Brief information about the project

Name of the project	AP19677835 «Research of models and development of an intelligent question-answer system based on semantic approaches for the state language in the field of legislation of the Republic of Kazakhstan» (0123PK00861)
Relevance	Modern systems and developed technologies in the field of natural language processing - NLP and artificial intelligence (AI) surprise with their achievements and speed of development and implementation in various professional fields of activity and in everyday human life. It is also necessary to take into account that behind the results obtained and achievements in NLP there is a lot of scientific work and research, technological solutions and opportunities. One of the most relevant and popular areas of NLP is dialogue and question-answer systems. These systems include a large number of subtasks, the results of which can be applied in various areas of artificial intelligence such as speech technologies, machine translation, search engines, etc. There are various scientific approaches and methods for solving question-answering systems. Some of them will be presented below. Software offered by various manufacturers, such as Currently, there are quite a large number of implementations of question-answering systems. The implementation of the question-answer system START deserves attention. Other interesting systems that claim open source status are the Cyc, OpenEphyra and PIQUANT systems. Also, researchers of the Semantic Web have proposed using semantic search technologies for the English language, such as OntoSearch, Semantic Wikis, Semantic Portals, multi-agent P2P semantic system of routes (queries) based on ontology, question-answer systems based on ontology. And also the well-known smart station “Alice” - a virtual voice assistant created by Yandex. Recognizes natural speech, simulates live dialogue, provides answers to user questions and, thanks to programmed skills, solves applied problems. Currently, various chatbot systems are used in the Kazakh language in the egov state portal and various banking systems. But unfortunately, they are limited by subject matter and response protocols, and also do not support voice functionality. Unfortunately, there are no analogues of these systems (open access and/or paid resources) for the Kazakh language. The above software products were developed for many resource (European, Slavic groups) languages such as English, Spanish, Russian, etc. Unfortunately, for

	Turkic languages (Kazakh, Kyrgyz, Turkish, Uzbek, etc.) there is currently no software available in the public domain implementation. The presented works and systems cannot be applied partially or completely to the Kazakh language. This is due to the fact that for each language separate resources are collected and processed, taking into account linguistic properties. Also, each language has its own dialect, semantic (cognitive) concepts, etc.
Purpose	The goal of the project is to study the problem and develop question-answering technology (algorithms, models, systems) in the Kazakh language in the field of legislation of the Republic of Kazakhstan, based on modern semantic approaches and neural network learning.
Objectives	The main objectives of the project are: • Development of an approach for collecting and processing text data from open electronic sources; • Development of a question-answer matrix for interaction; • Development of a method for synthesizing an answer in Kazakh, taking into account semantics based on machine learning. • Development of a method for predicting responses by building a classification model and neural network training • Development and implementation of a module for supporting the system's voice interface in the Kazakh language
Expected and achieved results	Expected results of the project: research of the problem and development of a prototype of intelligent question-answering technology (algorithms, models, systems) based on neural network learning with the ability to conduct dialogue in the Kazakh language. It is planned to publish articles in publications indexed in the Science Citation Index Expanded of the Web of Science database and (or) have a CiteScore percentile in the Scopus database of at least 35 (thirty-five) and publication) of the article in a peer-reviewed foreign or domestic publication recommended by RK. Work completed and results obtained for 2023: - An analytical review of the research topic was carried out. Modern foreign analogues of question-answer systems, chatbots and web portals related to legal and regulatory documents were researched and analyzed. The properties of ChatGPT and its analogs are described, such as: "Alphachat", "Law ChatGPT", "DoNotPay",

	"LawDroid", "Ross Intelligence", "CaseMine". Their differences and the applied methods of supervised learning and reinforcement learning are described. - The types of regulatory documents have been studied, and they have been classified into basic and derivative types, depending on the development and procedure for the adoption of administrative and legal acts of the Republic of Kazakhstan. ;- The hierarchy of legislative documents of the Republic of Kazakhstan is described. Based on the hierarchy, data is collected from legislative documents to form a database. - The structure of the Kazakh language of legislative documents is determined on the basis of formal grammars. For this purpose, an analysis of texts and sentences of various types from legal documents in the Kazakh language was carried out. - Data was collected from articles on legislative websites of the Republic of Kazakhstan. The main source of information is the website https://adilet.zan.kz/kaz/. The data was uploaded in html format and then converted to text format (txt). To collect data, we used a universal multi-threaded website parser for Windows with a large number of functions and the ability to fully control and configure all stages of working with WEB content - Content downloader.
Research team members with their identifiers (Scopus Author ID, Researcher ID, ORCID, if available) and links to relevant profiles	Project leader: PhD. Rakhimova Diana Ramazanovna h-index: 4, Scopus author ID: 55682794500 Web of Science ResearcherID D-8421-2012 Research team members: 1. Doctor of Technical Sciences, prof. Tukeyev Ualsher Anuarbekovich h-index: 5. Scopus author ID 55701639900 https://orcid.org/0000-0001-9878-981X 2. PhD Shormakova Assem h-index:2, ORCID ID: 0000-0002-1637-4643, Author ID Scopus: 55786942400, Researcher ID Web of Science: D-5836-2015 3. PhD Karibayeva Aidana h-index: 3 orcid id: 0000-0002-2023-1573, Scopus id: 57196004542, ResearcherID: AAR-4134-2020 4. PhD Karyukin Vladislav, h-1, https://orcid.org/0000-0002-8768-0349, Scopus Author ID: 57218952479 SciProfiles: 2410440 5. Master Turarbek Assem h-1, Scopus id 57031944900, https://orcid.org/0000-0002-4793-0446

	6 Master Amirova Dina h-index: 1, ResearcherID: P-5668-2017, Orcid id: 0000-0002-0728-905X, Scopus ID: 57196009653
List of publications with links to them	List of publications 2023: 1.Ualsher Tukeyev and etc. Kazakh-Tatar Machine Translation on the Base of Complete Set of Endings Model. Chapter of the book «Machine Learning». Editors: Burcu Arsan Ph.D(c) | Prof. Dr. Natalya Ketenci ©Yeditepe University Press, pp. 73-90. ISBN: 978-975-307-139-0 Istanbul, 2023 2.Diana Rakhimova, Yntymak Abdrazakh. Study and development of an approach for identifying incorrect words for the Kazakh language in semi-structured data. Chapter of the book «Machine Learning». Editors: Burcu Arsan Ph.D(c) | Prof. Dr. Natalya Ketenci ©Yeditepe University Press, pp. 23-39. ISBN: 978-975-307-139-0 Istanbul, 2023. 3.Тукеев У.А. Реляционные модели обработки тюркских языков. VIII — Международную научно-практическую конференцию «Информатика и прикладная математика» с 26 по 27 октября 2023 г. Алматы. 85-89 с.
Patents	-