В последнее время общее количество обращений клиентов банков по продуктам очень сильно возросло. В банковской среде возникает потребность в решении задачи *поиска корневой причины* обращений. В настоящее время эта задача решается вручную или полуавтоматически.

Целью этой работы ставится создание полностью автоматической системы, так как это приведет к уменьшению расходов на штат и увеличит качество предоставления услуг.

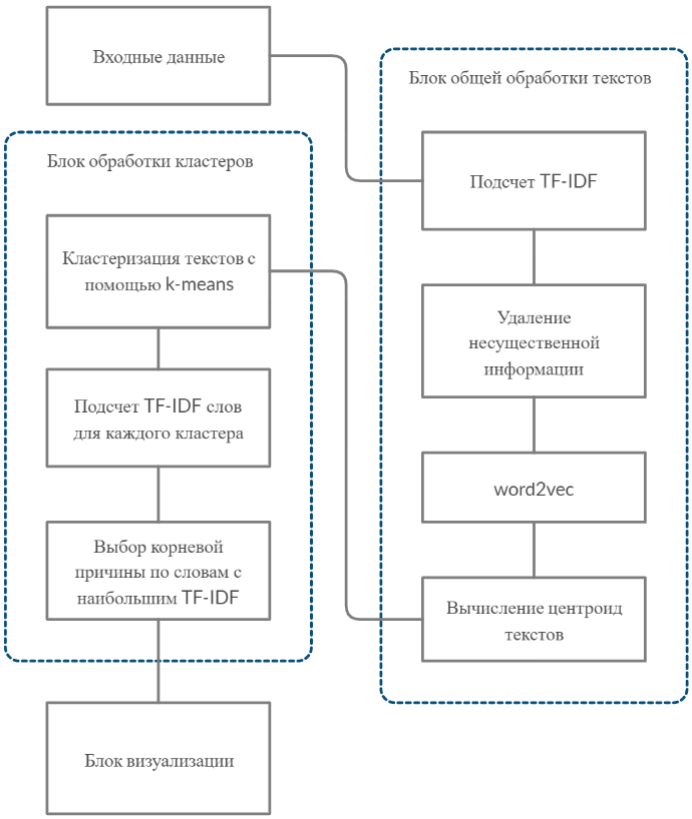

Рис. 1. Функциональная схема работы

Автором предложен следующий алгоритм решения задачи: для начала нужно произвести нормализацию исходных данных, то есть приведения текста обращения в нормальную (словарную) форму. Это нужно для того, чтобы одно слово в разных словарных формах (например, падежах) считалось системой за одну и ту же смысловую единицу. Далее считается их характеристика TF-IDF, далее удаляются слова с самым низким и самым высоким показателем TF-IDF, так как установлено, что таковые очень часто являются несущественными. Далее слова преобразуются в вектора с помощью word2vec, эти векторы будут отображать эмбеддинги (смыслы) слов. После этого по векторам вычисляется центроиды текстов, к которым применяется алгоритм k-means для кластеризации по корневым причинам. Для каждого кластера снова считается TF-IDF слов. Несколько (2-5) слов с наибольшим TF-IDF выбираются в качестве корневой причины. Также предусмотрена визуализация для мониторинга основных корневых причин.

Алгоритм реализован на языке программирования Python3, с помощью модулей Pymorphy2, nltk и gensim. В данный момент разработка находится на стадии тестирования и получения численных оценок качества работы.