Автоматическая классификация обращений граждан

А.Р. Мангушева[1], А.Г. Кварацхелия[2], Д.Ф. Рахимов[3], К.А. Григорян[4]

[1]*Казанский национальный исследовательский технологический университет, Карла Маркса 68, Казань, Россия, 420015*
[2]*АО «Барс Груп», Некрасова, 9, Казань, Россия, 420012*
[3]*ГКУ «Центр цифровой трансформации Республики Татарстан», Петербургская 56, Казань, Россия, 420074*
[4]*Казанский федеральный университет, Кремлёвская, 18, Казань, Россия, 420008*

Аннотация

В данной статье приведены результаты работы над сервисом, позволяющем классифицировать обращение граждан. Обращения граждан представляют собой неструктурированный текст, написанный на естественном языке. Задача состоит в том, чтобы по тексту определить в какой исполнительный орган должна быть отнесена заявка. На основе документов текстового набора строится матрица расстояний с использованием алгоритмов word2vec и Word Mover's Distance. Для каждого нового документа рассчитывается вектор расстояний от данного документа до каждого документа из обучающего набора. Применение классификатора K-ближайших соседей с параметром количества соседей, равным 10-и. позволяет определить тематику и исполнителя обращения.

Ключевые слова

Машинное обучение, обработка естественного языка, токинайзер, анализ данных

1. Введение

Научная новизна и практическая значимость настоящей разработки заключается в комплексном подходе к извлечению данных из текстов, составленных на естественном языке, в том числе с применением подходов машинного обучения, статистического и лингвистического анализа данных. Естественный язык (ЕЯ) — сверхсложная семиотическая система, состоящая из неограниченного числа подсистем, каждая из которых конечна, а потому формализуема, при этом сам язык — незамкнутая система, которая не может быть формализована до конца [3].

Согласно [4], обработка естественного языка (Natural Language Processing, NLP) — общее направление искусственного интеллекта и математической лингвистики. На текущий момент достаточно много достижений в рамках данного направления. Яркими примерами являются различные голосовые помощники, автоматические переводчики, чат-боты. В рамках NLP существует достаточно много методы с данными, причем для каждой задачи выбирается свой способ решения. Существуют подходы, основанные на правилах, где необходимо определить наборы языковых правил для описания данных. В рамках данного исследования, описание правил является достаточно проблематичным способом решения поставленной задачи. Тексты, находящиеся в обращениях граждан совершенно не структурированы. использование правил может быть применено локально для поиска конкретных слов, но данный способ не даст необходимой скорости обработки данных. Обработать большой массив данных согласно правилам - задача слишком затратная по времени. Следует понимать, что самой важной частью при обработке текста на естественном языке являются сами данные. Их изучение требует достаточно скрупулезного подхода. Специфика данных также не облегчает эту задачу.

2. Описание решемой задачи

Существует множество разрозненных источников поступления обратной связи от населения жалобы, предложения . Заявки поступают в различном формате в различные инстанции,

которые, в свою очередь, не всегда компетентны по поступающему вопросу. Поступающая информация часто теряется, что ведет к нерешению проблем, росту однотипных повторяющихся заявок от населения и, как следствие, к повышению неудовлетворенности населения. Целью реализация сервиса с использованием методов машинного обучения, которая на основании содержания текста обращения могла бы определять тему обращения и ответственный орган власти является сокращение потери заявок, поступающих от граждан.

3. Методы и инструменты реализации сервиса

Пользователь вводит текст обращения. По введенным данным автоматически определяется поле theme, которое содержит ключевые факты из обращения. Затем, на основании данных фактов, генерируется поле category - предложение по отнесению обращения к определенной категории из справочника. После этого последним этапом работы системы является генерация поля executor - предполагаемый исполнитель, которому должно быть адресовано обращение. Сервис выводят предоставляет информацию о первых трех возможных варианта тема, категории и исполнителей с сортировкой в порядке убывания вероятности совпадения.

Для решения данной задачи были использованы такие алгоритмы, как word2vec [1] и Word Mover's Distance (WMD) [2]. Word2vec в качестве входных данных принимает большой текстовый корпус и некоторые гиперпараметры. Далее названия гиперпараметров и их дефолтные значения будут указаны в соответсвии с реализацией Word2Vec на open-source платформе Gensim. Модель реализована на Python с использованием библиотек NumPy, SciPy. Алгоритм каждому слову сопоставляет вектор. Причем близкие слова соответствуют близким векторам. Мерой близости слов выступает их контекстная близость: близкие слова встречаются в тексте рядом с одинаковыми словами. А расстоянием между векторами измеряется при помощи косинусного сходства (cosine similarity). Обучая нейронные сети, Word2Vec максимизирует косинусную меру близости между векторами слов, которые встречаются в похожих контекстах и минимизирует косинусную меру между словами, которые не встречаются рядом. На выход Word2Vec передает координаты векторов, соответствующих данным словам.

Для того, чтобы классифицировать обращения был использован следующий алгоритм:

а) вычисляется попарное WMD-расстояние от текста обращения до каждого текста из корпуса модели;

б) применяется алгоритм k-ближайших соседей [7] для вычисления вероятностей принадлежности к классам категорий, исполнителей, тем.

Также были использованы такие библиотеки и компоненты, как: NTLK [6], Gensim, RNNmorph (TensorFlow), Scikit-learn, семантические модели русского языка [5].

Данное решение было внедрено в несколько регионов РФ.

4. Литература

[1] Mikolov, T. Efficient estimation of word representations in vector space // ArXiv preprint: 1301.3781. – 2013.
[2] Kusner, M. From word embeddings to document distances // International conference on machine learning. PMLR. – 2015. – P. 957-966.
[3] Воронина, И.Е. Компьютерное моделирование лингвистических объектов, 2007.
[4] Википедия [Электронный ресурс]. – Режим доступа: https://ru.wikipedia.org.
[5] Сервис RusVectōrēs [Электронный ресурс]. – Режим доступа: https://rusvectores.org/ru/models/ruscorpora_upos_cbow_300_20_2019 (27.01.2021).
[6] NLTK 3.5 documentation [Электронный ресурс]. – Режим доступа: https://www.nltk.org/ (27.01.2021).
[7] Алгоритм k-ближайших соседей [Электронный ресурс]. – Режим доступа: http://datascientist.one/k-nearest-neighbors-algorithm (27.01.2021).