Д. А. Свядыш

ИЗВЛЕЧЕНИЕ ИНФОРМАЦИИ О ПРОИСШЕСТВИЯХ ИЗ ПУБЛИКАЦИЙ В СОЦИАЛЬНЫХ СЕТЯХ

Рассматривается проблема извлечения информации из публикаций в социальных сетях. Предлагается архитектура системы, которая решает поставленную задачу. Проводится сравнительный анализ инструментов для извлечения данных из текста.

Введение

Современное общество так же, как и прежде, не защищено от чрезвычайных ситуаций, каких-либо происшествий. Важным является оперативное получение актуальной информации о таковых. Возможным решением данной задачи является создание автоматизированной системы, которая будет способна предоставлять актуальную информацию о происшествиях.

I. Требования к разрабатываемой системе

Основными требованиями к разрабатываемой системе являются:

- извлекать, обрабатывать и сохранять в базу данных последние публикации из социальных сетей за n часов;
- предоставлять пользователям возможность регистрироваться в системе, а также предоставлять различные роли;
- предоставлять пользователям возможность просматривать список последних публикаций, с возможностью применять фильтры;
- предоставлять пользователям возможность подписываться на рассылку отчетов о последних публикациях;
- предоставлять администратору возможность настройки параметров системы.

На рисунке ниже приведена предлагаемая архитектура разрабатываемой системы.

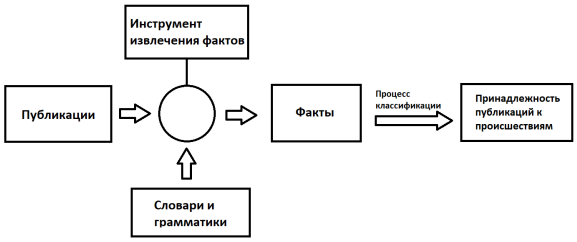

Рис. 1 – Архитектура разрабатываемой интеллектуальной системы

II. Анализ инструментов для извлечения данных

В качестве инструментов для извлечения структурированных данных из текстов на естественном языке рассмотрим Томита-парсер от yandex и Google Cloud Natural Language.

Томита-парсер – это инструмент для извлечения структурированных данных (фактов) из текста на естественном языке [1]. Извлечение фактов происходит при помощи контекстно-свободных грамматик и словарей ключевых слов. Парсер позволяет написать свою грамматику, добавить свои словари и запустить на текстах.

Google Cloud Natural Language API выявляет структуру и значение текстов, используя мощнейшие модели на базе технологий машинного обучения, которые упрощают работу с REST API [2]. Таким образом, разработчики получают возможность использовать данные о людях, местах, событиях и прочих реалиях, которые когда-либо упоминались в текстах новостей, статей и блогов.

Для разрабатываемой системы можно будет выбрать один из предложенных инструментов для извлечения фактов, также возможно их одновременное использование.

III. Выводы

Поиск и отслеживание актуальной информации о последних происшествиях является важным элементом работы многих структур, так что создание и улучшение систем, решающих данную задачу является важным. Предлагаемая система позволит достаточно точно находить искомую информацию из публикаций в социальных сетях.

1. Томита-парсер. https://tech.yandex.ru/tomita/
2. Cloud Natural Language. https://cloud.google.com/natural-language/

Свядыш Дмитрий Алексеевич, магистрант кафедры интеллектуальных информационных технологий БГУИР, dmitry.svyadysh@gmail.com.

Научный руководитель: Сердюков Роман Евгеньевич, кандидат технических наук, доцент кафедры интеллектуальных информационных технологий БГУИР, rserdyukov@gmail.com.