**УДК 004.023**

# МЕТОДЫ АГРЕГАЦИИ НОВОСТНОГО КОНТЕНТА НА ПРИМЕРЕ ПРОЕКТА ONE MEDIA

**Михайлов Д.И.** (Университет ИТМО), **Лучко С.Ю.** (Университет ИТМО), **Лучко А.Ю.** (Университет ИТМО)
**Научный руководитель – кандидат культурологии, доцент ИМРиП Пучковская А.А.** (Университет ИТМО)

В проекте One Media разрабатываются инструменты, предназначенные для сбора, анализа и мониторинга контента, который публикуется и распространяется как в сетевых СМИ, так и в целых платформах и медиаэкосистемах. Одной из важнейших частей проекта являются инструменты, предназначенные для автоматического извлечения и архивирования данных широкого перечня сетевых источников. В данном исследовании представлены основные методы и практики, на основе которых мы построили архитектурное решение сканирования и извлечения веб-данных.

**Введение.** Сегодня интернет-СМИ и медиапространство в целом представляют собой постоянно изменяющуюся сферу. Тот факт, что все основные СМИ сегодня перешли от печатного формата к формату сетевых изданий является очевидным, кроме того, взаимодействуя и публикуя новостные сообщения в социальных сетях и платформах СМИ могут организовывать целые медиаэкосистемы. Все это влечет за собой производство огромного количества контента самого разного содержания и отражающего в себе самые различные явления. Исходя из этого, основной задачей проекта One Media является разработка инструментов и решений, которые позволяют изучать и осуществлять мониторинг этой экосистемы онлайн-медиа.

**Основная часть.** Наше основное решение представляет специальное приложение с помощью которого мы собираем web-данные из новостных источников, затем обрабатываем их, сохраняем, а в последствии делаем доступными через API. Основными строительными блоками инфраструктуры представленного приложения являются следующие элементы:

- Пауки для обнаружения и извлечения необходимых данных.
- Система развертывания и управления сбором данных.
- Управление прокси и обходом возможных блокировок.
- Постобработка данных.
- Обеспечение качества данных.
- Связанные утилиты и технологии.

Инфраструктура приложения реализована с помощью языка программирования Python, а также СУБД PostgreSQL.

**Выводы.** В результате проведенного тестирования и апробации на извлеченных данных из реальных источников, представленное решение способно в автоматическом режиме извлекать всю необходимую информацию из более чем 1000 новостных источников, при этом с возможностью дальнейшего расширения.

| | |
|---|---|
| Михайлов Д.И. (автор) | Подпись |
| Лучко С.Ю. (автор) | Подпись |
| Лучко А.Ю. (автор) | Подпись |
| Пучковская А.А. (научный руководитель) | Подпись |