ABSTRACT

The thesis is presented in 45 pages. It contains 2 appendixes and bibliography of 26 references. Three figures and 6 tables are given in the thesis.

The goal of the thesis is to develop mathematical and software tools for solving the problem of building specific dictionaries for recognition of speech signals for various subjects.

In the thesis, existing systems Continuous Speech Recognition and related dictionaries are analyzed. Was done comparing them in terms of the accuracy of the resulting solutions, value, purpose and scope of the dictionary.

Formed requirements for normalization of texts. For each subject creating corresponding dictionaries. Developed an automated system that builds thematic dictionaries. Tests developed system.

Keywords: text corpus, Files keywords, normalization, Lemmatisation, linguistic model, TFIDF metric, clustering.