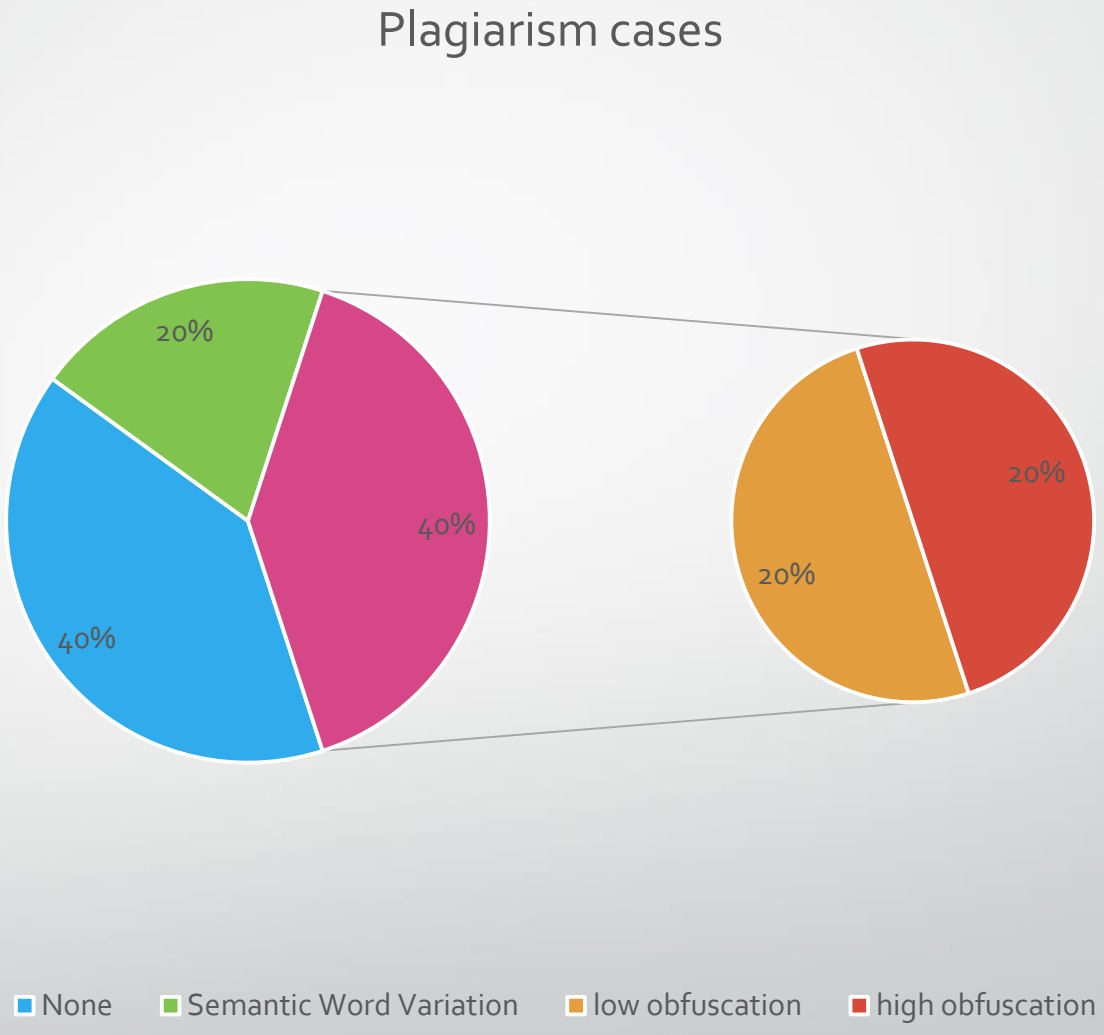
Types of obfuscation in plagiarism cases

40 percent of the plagiarism cases have no obfuscation. These cases are especially appropriate for evaluating intrinsic plagiarism detection.



