ABSTRACT

The thesis is presented in 55 pages. It contains 2 appendixes and bibliography of 14 references. Five figures and 2 tables are given in the thesis.

The goal of the thesis is to create a system for analyzing air crashes data to identify their causes in large datasets since 1919 year.

In the thesis, existing software solutions for processing text information and clustering methods are analyzed, such as hierarchical clustering, clustering of k-means, expectation-maximization method (EM method) and DBSCAN method. They are compared in terms of their speed, efficiency, type of model, the ability to process a large amount of data and the complexity of the algorithms. In the thesis, the method of clustering of k-means is chosen. To convert text descriptions into vectors, the method of Bag-of-words using the TF-IDF measure was chosen.

An automated system that implements selected methods is developed. The testing of the developed system is carried out.

Keywords: text descriptions of air crashes, text vectorization, document body, clustering, cluster annotation.