The aim of the project is to develop methods of automated subject area classification of natural language documents (i.e. papers) based on document contents analysis done by a computer algorithm.
Improvements in the currently developed methods of solving this problem are proposed. The most significant concerns improvement of document recognition of subject areas other than those that were previously used in the learning process (the existing approaches tend to assign such documents to the nearest of the learned classes, even if the document is not similar to it).

From the methodological point of view realization of the project goals needs the development of classification methods based on high-dimensional data. Characteristics of the text documents determined by computer algorithms in automatic analysis tend to have a very high dimensionality. Moreover, the classifier should not only assign the recognized document to the nearest class known to the system, but should check if the document is sufficiently similar in subject to the selected class. In this way, the algorithm could classify as "other"/"unrecognized" documents that thematically are far away from the classes known to the system (this property is known as “open-set” classification).

The proposed project tasks will be performed for documents in Polish and English. The project, in addition to the research and methodological goals, has the practical value. Improvement in methods of automated subject recognition of documents could simplify the process of document retrieval/ information search in the growing in size repositories of documents (e.g. in the collections of newspaper articles, scientific papers or open Internet repositories).