ПРОГРАММНОЕ СРЕДСТВО АНАЛИЗА ТЕКСТОВ ПУБЛИКАЦИЙ НА ЕСТЕСТВЕННОМ ЯЗЫКЕ С ИСПОЛЬЗОВАНИЕМ ТЕХНОЛОГИЙ FLUTTER И JAVA SPRING BOOT

Синицкая К.Д.

Белорусский государственный университет информатики и радиоэлектроники
г. Минск, Республика Беларусь

Сурков Д.А. – ст. преподаватель

В рамках данной работы рассматривается программное средство, позволяющее решать задачу получения новостных публикаций из различных источников, автоматической группировки и отображения их по темам.

На практике есть огромное множество задач, связанных с анализом текста на естественном языке. В рамках данной работы под анализом текста понимается процесс получения, изучения и обработки неструктурированных данных, представленных в виде текста.

Среди подзадач задачи анализа текста можно выделить распознавание текста, исправление ошибок, генерацию текста, выделение частей текста, устранение двусмысленности, перевод, анализ тематики, информационный поиск и многое другое. Выделяется среди этих подзадач проблема получения новостных публикаций из различных источников, автоматической группировки и отображения полученных новостных публикаций по темам. В рамках этой задачи также может производиться оценка эмоциональной окраски публикации, а, в связи с этим, достоверности и правдивости рассматриваемой новости. Возможность настройки анализа эмоциональной оценки может быть полезна для тех людей, которые изначально хотят знать, какие публикации по тональности будут совпадать с их мнением, а какие будут ему противоречить. Это позволяет иметь первоначальную оценку публикации, отталкиваясь от которой любой человек может решить, стоит ли ознакомляться с материалами статьи или он считает, что материалы публикации могут оказаться недостоверными, содержащими только субъективное мнение автора, не подкрепленное фактами.

Для решения проблемы группировки текстов публикаций по темам существует несколько основополагающих подходов, среди которых хорошо зарекомендовал себя подход, основанный на использовании морфологии, выделении частей текста, словосочетаний и морфем. На практике есть несколько вариантов реализации данного решения, среди которых основными являются использование нейронных сетей, создание специализированного ПО с жесткими настройками и комбинированный метод.

Подход с использованием нейронных сетей не оправдывает себя в ситуациях, когда может понадобиться быстрая подстройка ПО, т.к. изначально нейронные сети необходимо обучить на некоторой выборке, а уже затем такое приложение можно будет использовать. Также нейронные сети не дают пояснений, почему они приняли то или иное решение. В свете того, что темы новостных публикаций быстро появляются и меняются, становится очевидно, что использование нейронных сетей не сможет позволить приложению быстро подстроиться и отображать актуальную информацию: пока пройдет процесс обучения нейронной сети на выборке, то может потребоваться переобучение сети для выделения все новых и новых тем.

Необходимость постоянной надстройки зачастую требует создания нового специализированного ПО, которое готово к постоянно изменяющимся требованиям, настройкам анализа и классификации, поэтому было принято решение о создании такого приложения. Учитывая, что большую часть информации в наше время люди получают благодаря интернету, а для быстрого просмотра такого рода информации, к которой относятся и новости, используют телефон, стало очевидно, что лучшим вариантом станет создание приложения с клиент-серверной архитектурой, где в качестве клиента выступит мобильное приложение, а сервер будет построен с использованием быстрых и мощных средств, как архитектурный стиль REST.

В результате было создано такое программное средство, которое получает новостные публикации, автоматически группирует их по темам и отображает пользователю с оценкой тональности каждой конкретной публикации. Данное приложение позволяет отображать конечному пользователю новостные публикации, сгруппированные по темам, показывает оценку тональности каждой публикации, позволяет производить свои настройки классификации, формировать объекты мониторинга, сохранять избранные публикации и т.д.

Список использованных источников:

1. Мэннинг, К.Д. Основы статистической обработки естественного языка / К.Д. Мэннинг, Х. Шютц. – Москва: Техносфера, 2013. – 1048 с.