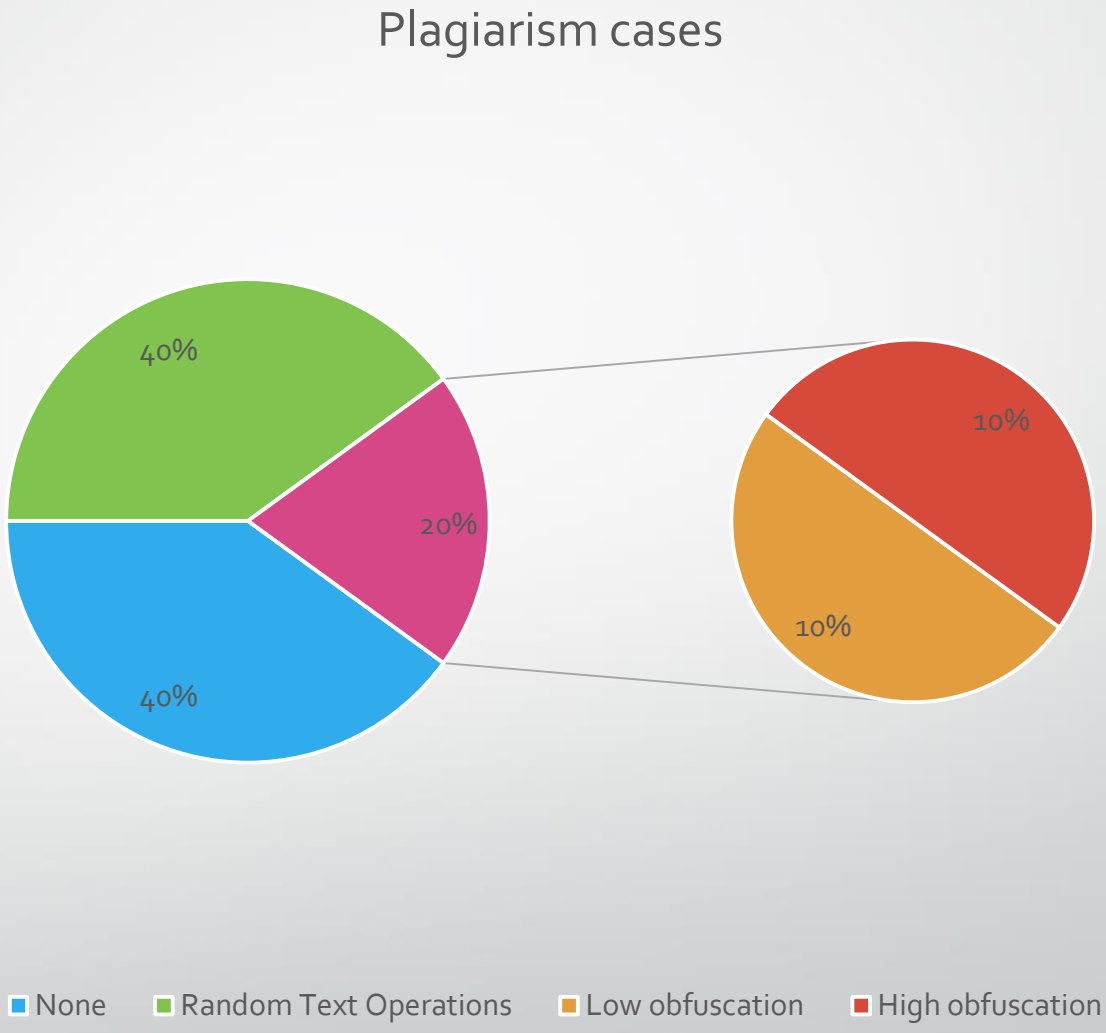
Random Text Operations

Random text operations are operations such as adding, deleting and substituting words, which are all done randomly.



Semantic Word Variation

Semantic word variation, is the random substitution of words with their synonyms.



SUMMARY

- Mahak Samim is a plagiarism corpus which can be used for evaluating both intrinsic and external plagiarism detection systems.
- In order to preserve overall balance, many factors – plagiarism per document, plagiarism case length, topic match, obfuscation type, and obfuscation degree – were taken into consideration while preparing each plagiarism case.